Global-TRANSIT: Modeling Global Historic Sailing Using A Least-Cost Surface Analysis

Lucinda Roberts[a], Joanna Merson[a], and Woan Foong Wong[b]

[a]InfoGraphics Lab, University of Oregon, Eugene, Oregon, US;
[b]Department of Economics, University of Oregon, Eugene, Oregon, US and Centre for Economic Policy Research, London, England

September 2024

ABSTRACT
Cost-surface analyses in geographic information systems (GIS) can be a useful tool for approximating the travel of historic sailing ships to fill gaps in the historic record. We present the Global-TRANSIT workflow, a least-cost surface raster analysis that uses wind speed and direction to estimate sailing routes and durations for ports globally. Our workflow makes four contributions relative to previously published toolkits. First, our workflow–freely-available as a Python notebook–estimates sail travel for ports at the global scale, while accounting for projection-related challenges. Second, our workflow evaluates origin and destination pairs in a many-origins-to-many-destinations matrix structure (compared to previous one-origin-to-one-destination relationship) which increases the scalability of our toolbox. Third, we adjust the overall workflow to allow for global routing, despite the tool's use of a flat planar projection. Fourth, our workflow replaces deprecated tools used in previous work with newer tools that reduce the grid-induced bias. Despite the expected limitations of modeling a complex phenomenon like sailing, we find a high correlation between our modelled estimates and historically observed sail duration and routes. The outputs of Global-TRANSIT provide an approximation of the likely duration and route of sail travel between worldwide ports, serving as a reference for understanding historic sail voyage patterns globally and as a benchmark for measuring the evolution of maritime shipping over time.

KEYWORDS
Cost-surface; raster analysis; path distance; least cost path; historic sail time

---

CONTACT: Joanna Merson jmerson@uoregon.edu. We thank Bruce Blonigen, Erik Steiner, and Carolyn Fish for their helpful comments and feedback. Peyton Carl, Jack Lei, Maxim Johnson, and Lauren Nguyen from the University of Oregon InfoGraphics Lab provided excellent research assistance. We are grateful to Gianmarco Alberti for providing us with access to the TRANSIT toolbox upon request. This research is based upon work supported by the National Science Foundation Social and Economic Sciences Grant under Award No. SES-1919290 and award PIs Woan Foong Wong and Bruce Blonigen.

1. Introduction

Maritime shipping has played an integral role in facilitating international trade and economic activity throughout history and into modern times. Despite its global importance, the literature on modeling historic travel in geographic information systems (GIS) has largely been devoted to terrestrial travel.[1] Expanding the use of geospatial analyses beyond terrestrial into maritime travel can be valuable, given the significant historic technological advancements in maritime shipping like the move from sail to steam ships.

We present Global-TRANSIT, a freely available least-cost surface raster analysis workflow which allows researchers to flexibly define a collection of input parameters (including wind speed and direction, ship characteristics, and the locations coastal trade ports) in order to estimate a set of sail travel duration and routes for ports worldwide. Global-TRANSIT, available as a python notebook for ArcGIS Pro, can be used by researchers to approximate historic sail travel in the absence of comprehensive data on historic sail records. For example, this tool can be used to fill in temporal or geographic gaps where historic data is limited to records for specific eras and parts of the world. Finding a high correlation between our modelled estimates and historically observed sail times and routes, our workflow produces reasonable comparisons to known historic records and can serve as a reference for understanding historic sail voyage patterns globally.

Our workflow estimates optimized paths, based on the premise that travelers will, over time, minimize the spatial costs of frequently traveled routes, creating what is termed a *least cost path* (LCP) (Herzog, 2013a). In the absence of this workflow, a researcher can perform a least-cost path raster analysis in order to determine a least-cost path (alternatively called a least-cost analysis or a cost-surface analysis in other studies). First, a researcher calculates a raster surface known as an *accumulated cost surface* (ACS), where each cell in the raster represents the accumulated cost of moving across cells to a designated origin location. Second, the researcher uses the ACS to calculate a LCP, which is the route that uses the fewest cumulative resources to travel from the specified origin point to a given destination point (Conolly & Lake, 2006). Overall, the generation of a navigational LCP is non-trivial for an average GIS user (Alberti, 2018; Herzog, 2013b); prompting the development a toolbox and methodology to accomplish this task.

Global-TRANSIT builds upon and makes four contributions relative to previously published toolkits on estimating historic sailing. First, our freely available workflow estimates sail travel times and routes for ports at the global scale. Previous work has either estimated historic sailing times for a specific region or a set of countries. Alberti (2018) published the GIS toolbox TRANSIT (Toolbox foR ANcient Sailing tIme esTimation) which estimates the sail travel times at a regional scale for the Mediterranean Sea. Pascali (2017) applies Dijkstra's algorithm to a nodular ocean grid of wind speed and wind direction to calculate optimized sailing times and routes of ocean transit between countries.[2] Additionally, while the TRANSIT toolbox is available upon request from the author, the workflow from Pascali (2017) provides a high-

[1] Historians and archaeologists have used GIS techniques to estimate historic travel corridors (Scherjon, 2013), the distribution of ancient settlements (Savage, 1990), and analyzing ancient road networks (Conolly & Lake, 2006).

[2] The modeled sail duration in Pascali (2017) is then compared against a compiled dataset of steam engine transit times to understand how the adoption of the steam engine impacted global patterns in maritime transit and trade.

level overview of the algorithm and inputs but does not share the specific tools or implementation steps required to work through the process which limits its use by other researchers. Our workflow is freely available as Python scripts for ArcGIS Pro at https://figshare.com/s/8517eb49981f658df6ff .

Our second contribution is that Global-TRANSIT evaluates origin and destination pairs within a matrix structure accommodating many origins and many destinations, compared to previous work which focused on one-origin-to-one-destination relationships. This feature increases the scalability of our toolbox. The Alberti (2018) TRANSIT toolbox automated the process of generating the ACS for a single origin. However, the goal of the TRANSIT tool is to estimate sailing times, not routes. Therefore, the second part of a second part of the least cost raster analysis—generating a LCP from the ACS—must be implemented by the individual researcher.This confines the user of this toolbox to generate one route at a time which limits the tool's ability to scale up and accommodate a large number of origins and destinations.

The third contribution is a solution to the limitations of the planar projection used in Esri's Distance toolkit. Planar projections model the globe as a flat plane, not allowing data to "wrap" across the edges of the earth. While the TRANSIT toolbox was not functionally limited to any spatial scale, the implementation in the distance tools precluded researchers from modeling routes which wrapped around the earth. In our TRANSIT-global model, we implement an option for researchers to run the model centered at antipodal longitudes and programmatically compare the modeled outputs. This allows for researchers to approximate routes at scales from a regional to global level.

Our workflow's fourth contribution is to update deprecated tools used in previous work with newer tools. The TRANSIT tool was built in ArcMap, which has moved into extended support and will ultimately be retired by 2026 (Esri, 2024). It also uses the deprecated Path Distance and Cost Path as Polyline tools, which will be removed from future software updates. Significant updates to these tools are currently required in order for the tool to be implemented. The lack of access to these tools in the future could pose significant additional hurdles for researchers (Gheorghiade & Spencer, 2024). Additionally, the newer tools— Distance Accumulation and Optimal Path as Line— allow for measurement using geodesic distances and reduce the grid-induced bias from the modeled travel across a raster surface.

In the rest of the paper, Section 2 provides background information to explain our model and the related literature, including an explanation of least-cost surface raster analyses (Section 2.1), relevant parameters for modeling sailing routes (Section 2.2), and an overview of the relevant historic records (Section 2.3). Section 3 provides an overview of the Global-TRANSIT workflow and how it expands on previously published models. Section 4 describes the calibration and results when validating our model against the assembled historic record. Discussion (Section 5) and conclusion (Section 6) follow.

2. Modeling of Historic Sailing in GIS

In this section, we present the background for Global-TRANSIT, a workflow for using a least-cost path raster surface analysis for a global set of shipping routes that can flexibly account for ship characteristics and projection constraints. This methodology builds on tools and workflows from previous researchers who have endeavored to model historic shipping, where their focus was on one local area (Alberti, 2018; Gheorghiade

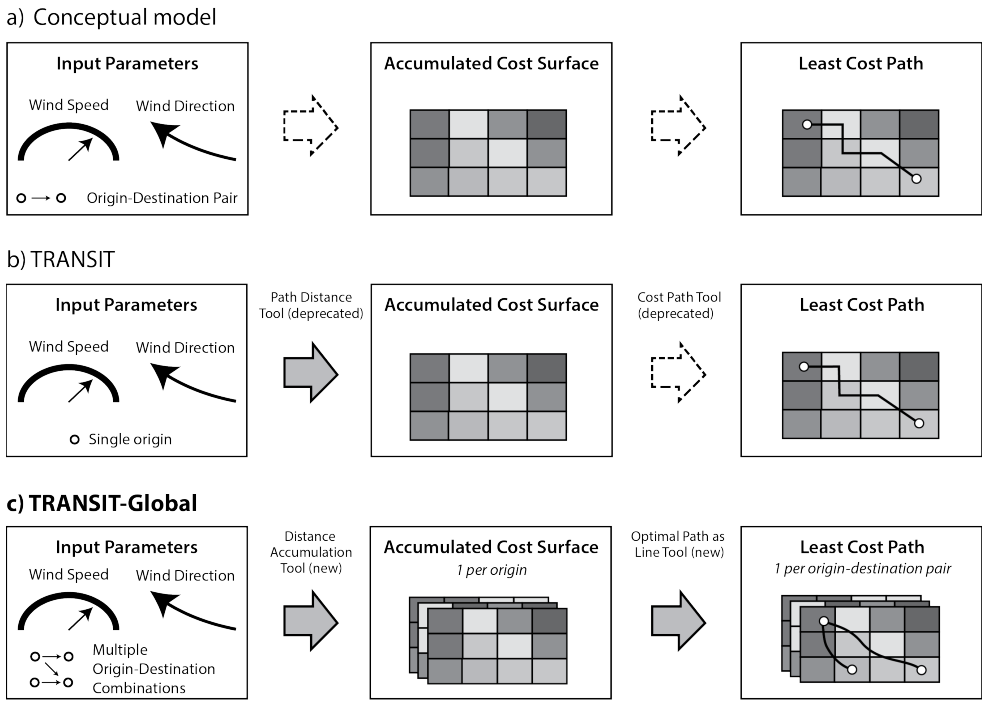


Figure 1. Conceptual diagram of a least cost surface raster analysis, TRANSIT model, and Global-TRANSIT model

***Notes*:** Panel (a) displays a conceptual diagram of a least cost surface raster analysis. Panel (b) displays how this analysis was implemented in the TRANSIT model. Panel (c) displays how this analysis is implemented in the updated Global-TRANSIT model, including major input parameters, tool updates, and scalability. Dashed arrows indicate a conceptual or manual process while filled-solid arrows indicate an automated process.

