Assessing ship movements using volunteered geographic information

Shaun Walbridge

December 11, 2012

1 Introduction

1.1 Shipping, who cares?

Shipping, the global transportation of people and goods, provides a useful window into critical issues for geographic information science (GIScience), and its understanding has implications for a variety of applied fields including marine conservation, marine spatial planning, transportation geography, and navigation. This work examines a synthesis of available shipping data.

GIScience seeks to formalize our understanding of geographic information, and shipping can provide insight into areas of the GIScience agenda. Ship observations are contributed by vessels that submit data on both a voluntarily and mandated basis (Freeman, 2010; Tetreault, 2002), affording an opportunity to advance our use of geographic quality checking methods (Goodchild and Li, 2012) against this volunteered geographic information (VGI) (Goodchild, 2007). And while individual observations have a simple representation (a location, time, and attributes), effective use requires multiple representations (Goodchild, 1992) and the spatio-temporal modeling techniques of time geography (Miller and Bridwell, 2008).

Shipping places a growing demand on the ocean, transporting $1.8 trillion dollars of goods annually, and providing a $300 billion dollar maritime industry, representing one of the largest economic users of the ocean. This activity impacts both human health and the environment, such as the $3 billion Exxon-Valdez spill[1], but largely the effects of shipping remain out of sight. Documented effects include greenhouse gasses, accounting for five percent of total man-made emissions, and causing 60,000 premature deaths per year (Corbett *et al.*, 2007). Ecologically, vessels cause ship strikes which can impair whale populations (Fujiwara and Caswell, 2001), transport invasive species, leading to widespread economic and biological losses, cause groundings with a host of direct effects like oil spills and habitat destruction, and produce noise pollution, leading to mortality events. Other ecological and human impacts are likely to exist, but remain unidentified, as shipping continues to be poorly understood (Davenport and Davenport, 2006). As we try to understand the health of the ocean (Halpern *et al.*, 2012), we need detailed data which poses an acute challenge to ocean scientists (Wright and Goodchild, 1997). This data is necessary to inform decision-making, such as marine spatial planning, but also is important to minimizing the regulatory burden on the shipping industry: by managing shipping as a complex system, efficiency can be tied to improved environmental conditions.

[1]The $5 billion liability for the spill was so great that a now infamous financial instrument was manufactured to absorb it: the credit default swap.

Another major, direct economic user of the ocean is fishing, where our understanding has been advanced by intensive investments to study its effects. Many fisheries are now managed systems (Worm *et al.*, 2009), in part due to frequent *tragedy of the commons* events, leading to the collapse of unmanaged fish stocks (Costello *et al.*, 2012). This same management is noticeably absent in the shipping industry.

Shipping is regulated by the International Maritime Organization (IMO), a branch of the United Nations which enforces rules and standards on the shipping industry. However, the transnational nature of the industry has led to low policy enforcement. As is common in consensus-based international bodies (Cogan, 2009), the IMO is slow to respond to problems of the day. For example, while acknowledging that 20% reductions in greenhouse gasses can be accomplished in this decade, without additional costs to operators(International Maritime Organization, 2009), it hasn't implemented new rules on such gasses. In light of the weak regulatory environment, alternatives, such as tying ship insurance rates to specific areas of the ocean, may be preferable.

1.2 Transportation networks

Transportation networks with geographically-fixed edges and nodes, such as road and rail, form the basis of transportation geography (Rodrigue *et al.*, 2009). Unlike these networks, shipping exists in between the extremes of a fixed network and two dimensional brownian bridges, as vessels must transport both goods and passengers to specific locations (primarily ports) but are free in choosing movement between destinations, except in near shore areas regulated by shipping lanes. Air transportation shares network similarities, but strong regulation has led to predetermined flight paths, with deviations generally limited to emergencies or extreme weather.

Because of inherent risks in air transportation, detailed, well-vetted information on flight paths is provided by government agencies (Guimera *et al.*, 2005), simplifying modeling. Ship movements have no equivalent top-down data collection effort, and while organizations such as the US Coast Guard have made efforts, the system remains rudimentary. Most information is provided via private contracts, through organisations such as Lloyd's of London, who have been recording information on shipping since 1774 (Lloyd's Register-Fairplay Ltd., 2010). As a result, limited public data is available on the shipping trade, despite being identified by the FGDC (Federal Geographic Data Committee, 2008) as a key transportation component to the national spatial data infrastructure (Currier *et al.*, 2012).