& Spencer, 2024) or at the country-level with limited implementation details (Pascali, 2017).

2.1. Least-Cost Path Raster Surface Analysis

The core heuristic of this workflow is a least-cost path raster surface analysis, (shown in Figure 1(a) Conceptual Model), a geospatial process for determining an optimum path which consists of two parts (Antikainen, 2013). First, a continuous space is divided into a regular set of areal units (typically nodes or a raster surface) and the surface is weighted according to a cost (such as time, distance, or money) that is associated with moving through that cell (Mitchell, 2012). This surface is used to calculate an ACS, where each unit in the surface represents the accumulated cost of moving across cells to a designated origin location (Alberti, 2018; Antikainen, 2013). Second, the ACS is then used to calculate a LCP, which is the route that uses the fewest cumulative resources to travel from the specified origin point to a given destination point (Conolly & Lake, 2006). The LCP represents an optimal path across a landscape which differs from a Euclidean (or shortest) distance. While an LCP Raster surface analysis can be used to model many modes of movement, the specific parameters defined in Global-TRANSIT allow a researcher to apply it specifically to historic sailing and those parameters are explained in the next section (Section 3.1).

2.2. Relevant factors for sailing navigation

In a literature review of relevant factors for modeling shipping routes, Alberti (2018) found that factors that can influence the cost of maritime transit include 1) sea-state, 2) human factors, 3) wind direction, 4) wind speed, 5) ocean-current conditions, and 6) ship characteristics.

The first two, sea-state and human factors, occur on a short time scale and can be anticipated by experienced sea navigators. Sea-state is a dynamic variable which can impact the route a given ship takes. Perttola (2022) adopted Alberti (2018)'s model to dynamically update wind data and account for sea-state at a more granular level. However, this methodology is computationally expensive, even for a relatively small number of origin-destination pairs. Because of the computational cost, we do not incorporate Perttola's methods; although, this could be a fruitful area for future research (see discussion in Section 5 for more details).

Human factors are also difficult to factor into a model; however, as Whitewright (2011) and Herzog (2013a) emphasize, the capacity to manage human contingencies and anticipate sea-state would have been a skill that a captain brought to the ship and, over time, the sailing routes would have naturally mimicked a least-cost path despite human and sea-state factors. As such, this workflow does not incorporate sea-state and human factors.

Instead, we incorporate the third and fourth factors above, wind speed and direction, as input parameters in our model. Overall global prevailing wind patterns have been relatively similar for the last 2500 years, despite the recent impacts of climate change. This allows other researchers to approximate historic and ancient sailing using modern winds (see, for example, Murray (1987), McGrail (2009), Alberti (2018), Gheorghiade and Spencer (2024)). We use the oldest-available data from our chosen wind speed and wind direction dataset (the Hersbach et al. (2018) data which was introduced in Section 3.1.1), the year 1940, as inputs for our wind parameters. Although imperfect, this dataset is the earliest readily available dataset with a fine enough spatial resolution for our purposes.

The fifth factor, currents, affect objects on a 1:1 ratio; meaning that any object with a significant portion above water will be more impacted by wind speed than ocean currents (Fitzpatrick & Callaghan, 2008). Additionally, oceanic navigation is largely driven by superficial ocean currents, which are in turn driven by wind speed and wind direction (Alberti, 2018; Fitzpatrick & Callaghan, 2008). Since we already incorporate wind speed and wind direction parameters, we do not include an additional parameter to represent ocean currents in our model of sailing navigation.

The sixth and final factor, ship characteristics, is incorporated into our model in two ways. The first way is by providing a maximum ship speed parameter, since a ship cannot go as fast as the wind it is sailing in (Alberti, 2018). Second, since sailing travel directly into, against, or at angle to the wind produces varying levels of relative challenge, a parameter is included to represent how sailing vessels travel. This is done using a frictional travel factor is referred to as the *horizontal factor*. Both the maximum ship speed and horizontal factor are described in more detail in Section 3.1.2.

2.3. Historic records for calibration and validation

Because of the number of interrelated factors involved in executing a least cost raster analysis, it is important to calibrate the model parameters and validate that the model is functioning successfully (Herzog, 2013a). In calibrating and validating our model, we

compared our workflow's outputs against a compiled list historic records of maritime sailing times between global ports.

In order to compare our model estimates against actual observed travel *times*, we compiled a list of historic travel records that has a global geographic distribution from online sources. To do so, we performed a keyword search into Google Search Engine and looked for recorded historic travel records which had been publicly published. Our compiled historic travel record validation set includes sources from Albion (1938), Chichester (1967), Gumport and Smith (2006), Kingsley (2020), and The Maritime Heritage Project. In total, we collected 102 recorded travel times (64 unique origin-destination port pairs) from the years 1400 to 1900s, which are detailed in Section 4.1.

Additionally, in order to compare actual travelled *routes*, we leverage the Climatological Database for the World's Oceans (*CLIWOC*), which was published through the Royal Netherlands Meteorological Institute. The database includes over 280,000 point records from the logbooks of European sailing vessels from 1750-1850. While this data is useful for visually comparing modelled routes against known historic records, its effectiveness for comparing against modelled travel *times* is more limited due to two reasons: because the records lack information on intermediate stops—both at port or at sea—which are typically part of actual routes, and because the records do not specify whether the time taken for these stops is included in the total travel time (Pascali, 2017).

3. Global-TRANSIT Workflow Overview

This section provides an overview of the Global-TRANSIT workflow, the driving input parameters, and how the workflow handles global modeling. Lastly, we discuss the benefits of migrating the model into Python and updating the underlying tools.

We created a workflow that can evaluate origin and destination port pairs in a many-to-many structure, with the framework of the TRANSIT toolbox (Alberti, 2018) at its core. Thus, it uses a similar workflow and inputs. Moving from a 1-to-1 origin-to-destination relationship in the TRANSIT toolbox to a many-to-many origins-to-destinations matrix increased the scalability of the toolbox. The updated Global-TRANSIT workflow also includes steps to account for projection limitations, and can optionally support spatial constraints for comparison.

Figure 2 displays the overall workflow, organized into three parts, each a python notebook: Panel (1) shows an optional wind pre-processing notebook (1. Process Wind Data), Panel (2) highlights the Global-TRANSIT toolbox (2. Global-TRANSIT), while Panel (3) shows the optional post-processing notebook for compiling the output travel-time data into comparable lists (3. Compile travel time list).

3.1. Input variables and parameters

Based on the reviewed literature (in Section 2.2), the input parameters driving the model are the 1) wind conditions, 2) the ship characteristics using max speed and a frictional factor (termed *horizontal factor*), and 3) trade port locations used as the origin and destination pairs for the LCPs. Details about each of these input parameters are outlined below:

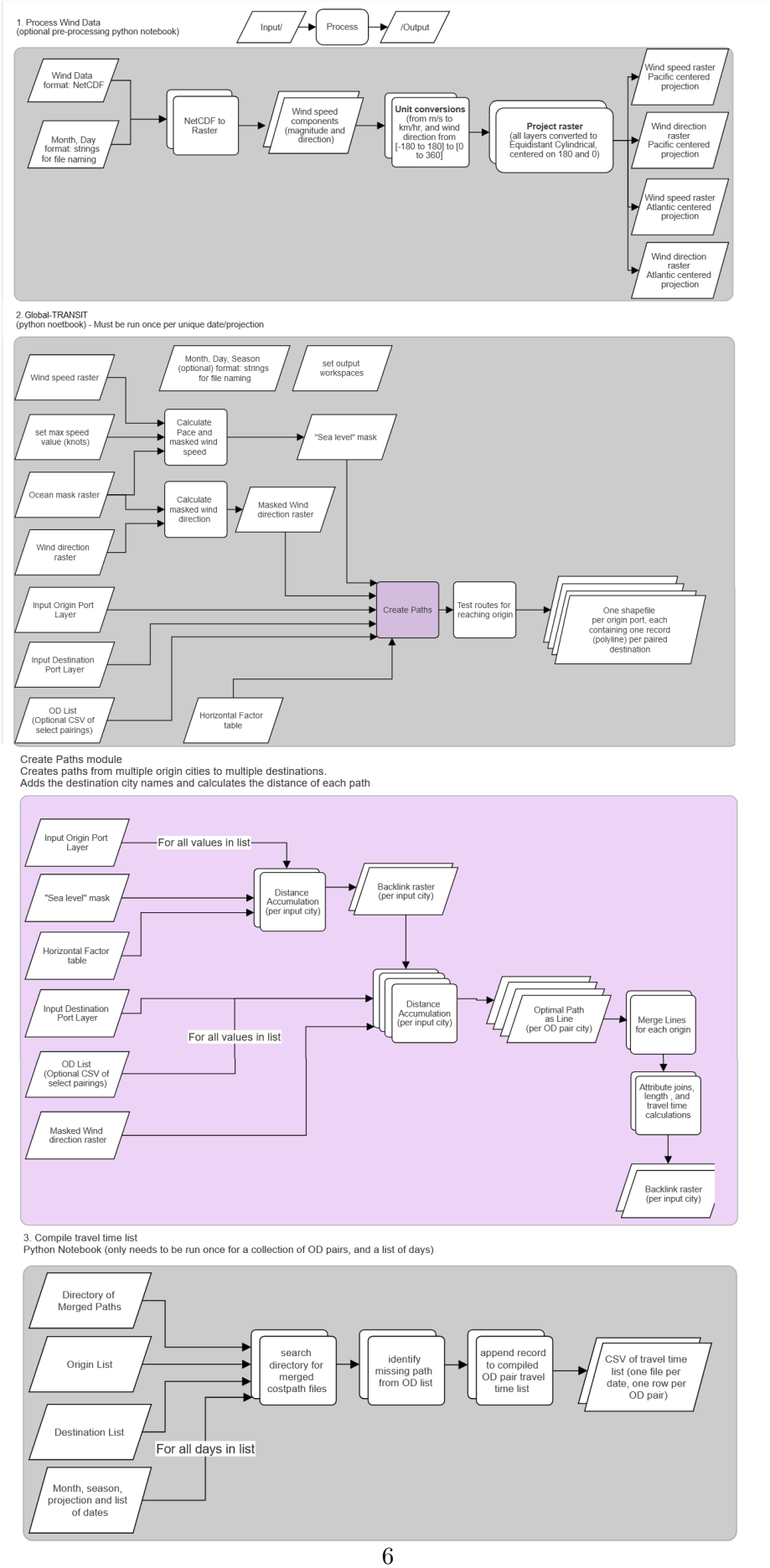


Figure 2. Global-TRANSIT workflow

***Notes*:** This figure includes Panel (1) an optional pre-processing wind conversion tool, Panel (2) the Global-TRANSIT toolbox, and Panel (3) a post-processing tool to quickly assemble values for comparison.

3.1.1. Wind conditions (Speed and Direction)

As the central variables identified for modeling sailing navigation, the input wind speed (magnitude) and wind direction are the main model inputs. The Global-TRANSIT model requires an input wind-speed and an input-wind direction raster with global coverage. As discussed in section 2.2, modern wind data (measured or modeled) can be used to approximate historical prevailing wind patterns. The wind conditions, and thus data which are selected, depend on the researcher's goal. Historic or modeled wind may be chosen to be representative of a specific storm season or the model can be re-run over a variety of days to represent a range of sailing conditions. These data can be derived from a variety of meteorological sources such as the World Oceanic Circulation Experiment's (WOCE) surface wind velocity data (Woods, 1985) or ERA5 model (Hersbach et al., 2018). The data which are typically available as a u- and v- components (meteorological convention) need to be converted to magnitude and direction and require unit conversion. A pre-preprossing python notebook for this conversion, from data in NetCDF format to individual, projected rasters, is provided in the workflow (Panel (1), Figure 2).

3.1.2. Ship Characteristics (Horizontal Factor and Max Speed)

For sailing vessels, travelling directly into, against, or at angle to the wind produces varying levels of relative challenge. This is considered an *anisotropic cost*, meaning the cost is related to the direction of movement (Mitchell, 2012), and is accounted for in an ACS analysis as a horizontal factor (HF). The HF we use for representing historic sailing vessels is provided as part of the workflow and its derivation is described in this section. Nevertheless, this table can be adjusted to represent different ship characteristics and could be calibrated for different scenarios in future research.

In Global-TRANSIT, the HF is an parameter of the ArcGIS Pro Distance Accumulation tool, set using a horizontal factor table, where the user specifies values from 0-180 degrees where, according to geographic (not maritime) conventions: 0 is directly with the wind and 180 is directly against the wind. A horizontal factor of 1 is a neutral factor and will leave the cost of traveling between cells unchanged. A factor less than 1 will decrease the associated challenge of traversing a cell in that direction, and a factor more than 1 will increase the challenge of traversing a cell in that direction. If a value is not provided in the HF table all the way to 180 degrees (i.e. if the user specifies values from 0-113), the tool assumes the value of the HF to be infinity (Esri). Alberti (2018) provided a horizontal factor following geospatial conventions based on a literature review with a HF value of 1 for running (0-34 degrees), 0.42 for broad reach (35-67 degrees), 1 for beam reach (68-90 degrees), 2.5 for close-hauled (91-113 degrees), and a high factor of 10 for the "no-go" zone (113-180 degrees). In the Distance Accumulation tool, the HF is assumed symmetrical for the remaining 180-360 degrees.

The *max scale value* represents the upper limit of how fast the model can approximate the ship's speed. The speed the ship is estimated to be able to travel across a given cell is driven by the wind speed, but does not exceed the maximum scale value, thus is capped based on the assumed max travel of a given vessel type.

Section 4.2.1 discusses our calibration choices for both the HF and max scale value.

3.1.3. Trade Port Locations

The Global-TRANSIT workflow can be run in a many-to-many relationship for all possible combinations of input ports or a specified subset. The user must specify an input layer containing origin ports and an input layer containing destination ports. Using the same layer for both inputs, produces LCPs for all potential port combinations, with all ports represented as both origin and destination, modelling routes both *to* and *from* each OD pair. In this case, the model runs with an n^2 processing speed with n being the number of ports. To speed processing, using different layers for the origin and destination inputs, the model runs with an mn processing speed – with m being the total number of origin ports, and n being the total number of destination ports. Additionally, the user can optionally specify a subset origin-destination pairs in a CSV list. See section 4.3.1 for more information on how we selected our list of ports by reviewing historic information.

3.2. Adjusting to global scale

Global-TRANSIT aims to create a versatile workflow for approximating sailing duration and routes from origin to destination ports worldwide. Researchers can utilize this workflow to approximate global historic sail voyage patterns and the evolution of maritime shipping. The original 2018 TRANSIT toolbox automated the process of generating the ACS for a single origin location, using the Mediterranean as an example. A researcher had to subsequently generate the LCP from the origin to a given destination following a workflow outlined in the publication. That limited the scalability of the TRANSIT model to only modeling single origin-destination pairs and manually generating the least cost paths.

The Global-TRANSIT model can be run in a many-to-many relationship because we incorporated two iterators into the Global-TRANSIT model (Panel (c), Figure 1). The first loops through the features in the input port origin layer and generates an ACS for each origin. The second loops through the input port destination layer to generate a LCP from each of those origins to each of their destinations. This adaptation of the tool was accomplished by migrating the workflow from ArcGIS Model Builder into ArcPy which we explain in further detail in Section 3.4.

Modeling at a global scale presents a new challenge, because the raster-based toolsets in Esri's desktop applications do not wrap around the edges of the projected map. Thus, the edges of the modeled world would not connect from one side of the given projection *off the edge* to the other side. To address this issue, we ran each model twice – once using a projection centered at the Prime Meridian (0 degrees longitude), and once using a projection centered 180 degrees of longitude, so each projection was centered on opposite locations on the globe. This allows for two potential routes to be taken between each OD pair. In some cases, such as from London to Lisbon, there is no difference in the modeled routes between the two projections. However, from a Western port, like San Francisco, to a port in Asia, like Manila, there is a large difference between the Pacific- and Atlantic-centered LCPs, as demonstrated in Figure 3.

Additionally, modeling sailing at a global scale with specific winds meant that with some input wind conditions, certain destinations could not be reached from certain origins. In these cases, the Optimal Path as Line (OPAL) tool creates an incomplete path which does not geographically connect the origin and destination, but no errors or warnings are output from the OPAL tool. To alleviate the need for the user to visually

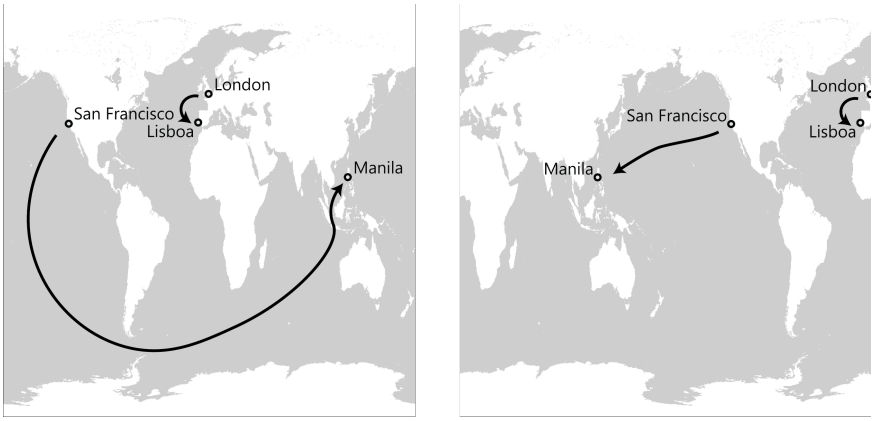


Figure 3. Potential routes from London to Lisbon and San Francisco to Manila displayed in the two projections used for Global-TRANSIT

inspect every output path in an analysis with many OD pairs, we built an automated sanity check: incomplete paths are flagged with an attribute so the researcher can quickly identify areas routes or ports that require interpretation.

3.3. Discussion of updates to the core tools

The original TRANSIT model is built on the Esri's desktop application Path Distance and Distance Accumulation tools from the Spatial Analyst toolbox. These tools have been deprecated by Esri and will not be included in future software releases (Gheorghiade & Spencer, 2024), so replacing them with the updated tools in Global-TRANSIT has multiple advantages: the new workflow has a longer operational life and the tools themselves have improvements to the underlying algorithms, such as mitigating systematic biases in the cost surface accumulation algorithm (TenBrink, 2019).

In the now deprecated Path Distance tool, the cell-to-cell movement uses a queen movement paradigm where, like in chess, the ship can move from any cell to any surrounding cell, including diagonal movement (Mitchell (2012), Panel (b) in Figure1). Unlike a queen in chess, however, the path distance tool can only move to the nearest 8 cells around the current location. Horizontal or vertical movement is calculated to be $resolution \times 1$ and the distance between adjacent diagonal cells is the raster resolution times the $\sqrt{1^2 + 1^2}$, or $resolution \times 1.41421$ (Mitchell, 2012) (Figure 4a). The forced movement along a gridded raster surface, such as the Path Distance tool uses, is one of the largest systemic biases in least-cost path analysis, and can lead to overestimating the optimum routes (Alberti, 2018; Herzog, 2013b; Perttola, 2022). To mitigate this error, some researchers have tried to use less-common raster surfaces, such as hexagonal raster grid or expanded the number of neighbors a computer may consider to allow for more angles in travel (Antikainen, 2013).