Transportation networks play an important economic role (Canning and Fay, 1993). Because of the major economic role, information on shipping is valuable, leading to a

cottage industry of information sales and limiting its public availability. Here, we look at using geographically-volunteered information to infer the movement patterns which in part define the shipping network.

1.3 Existing work

Much research has focused on specific areas within shipping, such as maritime awareness to detect anomalous behavior (Tun *et al.*, 2007), tracking ship-whale interactions (Jensen *et al.*, 2004), or shipping lanes to protect a species (Lagueux *et al.*, 2011; McKenna *et al.*, 2012). Few papers, however, tackle shipping as a global phenomena. Two important papers from this limited body of research will now be discussed.

The first work to examine the global shipping trade in an ecological context was a synthesis effort to collect information across habitat types and impacts (Halpern *et al.*, 2008). The work evaluates many human uses of the ocean, and gathers together global data to examine the cumulative impact humans have, ranging from changes in sea surface temperature due to radiative forcing, to land-based nutrient runoff. The data used in this analysis to represent shipping, however, has limitations: it ignores vessel classes, and contains information on only 12% of the fleet. The ships it does include are a spatially- and statistically-biased sample of the population (Wang *et al.*, 2007). The data is variable between ships, so many areas where there is active use are missing from the model. Finally, the paper does not tackle modeling issues, but instead provides patterns without the context necessary to address deeper questions.

A more recent paper (Kaluza *et al.*, 2010) takes a different approach, by focusing on shipping as a network, with an analysis geared toward a network theory audience. Licensing data from Lloyd's of London, the authors aggregate port of call records, which are sequential lists of vessel location. They then extrapolate port to port links into routes, by combining land-based barriers with great circle distance calculations to select the 'path' of a ship between two ports. This novel approach provides a measure of connectivity between port locations, and improves on the gravity model widely used to predict connectivity. However, this model provides insufficient detail to extract ship movement paths, necessary for resolving many of the conflicts listed above. Instead, we need geographically referenced facts about movement. Additionally, the data used requires expensive licensing, making extending the work financially prohibitive. Despite largely ignoring geography, this paper is currently the best examination of the global shipping trade.

1.4 Where this work comes in

Fusing together data from multiple sources, here I build a dataset useful for interpreting the global shipping trade. While high-resolution (spatially and temporally) volunteered information exists, its use has been limited to regional problems, and existing global work used poor data (Corbett *et al.*, 2007; Halpern *et al.*, 2008). This paper contributes data predominantly within 100km of shores, where most human and biological users of the ocean persist, and building our knowledge of these areas is particularly valuable.

As noted by Goodchild, "changing technology and economics are moving map production from a system of unified central production to a local patchwork, and the old radial system of dissemination is being replaced with a complex network."(Goodchild, 1999) By using quality assurance methods from both computer science and geography, such as record linage and geographic validation, we can filter unreliable inputs. The near future will involve global, real-time, high resolution ship data (Jones, 2012; Carson-Jackson, 2012), but we continue to need methods which accommodate data curation and integration, and allow us to address specific hypotheses. This is a key appeal of volunteered information: some problems of uncertainty become tractable with sufficient observation volume, as we can evaluate the distribution directly instead of relying on sampling methods.

2 Methods

2.1 Volunteered vessel information

Historically, ship data was collected both by governments for internal use, and by private corporations with the intent to sell. As elsewhere in the production of geographic facts, a shift is underway which moves the emphasis away from top-down primary data collection, to relying on observations from a multitude of sources (Goodchild, 2007; Elwood *et al.*, 2011). This new information has the potential to fundamentally change our use of the ocean.