We have opted to continue to use the more traditional square raster grid, which matches the format of most available wind data,upgrading from the deprecated Path Distance tool to the newer Distance Accumulation tool. The updated algorithm does not constrain the object (the ship) to a cell-to-cell movement, as a "network problem" (Esri, ArcGIS Pro 3.3b), and instead allows it to travel along diagonal angles that more closely approximate the ways that ships would travel (TenBrink, 2019). The updated algorithm uses concepts from differential geometry to remove the 8-way direction problem (where least cost paths start and end at cell centers). Instead, the Distance

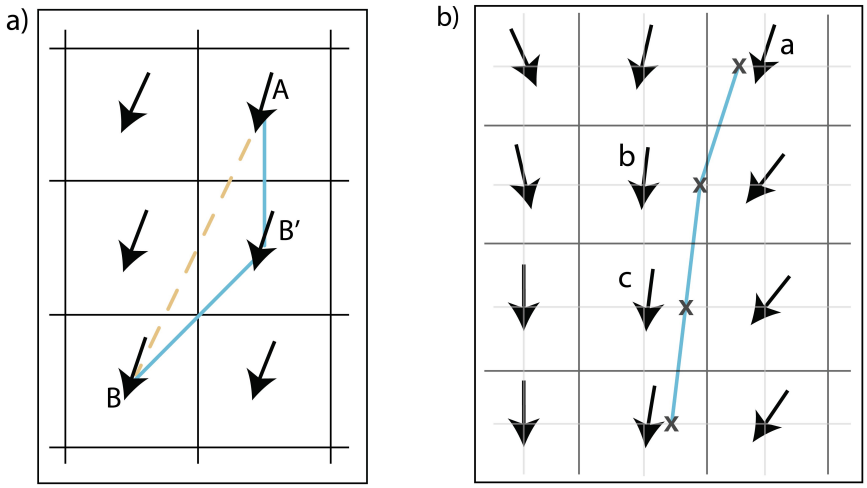


Figure 4. A comparison of the paths found using the deprecated Cost Path tool and the distance accumulation tool. Adapted from Esri. (arcGIS Pro 3.3) Distance accumulation algorithm.
***Notes*:** Panel (a) shows cell boundaries in black; cell centers as azimuth arrows. The shortest straight line path from A to B (orange line) will not be discovered by the deprecated tools. It will instead use A-B'- B as the shortest path (blue line), because it can only move in eight directions from cell-center to adjacent cell-center. Panel (b) shows the cell boundaries in dark gray and the cell values are shown as azimuth arrows at each cell center. The lattice connecting cell centers is shown in light gray. As the least cost path crosses lattice lines, it uses the azimuth in the closest back direction cell in the direction of travel to update its direction. At the top path vertex next to cell "a", the back direction value stored in "a" will be used to direct the line leaving that vertex. The next lattice line to be crossed is closest to cell "b", so that azimuth will be used to exit the second vertex, and so on. The path discovered by the toolset is the blue line.

Accumulation path vertices can be anywhere on the lattice of horizontal and vertical lines running through the cell centers. (Esri (ArcGIS Pro 3.3b), Figure 4b). The result is that the outputs from the Distance Accumulation tool travel more "smoothly" across space and minimizes grid-induced bias from the modeled travel across a raster surface (Esri, ArcGIS Pro 3.3b, ArcGIS Pro 3.3c). Another benefit of the algorithm used by the Distance Accumulation tool includes using geodesic distances instead of Euclidean distance metrics (TenBrink , 2019a). The Esri documentation (Esri, ArcGIS Pro 3.3b, ArcMap 10.8b) explains the differential geometry algorithm in considerably more depth, and as an implementation of a combination of techniques described in Sethian et al. (1999) and Zhao (2005).

3.4. Benefits of python-based model

The original TRANSIT toolbox was built in Model Builder and migrating the Global-TRANSIT toolbox to Python had several benefits for the functionality and scalability, of the model.

Most of Esri's geoprocessing tools can be executed manually individually or linked together using Model Builder – a graphical user interface (GUI) –, or scripted using python, specifically Esri's python package ArcPy. As noted earlier, two of our main goals in creating the Global-TRANSIT model are to increase scalability and allow for many-to-many relationships. This involved constructing a set of nested iterators to loop through the input origins and input destinations and we had to nest multiple tools which independently called each other. The Model Builder GUI environment made the implementation of iterators challenging, as it does not smoothly allow for nested iterators and changes to intermediate paths and input parameters did not always

communicate smoothly between the tools and across users due to known limitations with iterators in ArcGIS Model Builder (Esri). Iteration and paramaterization in the Python environment are notably more transparent and stable.

Additionally, processing power was increased by developing the model in ArcPy. Since the model can run at up to an n^2 runtime when it is comparing all ports as origins and destinations, the number of input ports can have a dramatic impact on the model's runtime. By developing the model in ArcPy, the tools are able to leverage parallel or multi-core processing, which utilize multi-core CPUs by dividing and performing operations across multiple processes to speed up the performance of geoprocessing tools (Esri).

Ensuring the validity of our modeled data requires applying a *horizontal factor* (HF) used to approximate the challenge of moving at varying angles with and against wind (see Section 3.1.2). However, when we developed the Global-TRANSIT toolbox in ArcGIS Pro, we noticed that when we changed the HF table, it did not modify the model outputs. Through correspondence with Esri, we determined this was a "known bug" in Model Builder where the HF table was silently reverting to a default setting (Personal communication, 19th April, 2023). This known bug is not replicated when the Path Distance tool is run in from Python.

Overall, creating Global-TRANSIT using Python notebooks, not only builds upon the methodology of previous work, but also updates our workflow to be more scalable, editable, updatable, transferable and more reliable.

4. Model Validation against Historic Sailing Routes

In this section, we test the estimates of our mode estimates against historically observed sailing records. We start by describing these records, then explain the calibration of our model, and then detail the validation exercises.

4.1. Historic Observed Sailing Routes

In order to test the validity of our model estimates, we collected observed duration data on historic sailing routes from several sources and compare our model-generated sailing duration to these routes. Our validation set includes Albion (1938), Chichester (1967), Gumport and Smith (2006), Kingsley (2020), and The Maritime Heritage Project, with the full list, including sources, listed in Appendix Table A.1. The set of historic routes includes 102 observations which contains 64 unique routes. There are 77 ports, of which 19 ports are both an origin and destination, 33 are an origin only, while 23 are a destination only. The historic observed set includes major global trade ports, such as London, Shanghai, New York/Newark, and Sydney, and numerous other globally distributed ports (Figure 5).

Our collected historic records represent a range of spatiotemporal coverage. Spatially, the identified historic origins and destinations span every continent except Antarctica, including the United States and Canada (13 observations), Central America and the Caribbean (5 observations), South America (5 observations), the Pacific Islands (2 observations), Australia (3 observations), Indonesia (2 observations), East Asia (3 observations), the Indian Subcontinent (2 observations), the United Kingdom (4 observations) and mainland Europe (2 observations). Temporally, we prioritized

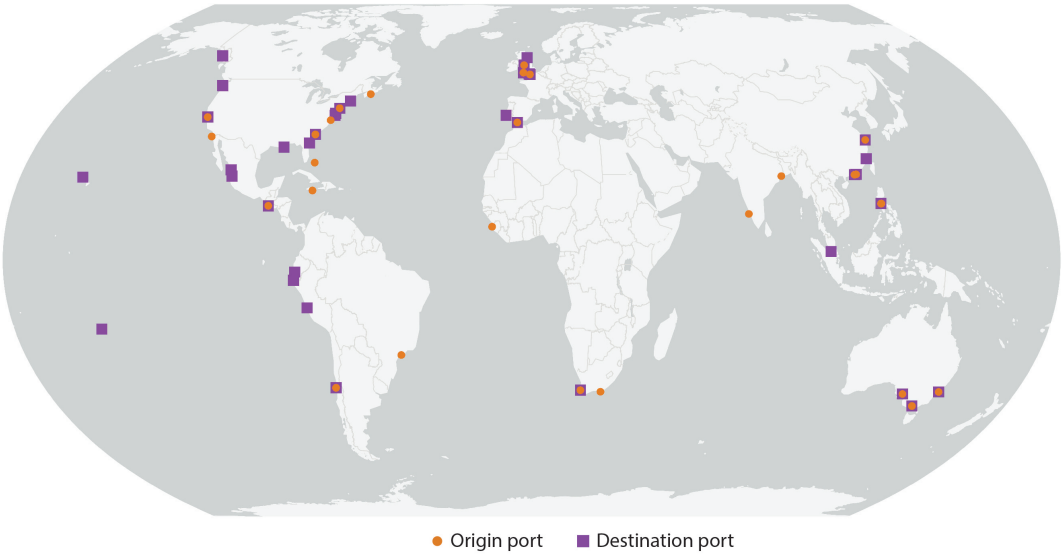


Figure 5. Map of the origins and destination ports in the validation set.

***Notes*:** There are 102 observations in the historic set. There are 77 ports in this set, of which 19 ports are both an origin and destination, 33 are an origin only, while 23 are a destination only.

identifying 19th century records from before the Suez Canal was opened (88 observations). In order to make sure we collected as many records as possible to validate our model, our records primarily range from 1492 to 1949.

4.2. Calibration and Model Output Validation

4.2.1. Calibration

The first parameter that was calibrated for our validation was the maximum speed. Among sailing ships in the mid-19th century, the fastest sailing ships were found to go up to 22 knots (Howe, 1986). Since we are approximating the average ship travel at the time, instead of the fastest sailing ships available, we calibrate 20 knots (37.04 kilometers per hour) as the maximum scale value for our analysis. This results in an implied average speed (the distance divided by duration) which ranged between 5-11 knots for our modeled routes, which closely approximates the implied speeds of historic sources (Casson, 1995).