Ship captains have long taken climate observations alongside known locations (Brohan *et al.*, 2009). Building on this history, the Voluntary Observing Ship (VOS, Fletcher, 2010) program collected a dataset spanning over 20 years and covering 10–20% of commercial traffic within each year. Ass the intention was to collect open-ocean climate observations, many of the records lack ship attribute information. The data is contained within the greater International Comprehensive Ocean-Atmosphere Data Set (ICOADS, Woodruff *et al.*, 2010), and though the data is both spatially- and statistically- biased (Wang *et al.*, 2007), it serves as a useful training dataset on ship movements in the open ocean. Here, I use VOS records from 2003–2011, supplemented with opportunistic observations collected by the vessel tracking site SailWX (Mueller, 2011). On the vessels that do report both attributes and locations, this information is captured at regular intervals, providing difficult to find open ocean observations.

The Automatic Identification System (AIS) (IALA AISM, 2004; Tetreault, 2002) was developed to improve maritime safety and prevent collisions, by providing mariners local situational awareness (Figure 1). By locating the ships via global positioning satellite (GPS) fix, and broadcasting the location alongside other attributes (Table 5) regularly via VHF transceiver, mariners gain real-time local traffic, invaluable in inclement weather or rescue operations (ITU, 2010). The International Maritime Organization (IMO) mandates that all ships $\geq$150 gross tonnage (GT) or ships bearing paid passengers carry AIS units (International Maritime Organization, 1980), which has lead to approximately 200,000 ships being outfitted with AIS equipment, including all licensed tankers, cargo ships and passenger vessels. Because the intent was to improve maritime safety, the system uses well-understood VHF radios, broadcasting to about 40km between ships. Since the system's implementation, land-based users, including ports, maritime professionals and amateurs, have set up VHF antennae, providing low-cost local ship traffic data at a range of up to 100km. Numerous sites now provide real-time feeds, such as MarineTraffic (Univeristy of the Aegean, 2010) and SailWX, aggregating the records from many land-based antennae (and more recently, satellite feeds) and displaying them over both

web maps and via Google Earth. These sites constantly augment their networks by providing in-depth data to those willing to set up AIS receiving stations in geographic areas not yet covered.

Figure 1: Automated Information System diagram. Ships communicate to one another via VHF broadcast. The same signals can be captured and stored by land- and satellite-based observers.

2.2 Data Collection

For this study, fifteen months of AIS data were collected, from November 2010–December 2011, aggregating records from three major online AIS providers: FleetMon, VesselTracker, and MarineTraffic. All three share Keyhole Markup Language (KML, Open Geospatial Consortium Inc.) files, intended for Google Earth. Examining data availability (Figure 14) showed these providers had considerably different coverage. At ten minute intervals over the study period, I automated downloading these KML files of real-time ship traffic from each of the providers. I then parsed the files to extract each observation within the dataset, wrote a library to normalize differences between the providers, and inserted the results into a spatial database, (PostGIS, Ramsey *et al.*), an extension providing support for OGC simple features (Open Geospatial Consortium Inc., 2010) on top of the PostgreSQL (PostgreSQL Global Development Group, 2012) object-relational database engine). Over the study period, this provided 2.37 billion observations which all include ship location and time, but frequently include additional attributes (Table 5). The VOS/SailWX dataset, consisting of 92.4 million records covering Feburary 1991–September 2011, was provided in a MySQL database dump, which was converted and imported into PostgreSQL.

Augmenting these ship observations, ancillary data was identified to provide validation against the raw observations, which is described below.

2.2.1 Vessel attribute data

Vessel data, tabular lists of ships alongside attributes, was collected from both authoritative and crowd-sourced media (Table 2) from a variety of sources, useful in validating the attributes provided in both VOS and AIS observations. Sources include:

1. Within the United States, the Federal Communications Commission (FCC) regulates the airwaves. In order to track and manage radio licenses, the FCC has developed the Consolidated Database System (CDBS), which includes information on all vessels with registered radios within the US.

2. The International Telecommunications Union (ITU) created the Maritime mobile Access and Retrieval System (MARS) database to provide the maritime community with the most up-to-date data for registered vessels. Thanks to their role as a regulating body, participant states are required to provide up-to-date information. This database is particularly useful as it includes details not available through other public means, such as the vessel owner, and passenger capacity.

3. DigitalSeas provided volunteered vessel attribute information, collected primarily through corrected AIS observations, but has since gone off-line.