The second parameter that was calibrated for our validation was the creation of the HF table. Given that ERA5 data has a relatively coarse resolution at 0.25° by 0.25°, we found that the HF of 10 for the "no-go" zone was too high to permit travel along challenging windy corridors, which resulted in the modeled routes being disconnected from destinations we know they were able to travel to, based on historic data. In actuality, we know that ships were able to tack back-and-forth when the winds were unfavorable. To better mimic this, we lowered the "no-go" zone factor to 5, meaning there was still a significant challenge to transit against the wind, but that these areas were still passable. Calibration of this modification was completed through a sensitivity

analysis of the modeled outputs against the historic observations.

With the singular modification to the "no go" zone, we otherwise kept the Alberti (2018) HF table constant in our validation. We maintain the Alberti (2018) HF table because it is suitable for the purposes of validating our heuristic model. However, we acknowledge that this parameter can be adjusted to represent different ship characteristics and could be calibrated for different scenarios in future research. It is our hope that as the literature on modeling historic sailing develops, future researchers will contribute more robust studies on different horizontal factors for different sailing and maritime transit modalities.

4.2.2. Model Output

To generate our model estimates for comparison with the historic sailing routes, we use the earliest year, 1940, of wind data from the ERA5 model (Hersbach et al., 2018). We pick the months of June and December to compare to the historic routes, rather than averaged meteorological data due to the computational complexity involved in the averaging process. The choice of these two months is to reflect the fact that different months within the year provide different regions with favorable sailing conditions. We also made this choice because, looking at records of past hurricane seasons, that year's hurricane season and weather patterns were notably less volatile than more recently modeled years. This relative stability may be due to the fact that, while climate change may not have drastically altered prevailing wind patterns, it has increased the frequency of large storms that would disrupt sail travel. For the Indian Ocean, April to October winds allow ancient sailors to go from Egypt to India (known as the south-west monsoon), while the November to March north-east monsoon winds powered the ships back to Egypt (Beresford, 2012). For the Mediterranean, the summer months of May to September are considered part of the best sailing season, while the late fall and winter months are when sailing is less optimal (Casson, 1995). Additionally, we avoid the months of August, September, and October due to hurricane season. Showing how our estimates, based on June and December wind data that provide different regions of the world with different sailing conditions, compare to historic records can help establish bounds on the validity of our model. For each June and December 1940 sail estimate, we take the average of 7 days at 5-day intervals: the 1^{st}, 5^{th}, 10^{th}, 15^{th}, 20^{th}, 25^{th}, and 30^{th}.[3]

To complete the Global-TRANSIT workflow, (Panel (2) in Figure 2) we run the first notebook, Process Wind Data, 12 times (6 days per two of the months). We then ran the second and third notebook, the Global-TRANSIT Model and Compile Travel Time List, 24 times (2 projections per unique date). There are 52 origin ports, so the core Distance Accumulation tool, which calculates the ACS, is automated to run for each origin port, was effectively run 1,280 times (24 x 52). The Optimal Path as Line tool, which calculates the LCP, is automated to run for each OD pair (102), was effectively run 2,448 times (24 x 102). Utilizing this model and workflow, the intermediary and output shapefiles and CSV of the historic sailing records are organized, reproducible, and easily compared. The following section details our comparison and validation against a historical set.

[3]For further robustness check, we also compare the median of our June and December estimates to the observed historic sailing route duration for each port-to-port combination. We find similar results.

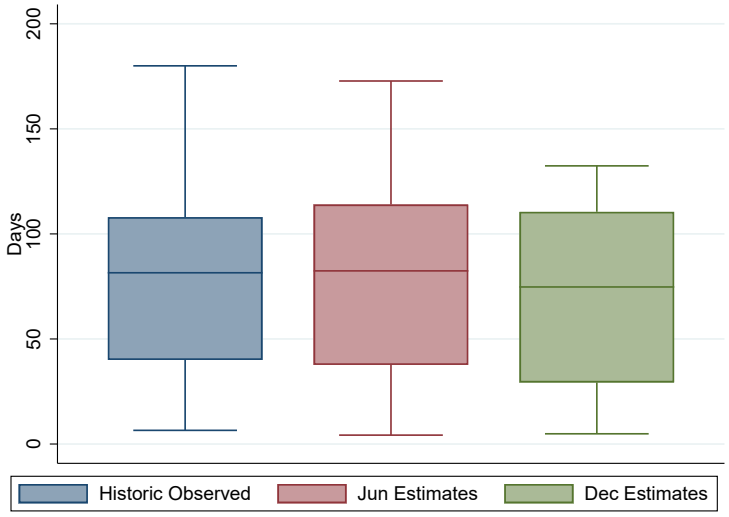


Figure 6. Box Plot Observed and Estimated Durations

***Notes*:** There are 102 historic observed route duration that are collected from historic records (see section 4.1 for further information). The model's June Estimates are the average of the model's estimates using 7 days of wind data in June 1940 (June 1st, 5th, 10th, 15th, 20th, 25th, and 30th). The model's December Estimates are the average of the model's estimates using 7 days of wind data in December 1940 (December 1st, 5th, 10th, 15th, 20th, 25th, and 30th).

4.2.3. Validation

Figure 6 reports the distribution of the observed duration and the model's June and December estimates for that same port-to-port route while the first five columns of Table 1 reports their mean, median, minimum, maximum, and standard deviation. While Figure 6 shows that the observed duration has a larger spread than both the model's June and December estimates, the middle range of the observed duration has significant overlap with both estimates. The mean of the observed duration is 76.6 days, which is in between the June and December estimates—lower than the 81.1 days mean from the June estimates but higher than the 70.9 days mean of the December estimates (Table 1, Column (1)). The median of the observed duration is 81 days, which is also in between both the model's June and December estimates (Table 1, Column (2)). The standard deviation for the observed duration is also similar to the standard deviation for the model's June and December estimates (Table 1, Column (5)).

Next, we utilize two statistical methods to evaluate the accuracy of our model's estimates compared to the observed historic values. First, we employ the root mean square error (RMSE) which calculates the error magnitude between the estimates and observed values by taking the square root of the average of the squared differences between estimated and observed values. A lower RMSE indicates smaller discrepancy between these values. Second, we use the correlation coefficient which quantifies the strength and direction of the linear relationship between the estimated and observed value. A higher positive value indicates a stronger positive relationship.

Table 1 reports the RMSE and correlation coefficients between the model's estimates and the observed historic values. Both model estimates report low RMSEs compared to the observed values. The model's June estimates has a RMSE of 0.355 while the

Table 1. Summary Statistics, Root Mean Squared Error, and Correlations of Observed and Estimated Durations

Duration (days)	Mean (1)	Median (2)	Min (3)	Max (4)	SD (5)	RMSE (6)	Corr (7)	Cargo Corr (8)
Historic Observed	76.3	81.5	6.5	180.0	41.5			
Jun Estimates	80.2	82.4	4.2	172.8	46.6	0.356	0.909	0.926
Dec Estimates	71.0	74.8	4.9	132.4	40.3	0.340	0.906	0.925

***Notes*:** There are 102 historic observed route duration that are collected from historic records (see section 4.1 for further information). The model's June estimates are the average of the model's estimates using 7 days of wind data in June 1940 (June 1st, 5th, 10th, 15th, 20th, 25th, and 30th). The model's December estimates are the average of the model's estimates using 7 days of wind data in December 1940 (December 1st, 5th, 10th, 15th, 20th, 25th, and 30th). Column (1) reports the average of each row, Column (2) reports the median, while Columns (3) and (4) report the minimum and maximum respectively. Column (5) reports the standard deviation of each row. Column (6) reports the Root Mean Squared Error (RMSE) between the observed duration in row 1 and each of the estimated duration in rows 2 to 4. Column (7) reports the correlation coefficient between the observed duration in row 1 and each of the estimated duration in rows 2 to 4. There are a subset of the historic records that are cargo trips. Column (8) reports the Pearson correlation coefficient of these cargo records between the observed duration in row 1 and each of the estimated duration in rows 2 to 4. Columns (6) to (8) are calculated with logged values.

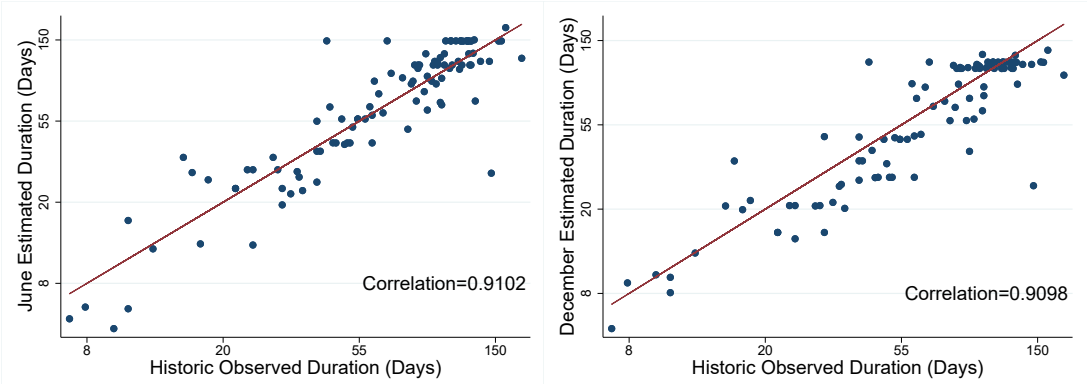
December estimates has a RMSE of 0.338 (Table 1, Column (6)). Both estimates also report high positive correlations with the observed historic values. The model's June and December estimates have a correlation of 0.910 with the observed historic values (Table 1, Column (7)).[4]

By the mid-19th century, most shipping countries had liners, or ships sailing by fixed and advertised dates, which had diversified into passenger and cargo routes (Dunkley & Stamper, 2016). While collecting the historic data, we found that a majority of the historic records that were for cargo (67 out of the 102 observations). Some of the other records are for passengers and others we do not have information on the trip purpose. As ships with passengers are typically more leisurely and involve multiple stops so that their passengers can visit more locations (Austin (2021); Duke University Digital Collections), the historic observed records with passengers may perhaps take less direct routes. Our workflow, which approximates a least-cost path estimates, is therefore a closer match to the historic observed duration for cargo records. Indeed, we find a slightly higher correlation between the model estimates with the observed historic values when restricting the sample to just cargo records. The model's June estimate has a correlation of 0.926 with the observed cargo values, while the December estimate with the cargo values has a correlation of 0.925 (Table 1, Column (8)).

We further visually depict the relationship between each of the model estimates and the observed historic values in a scatter plot (Figure 7). In each scatter plot, we include the 45-degree line, which indicates the line of equality. Points that are closer to the 45-degree line are more equal in value. By comparing the distribution of the data points to the 45-degree line, we can visually assess the relationship strength between the model estimates and the observed historic values. Panel (a) in Figure 7 shows the scatter plot between the model's June estimates and the historic observed values while Panel (b) in Figure 7 shows the scatter plot between the model's December estimates and the historic observed values. While there are some data points that are further

[4]While both correlation coefficients have the same value at the third decimal place, they are different at the fourth decimal place. The model's June estimate has a correlation of 0.9102 with the observed values, while the December estimate has a correlation of 0.9098 (Figure 7).

away from the 45-degree line, most of the data points are distributed along it. Delving into some of these outliers, the point that is furthest to the right of the 45-degree line with a very long historic observed day of almost 150 days is a route from London UK to Norfolk US which was transporting passengers. Since passenger transport sometimes makes multiple stops, this can explain why this historic records have a much longer duration. This further contributes to the higher correlation we obtain when restricting our records to just cargo records (Table 1, Column (8)).



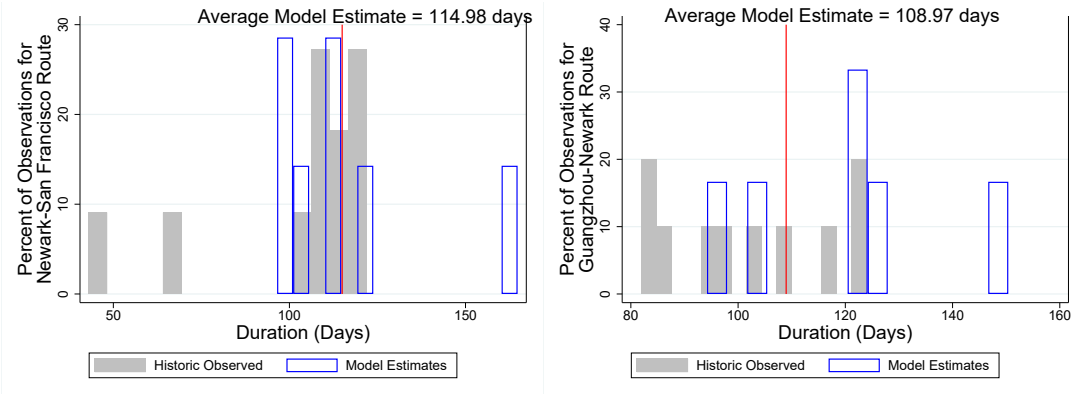
(a) Jun Estimated and Historic Observed Durations (b) Dec Estimated and Historic Observed Durations

Figure 7. Scatter Plot between Estimated and Observed Durations

***Notes*:** There are 102 historic observed route duration that are collected from historic records (see section 4.1 for further information). The model's June estimates are the average of the model's estimates using 7 days of wind data in June 1940 (June 1^{st}, 5^{th}, 10^{th}, 15^{th}, 20^{th}, 25^{th}, and 30^{th}). The model's December estimates are the average of the model's estimates using 7 days of wind data in December 1940 (December 1^{st}, 5^{th}, 10^{th}, 15^{th}, 20^{th}, 25^{th}, and 30^{th}). Panel (a) reports a scatter plot of the model's June estimates against the observed duration and the red line indicates the 45 degree line. Panel (b) reports a scatter plot of the model's December estimates against the observed duration and the red line indicates the 45-degree line. All values are logged.

There are multiple entries in our historical records that pertain to the same route. By isolating these routes with the highest number of entries and comparing the distribution of their entries to distribution of our estimates for each of the 7 days per month, we can perform an additional validation exercise–to assess how well our estimates align with historical records within each route. We focus on the two routes where we have the highest number of historic observations. The first route is Newark to San Francisco with 11 historic observations. Figure 8a shows the distribution for all 11 historic records and their durations (represented by the gray bars). Our average model estimate for December is 114.98 days and is plotted in red in this histogram. Despite the variation in the historic records, our average model estimate falls within a 15-day range of the majority of these estimates, covering over 81 percent of the data points (9 out of 11 observations). On top of this, we include each of our model estimates for the 7 days in December to illustrate their distribution (represented by the blue bars, Figure 8a). We find that our day-level model estimates generally fall within the range of the historically observed durations. To highlight a second example, we have 10 observations for the Guangzhou to Newark route. Figure 8b highlights the distribution of all 10 historic records and their durations (gray bars). Our average June model estimate is 108.97 days and is plotted in red in this histogram. Although there is some variability in the historic observed durations, our average model estimate is within a 15-day range of 70 percent of these estimates (7 out of 10 observations). Additionally, the estimates for the 7 days in June also fall within the range of the

historic observed durations (blue bars, Figure 8b).



(a) Newark to San Francisco

(b) Guangzhou to Newark

Figure 8. Distribution of Estimated and Observed Durations for Particular Routes

***Notes*:** Panel (a) shows the distribution of the 11 observed historic records and 7 model estimates for the route from Newark to San Francisco using December wind data. Panel (b) shows the distribution of the 10 observed historic records and 7 model estimates for the route from Guangzhou to Newark using June wind data. The red line in both panels show the average model estimate for each.

Last but not least, we can conduct an additional validity check on the implied speed of our estimates. Since our model generates routes and therefore distance estimates, we can combine it with the historic observed duration of these routes to calculated an implied average speed—dividing the estimated distance by historic observed duration. The 14 estimates (7 days in both June and December months) have implied speeds that average that range between 5-11 knots. This range covers the historic observed ranges for which data and logbooks are available: 9-12 knots for *Kyrenia II*'s return voyage to Greece (Beresford, 2012), and 4.6-6 knots for a subset of Mediterranean voyages (Casson, 1995).

5. Discussion

The results of the model validation show that the workflow provides reasonable estimates of historic sailing times. In the original TRANSIT model, modeled times were found to be slightly inflated relative to the historic observations (Alberti, 2018). As can be seen in Figure 6, while many individual factors could influence the journey of any particular sailing vessel, the results of the Global-TRANSIT workflow can be considered reasonable approximations of the historically observed travel times.

Additionally, the success of the modeled geographic routes can be seen in Figure 9. The routes in black display historic routes from the CLIWOC dataset, described in Section 2.3, which traveled between London and Cape Town. The green routes display the modeled outputs of the Global-TRANSIT workflow. Overall, the general shapes and variation of the wind-driven modelled routes approximate the shape and variation of the historic sailing routes. In both, there is significant variation depending on the selected day. Notably, our routes tend to travel slightly further south and west than the historic routes, influenced by the strength of the wind on the days modelled. Of note, the CLIWOC dataset is limited to Dutch, English, French and Spanish ships. Brazil

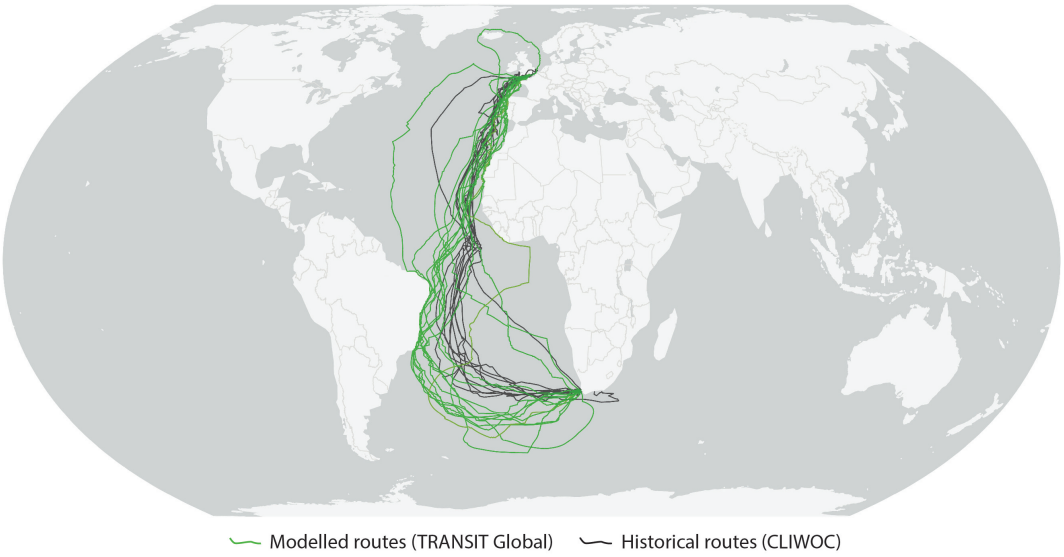


Figure 9. Map of model-estimated compared to observed historic routes - Southampton to Cape Town
***Notes*:** For more information on CLIWOC data, refer to Section 2.3.

was a Portuguese colony at the time, and therefore, this historic record is biased towards voyages which would not have stopped in Brazil. Overall, this comparison demonstrates a good agreement with both the overall trends in sailing navigation and how geopolitical considerations can alter the transit of global maritime shipping.

A single modeled route shows an interesting deviation from the historic data. Note the route which goes around the United Kingdom in Figure 9. This highlights a limitation of the least cost path algorithm. Given particularly non-optimal wind conditions to enter the English Chanel, the optimal Path as Line tool, thus model generates a substantially longer, but compatible route the "long way around". In reality, a ship might instead wait a day or two, at port or at sea. In that time the storm may have passed and the geographic route would continue along the expected shorter path. This is a known limitation of the model which is why we do not rely on the any single route alone, but even with this limitation our model duration estimates generally have a high correlation with the historic validation set (Table 1). Additionally, we reduce the impact of these outliers by evaluating our model estimates over several days across June and December and taking the average of the completed days. This is what Gal, Saaroni, and Cvikel (2023) directly quantify what they describe as "waiting days" in their model, and an alternate method could be introducing a moving temporal window to the model, as Perttola (2022) did, although this introduces substantial processing costs.[5]

We further highlight three main reasons for the discrepancies between our model estimates and the observed historical records: (1) passenger transport in the historical records, (2) multiple historic entries for the same route, and (3) extraordinary cases like speed records, war routes, as well as exploratory routes. First, the discrepancies

[5]In an separate analysis, port and at-sea stops as well as their stop duration could be approximated, and then the total travel times can be modified accordingly. However, that analysis is outside the scope of this project.

observed can be attributed to historic routes dedicated solely to passenger transportation. Our historical data includes eight observations that exclusively transported passengers. One example of this route, London to Norfolk, was highlighted earlier. Additional examples include routes from Guangzhou to Sydney, Gibraltar to Nassau, and Newark to Liverpool. Our model predicted shorter sail times than the historic records for these passenger-only routes. This is in part due to passenger ships often operating at a more leisurely pace and making multiple stops to allow passengers to visit various locations. Consequently, the historical records of journeys involving passengers may reflect less direct and more circuitous routes. This tendency towards longer and more scenic voyages likely contributes to the differences observed between our model's estimates and the actual historical data. Indeed, we find a slightly higher correlation between our estimates with the observed historic values when restricting the sample to just records that transport cargo exclusively (our baseline correlations are 0.909 and 0.906 for both June and December respectively, see Table 1 Column (7)). When restricting the sample to just cargo-transporting routes, the model's June estimate has a correlation of 0.926, while the December estimate has a correlation of 0.925 (Table 1, Column (8)). This higher correlation suggests that the model more accurately captures the nature of cargo transport, where routes are typically more direct and optimized for efficiency.

Second, there are multiple entries in our historical records that pertain to the same route. For these routes, our model generates the same sail time estimate since our estimate takes the average of 7 days at 5-day intervals for each month–June and December. The entries for these routes can be seen visually as horizontal dots in both panels in Figure 7, and contribute to a lower correlation overall between our model estimates and the historic durations (the dots are horizontal since our estimates are plotted on the y-axis). The top two routes with the highest number of historic observations have 11 and 10 observations each–Newark to San Francisco and Guangzhou to Newark respectively). When restricting our sample to routes that have 2 entries or less (thereby removing routes with many repeated entries that include these two routes), our June correlation improves slightly to 0.911 while our December correlation goes up to 0.908. We also utilize these routes for an additional validity check (Figure 8).

Third, our historic data includes some extraordinary cases like speed records and war routes, as well as exploratory routes. The routes that set out to break speed records and routes that set sail for war are likely to be much faster than our estimates, while routes that are exploratory are likely to be much slower. As examples, the races to transport tea in the Great Tea Race of 1866 (Dash, 2011) include the route from Fuzhou to London which took 97 days while in comparison our model estimates were 11-14 percent longer in duration, as well as Guangzhou to Bristol which took 99 days while in comparison our model estimates were 17-21 percent longer in duration. On the other hand, exploratory routes take longer. As an example, the exploratory route from Bristol to Halifax took 33 days (Cartwright, 2020) while our model estimates are 48-52 percent shorter. When restricting our sample to routes that are not speed records, war-driven, or exploratory, our model estimates have a higher correlation with the historic observed routes: the June estimates have a higher correlation of 0.944 while the December estimates have a higher correlation of 0.943.

Overall, if we limit our sample to routes with few repeated observations and by excluding extraordinary routes, our overall correlation coefficients improve by much more to 0.9622 for the June estimates, and to 0.9571 for the December estimates. This improvement indicates that the discrepancies between the model's estimates and historical records are largely due to the factors we outlined earlier.

Despite the known and expected limitations of modeling such a complicated phenomenon as sailing, the Global-TRANSIT workflow proves useful as a method of approximating historic sail movement through space and allowing researchers to comparatively investigate the impact of changing input parameters on shipping routes. One promising new measurement from the updates spatial analyst tools, for example, is the addition of optional input "barrier" data to the Distance Accumulation tool. While we do not integrate this feature into our validation, this opens opportunity for future researchers to investigate the impact of changing oceanic barriers. Additionally, as mentioned earlier, future researchers could evaluate different approximations of the horizontal factor table, which could approximate the movement of different vessel types. Since the publication of the TRANSIT model, for example, the model has been used in conjunction with other records to approximate the impact of factors such as seasonality on seafaring (Gheorghiade & Spencer, 2024). The Global-TRANSIT model expands the opportunity to model and approximate changing factors in maritime voyage patterns globally.

6. Conclusion

This paper provides a stable and replicable methodology for modeling historic sailing times. Global-TRANSIT is a comprehensive workflow that uses cost-surface analysis to create travel time estimates for multiple origin-destination pairs globally. This workflow builds on and extends the TRANSIT workflow (Alberti, 2018) by increasing the spatial and relational scalabiliy of the model and updating deprecated tools that the previous TRANSIT model was built on. We maintain the ability to accommodate varying ship characteristics and add the ability to run Global-TRANSIT within a larger workflow that accounts for projection limitations and supports spatial constraints (such as the opening of international waterways) as inputs. These inputs can be adjusted to evaluate their overall impact on global maritime shipping patterns. Our goal with providing this workflow is to create a replicable tool for researchers in the social sciences to assess the relative accessibility of trade ports at a global scale.

For demonstration, the model was first compared against historically observed travel times. Our modeled data falls within the distributions of known historic travel times and we find a high correlation between them. Additionally, we show that the implied speeds from our modeled data match historic sail speeds and our modeled geographic routes are similar to historically observed routes. The outputs of Global-TRANSIT provide an approximation of the likely duration and route of sailing journeys between origin and destination ports worldwide. These findings can serve as a valuable reference for understanding historic patterns of sail voyage globally, and as a benchmark for assessing the evolution of maritime shipping over time.

Data Availability Statement

Our paper presents the Global-TRANSIT workflow, a least-cost surface raster analysis that uses wind speed and direction to estimate sailing routes and durations for ports globally. Our workflow is openly available as Python scripts for ArcGIS Pro in a public repository at https://figshare.com/s/8517eb49981f658df6ff.

References

Alberti, G. (2018). Transit: a gis toolbox for estimating the duration of ancient sail-powered navigation. *Cartography and Geographic Information Science*, *45*(6), 510–528.

Albion, R. G. (1938). Square-riggers on schedule: The new york sailing packets to england, france, and the cotton ports.

Antikainen, H. (2013, February). Comparison of Different Strategies for Determining Raster-Based Least-Cost Paths with a Minimum Amount of Distortion: Determining Raster-Based Least-Cost Paths. *Transactions in GIS*, *17*(1), 96–108. Retrieved 2023-04-13, from `https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2012.01355.x`

Austin, D. (2021). *The history of the world's first cruise ship built solely for luxurious travel.* https://www.smithsonianmag.com/history/history-worlds-first-cruise-ship-built-solely-luxurious-travel-180978254/. (Smithsonian Magazine, Accessed on April 6, 2024)

Barboza, R. (2017). *Fairhaven's famous clipper ship skipper, captain alexander winsor (1810 – 1890).* https://www.southcoasttoday.com/story/news/local/advocate/2017/06/30/fairhaven-x2019-s-famous-clipper/20399566007/. The Standard Times.

Beresford, J. (2012). *The ancient sailing season* (Vol. 351). Brill.

Braga, J. (1955). *China landfall, 1513 : Jorge Alvares' voyage to China, a compilation of some relevant material.* Macau: Imprensa Nacional. https://nla.gov.au/nla.obj-239881932.

Calmon, P. (2024, February). *Pedro Álvares cabral.* Retrieved 2024-04-12, from `https://www.britannica.com/biography/Pedro-Alvares-Cabral` (Encyclopedia Britannica, Accessed on August 6, 2024)

Cartwright, M. (2020). *John Cabot.* Retrieved 2024-04-12, from `https://www.worldhistory.org/John_Cabot/` (World History Encyclopedia, Accessed on August 6, 2024)

Casson, L. (1995). *Ships and seamanship in the ancient world.* JHU Press.

Chichester, F. (1967). *Along the clipper way* (1st American ed.]. ed.). New York: Coward McCann.

CLIWOC. (n.d.). Retrieved 2023-06-23, from `https://www.historicalclimatology.com/cliwoc.html`

Conolly, J., & Lake, M. (2006). Section 11: Routes: networks, cost paths, and hydrology. In *Geographic Information Systems in Archaeology.* Cambridge, UK: Cambridge University Press.

Dash, M. (2011, December). *The Great Tea Race of 1866.* Retrieved 2024-04-06, from `https://www.smithsonianmag.com/history/the-great-tea-race-of-1866-8209465/` (Section: History, World History, , Blogs, , Past Imperfect, , Articles)

Duke University Digital Collections. (n.d.). *Brief history of the passenger ship industry.* https://blogs.library.duke.edu/digital-collections/adaccess/guide/transportation/passenger-ships/. (Duke University Library Digital Collections, Research and text by Lydia Boyd, Accessed on April 6, 2024)

Dunkley, M., & Stamper, P. (2016, July). *Ships and Boats: 1840-1950* (Tech. Rep.). Greak Britain: Historic England.

Esri. (2024). *Esri product lifecycle support policy.*

https://downloads2.esri.com/support/TechArticles/Product-Life-Cycle.pdf.

Esri. (ArcGIS Pro 3.3a). *Arcgis pro parallel processing factor (environment setting).* https://pro.arcgis.com/en/pro-app/latest/tool-reference/environment-settings/parallel-processing-factor.htm. (ArcGIS Pro 3.3 Tool Reference, Accessed on 2024-04-10)

Esri. (ArcGIS Pro 3.3b). *Distance accumulation algorithm.* https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/distance-accumulation-algorithm.htm. (ArcGIS Pro 3.3 Distance Toolset Concepts, Accessed on 2024-04)

Esri. (ArcGIS Pro 3.3c). *Distance accumulation function.* https://pro.arcgis.com/en/pro-app/latest/help/analysis/raster-functions/distance-accumulation-global-function.htm. (ArcGIS Pro 3.3 Distance Toolset Concepts, Accessed on 2024-04)

Esri. (ArcGIS Pro 3.3d). *Iterators.* https://pro.arcgis.com/en/pro-app/latest/help/analysis/geoprocessing/modelbuilder/iterators-for-looping.htm. (ArcGIS Pro 3.3, Accessed on 2024-08-15)

Esri. (ArcMap 10.8a). *How the horizontal and vertical factors affect path distance.* Retrieved from `https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/how-the-horizonal-and-vertical-factors-affect-path-distance.htm` (ArcMap 10.8, accessed in 2024-04-10)

Esri. (ArcMap 10.8b). *How the path distance tools work.* Retrieved 2023-01-09, from `https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/how-the-path-distance-tools-work.htm#ESRI_SECTION1_864C70F88FEC40BAACD8ECB6CC1EEADB` (ArcMap 10.8, accessed in August 2024)

Fitzgerald, S. (1997). *Red tape, gold scissors: The story of sydney's chinese.* State Library of New South Wales Press. Retrieved from `https://books.google.com/books?id=LrZJAAAACAAJ`

Fitzpatrick, S. M., & Callaghan, R. (2008, June). Seafaring simulations and the origin of prehistoric settlers to Madagascar. In G. Clark, F. Leach, & S. O'Connor (Eds.), *Islands of Inquiry: Colonisation, seafaring and the archaeology of maritime landscapes* (1st ed.). ANU Press. Retrieved 2023-05-03, from `http://press-files.anu.edu.au/downloads/press/p26551/pdf/ch0318.pdf`

Gal, D., Saaroni, H., & Cvikel, D. (2023, June). Mappings of Potential Sailing Mobility in the Mediterranean During Antiquity. *Journal of Archaeological Method and Theory*, *30*(2), 397–448. Retrieved 2024-03-19, from `https://link.springer.com/10.1007/s10816-022-09567-5`

Gheorghiade, P., & Spencer, C. (2024, January). Modelling the Cost of the Wind: A Preliminary Reassessment of Networks of Mobility in the Late Bronze Age Mediterranean. *Journal of Computer Applications in Archaeology*, *7*(1), 36–53. Retrieved 2024-02-26, from `http://journal.caa-international.org/articles/10.5334/jcaa.119/`

Gumport, R. K., & Smith, M. M. (2006). *The china trade, 1830 to 1860.* University of Illinois at Urbana-Champaign. Retrieved from `http://teachingresources.atlas.illinois.edu/chinatrade/resources/resource1_7.pdf`

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., ... Thépaut, J.-N. (2018). *ERA5 hourly data on single levels from 1940 to present.* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Retrieved from `https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.adbb2d47?tab=overview` (last accessed December 11, 2023)

Herzog, I. (2013a). Least-cost Networks. In G. Earl et al. (Eds.), *Archaeology in the Digital Era* (pp. 237–248). Amsterdam University Press. Retrieved 2023-06-12, from http://www.jstor.org/stable/j.ctt6wp7kg.28

Herzog, I. (2013b, August). The Potential and Limits of Optimal Path Analysis. *Computational Approaches to Archaeological Spaces*, *60*.

Howe, O. T. (1986). *American clipper ships, 1833-1858*. New York : Dover. Retrieved 2024-03-20, from http://archive.org/details/americanclippers0000howe

Kingsley, M. H. (2020). *Travels in west africa (congo francaise, corisco and cameroons)*. BoD–Books on Demand.

Kraus, H. (n.d.). *The Cadiz Raid.* Retrieved 2024-04-08, from https://www.loc.gov/collections/sir-francis-drake/articles-and-essays/drake-biography/the-cadiz-raid/ (Library of Congress)

Leon-Guerrero, J. (2024, November). *Navigation and Cargo of the Manila Galleons.* Retrieved 2024-04-12, from https://www.guampedia.com/navigation-and-cargo-of-the-manila-galleons/ (Guampedia Foundation and University of Guam, Accessed on August 6, 2024)

MacGregor, D. R. (1983). *The tea clippers: their history and development, 1833-1875*. Conway Maritime Press Limited.

Marks, A. A. (n.d.). *Jamestown Questions and Answers.* Jamestown-Yorktown Foundation. Retrieved 2024-04-08, from https://web.archive.org/web/20220623062701/https://jyfmuseums.org/pdf/Curriculum-Materials/JamestownQuestionsandAnswers.pdf (Jamestown-Yorktown Foundation, Accessed on August 6, 2024)

McGrail, S. (2009). *Boats of the World: from the Stone Age to Medieval Times*. Oxford [U.K.]: Oxford University Press.

Mitchell, A. (2012). *The Esri guide to GIS analysis. Volume 3, Modeling suitability, movement, and interaction*. Redlands, California: Esri Press.

Murray, W. M. (1987, December). Do modern winds equal ancient winds? *Mediterranean Historical Review*, *2*(2), 139–167. Retrieved 2023-06-14, from http://www.tandfonline.com/doi/abs/10.1080/09518968708569525

Nantucket Historical Association. (n.d.). *Log of the brig clio, 1830.* https://fromthepage.com/nharl/logs/ms220-log390?page=21. Ships' Logs Collection. (Accessed on August 6, 2024)

New York Times. (1861). *From hong kong to san francisco: Voyage of the mary whitridge.* https://www.nytimes.com/1861/04/21/archives/from-hongkong-to-sanfrancisco-voyage-of-the-mary-whitridge-with-a-c.html. (Accessed on August 6, 2024)

Pascali, L. (2017). The wind of change: Maritime technology, trade, and economic development. *American Economic Review*, *107*(9), 2821–54.

Perttola, W. (2022, June). Digital Navigator on the Seas of the Selden Map of China: Sequential Least-Cost Path Analysis Using Dynamic Wind Data. *Journal of Archaeological Method and Theory*, *29*(2), 688–721. Retrieved 2023-03-06, from https://link.springer.com/10.1007/s10816-021-09534-6

Piccotti, T. (2023). *Christopher columbus.* https://www.biography.com/explorer/christopher-columbus: :text=On Biography.com.

Royal Museums Greenwich. (2018, November). *18th century sailing times between the English Channel and the Coast of America: How long did it take?* Retrieved 2024-04-06, from https://www.rmg.co.uk/stories/blog/library-archive/18th-century-sailing-times-between-english-channel-coast-america-how

Sandström, F. (2000). *Data of the ship arrow, 1902.* https://web.archive.org/web/20110427100736/http:/sailing-ships.oktett.net/4.html. Database of Historic Sailing Ships. (Accessed on August 6, 2024)

Savage, S. H. (1990). Modelling the Late Archaic social landscapes. In K. M. Allen, S. W. Green, & E. B. Zubrow (Eds.), *Interpreting Space: GIS and Archaeology* (pp. 330–355). Bristol, PA: Taylor & Francis.

Scherjon, F. (2013). Stepping in-modern humans moving into europe-implementation. In *Proceedings of the 40th conference on computer applications and quantitative methods in archaeology southampton, 26–30 march 2012* (pp. 105–117).

Sethian, J. A., et al. (1999). *Level set methods and fast marching methods* (Vol. 98) (No. 2). Cambridge Cambridge UP.

Sobel, D. (2005). *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time.* Macmillan.

South Carolina Historical Society. (1685). *Tales of their travels: A glimpse at colonial migration and movement in south carolina.* https://schistory.org/exhibit/tales-of-their-travels/. (Letter by Robert Quarry, Accessed on August 6, 2024)

Stark, W. F. (2009). *The last time around cape horn: The historic 1949 voyage of the windjammer pamir.* Hachette UK.

TenBrink, J. (2019). *What's new in the spatial analyst distance toolset in pro 2.5.* Retrieved 2024-04-06, from `https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/whats-new-in-the-spatial-analyst-distance-toolset-in-pro-25/`

The Maritime Heritage Project. (n.d.). *Merchant ships in port.* https://www.maritimeheritage.org/inport/1849.htm. (Accessed on August 6, 2024)

The Maritime Heritage Project: SS China. (n.d.). *Ss china passenger lists: San francisco 1800s.* https://www.maritimeheritage.org/passengers/SS-China.html. (Accessed on August 6, 2024)

White, R. C. (2016). *American Ulysses: A Life of Ulysses S. Grant.* Random House Publishing Group. (Google-Books-ID: TPNRCwAAQBAJ)

Whitewright, J. (2011, March). The Potential Performance of Ancient Mediterranean Sailing Rigs. *International Journal of Nautical Archaeology*, *40*(1), 2–17. Retrieved 2023-06-14, from `https://onlinelibrary.wiley.com/doi/10.1111/j.1095-9270.2010.00276.x`

Williamson, W. (1846). *Journal of the ship recorder bound to the cape of good hope with emigrants.* https://www.geni.com/projects/British-Ships-to-South-Africa-in-the-1800-s-Recorder/38414. British Ships to South Africa in the 1800's - Recorder. (Accessed on August 6, 2024)

Woods, J. D. (1985). The World Ocean Circulation Experiment. *Nature*, *314*, 501–511.

Zhao, H. (2005). A fast sweeping method for eikonal equations. *Mathematics of computation*, *74*(250), 603–627.

Appendix

Table A.1. Sources of Historic Observed Routes

Sources (1)	Number of Routes (2)
Gumport and Smith (2006)	46
The Maritime Heritage Project	22
Albion (1938)	5
Chichester (1967)	4
Williamson (1846),Leon-Guerrero (2024), The Maritime Heritage Project: SS China	2 each
South Carolina Historical Society (1685), New York Times (1861), Braga (1955), MacGregor (1983), Fitzgerald (1997), Sandström (2000), Nantucket Historical Association, Stark (2009), Kingsley (2020), Marks, Piccotti (2023), Kraus, Calmon (2024) Cartwright (2020) Sobel (2005), Royal Museums Greenwich (2018), Dash (2011), White (2016), Barboza (2017)	1 each
Total of 102 Port-Pair Routes	

***Notes*:** This table lists the sources of the 102 historic observed port-pair routes used in our validation in Section 4. Column (1) describes the sources and Column (2) lists the number of port-pair routes that come from each source.