4. VesselTracker includes ship tracking, reporting and vessel records, alongside real-time AIS position data. Both their vessel data, and their AIS observations were recorded in this study. The dataset shows particular strength in European waters. Vesseltracker has also developed a ship routing network, but doesn't provide public access to this resource.

2.2.2 Land-sea mask

A high resolution land-sea mask, derived from the Shuttle Radar Topography Mission (SRTM) Water Body Data (SWBD, Slater *et al.*, 2006), classifies the world into either land or sea at three arcsecond resolution (~90m) for much of the world (56° S to 60° N). It was a by-product of the SRTM digital terrain project (Rabus *et al.*, 2003), and has the advantage that the data was collected over a very short period of time, increasing self-consistency.

For the areas beyond that covered by SRTM, the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS, Wessel and Smith, 1996) was used, which

is an amalgamation of publicly available shoreline data. This data is lower resolution than SWBD, but the vast majority of my observations come from within the SRTM study area. The transition zone between these two layers was manually corrected to make a single, consistent, high-resolution land-sea mask at three arc seconds.

2.2.3 Other data

Port databases were collected, containing coordinates and berth details for ports globally. Approximately 5,000 ports were identified from a range of volunteered and authoritative sources. Approximate information on ship movement patterns was also collected, based on historical charts such as a CIA vessel movement chart from the Cold War (Figure 15). The original ship model produced as part of the previous modeling effort (Halpern *et al.*, 2008) was helpful for comparison.

2.3 Validation

The raw observations are full of caveats: because of inherent limitations in protocol design, there is no direct way of validating an incoming packet (Raymond and Schwehr, 2012). So, as seen in (Figure 11), many locations in the middle of major landmasses have observations, including a particularly thick band centered around the prime meridian. These records are more likely due to corruption of the longitude coordinate than reverence for Greenwich. Alongside transmission errors are operator error: the attributes which are sent alongside the time and position information are input by the mariners, and they may introduce errors in entry, or fail to update the attributes which change over time, such as the destination field. Each observation within this dataset is suspect, and the data should be treated as guilty until proven innocent.

While there are numerous problems with the data, the volume of data, and the compiled ancillary datasets, provide ways to use the data via cross-correlation. This improves accuracy and minimizes the observations required to build a model representation of the data. Two areas relevant to this problem are geographic data mining (Miller and Han, 2009), and the recent work of Goodchild and Li (2012). Here, I borrow the framework described in the latter work, and explore three avenues of quality assurance: crowd sourcing, social, and geographic approaches.

2.3.1 Crowd-sourcing

While Goodchild and Li (2012) found that crowd-sourcing was commonly inadequate for VGI, it can function when the domain is limited and the pool of expertise is vast. In

one sense, crowd-sourcing becomes useful here when radio towers of differing quality receive records. Cross-referencing these sources then provides and additionaly layer of consistency, and although this doesn't rely on citizen participation, it does minimize instrument error.

2.3.2 Social

Mariners provide regular updates to online services, so attribute data from these sources tends to be high. The ship operators frequently have the best working knowledge of a ships' vitals, much like someone who lives in a specific neighborhood is likely to have a better understanding of local geography. The ship operators can then communicate the data up to various shipping related aggregation sites, who operate the higher levels of the hierarchy, and rely on a group of trusted users for vetting incoming updates.

2.3.3 Geographic

This brings us to perhaps the most important validation technique: using geography to validate the records. Here, the individual observations are points with time, and I can rely on a few simple tests. In addition to our point location $\langle x, t \rangle$, have multiple attributes $\langle z \rangle$ provide us a spatio-temporal observation $\langle x, t, z \rangle$. By cross-referencing the attributes, the probability of each attribute-only component can be computed, and from that, infer a likelihood on geometry and time.

I also impose basic validation on the geometries by referencing other geographic facts, as is used in the most pedestrian of GIS functions, the overlay analysis. By checking the data against a land-sea mask, estimations are created as to when the provided location is a physical impossibility. One trick to this is that many ships do travel by river, so these rules must be careful to define what is a traversable area. Additionally, shallow water bodies impose additional constraints on many ships with positive draft. This could be validated (showing, for example, the distance oil tankers keep from land), but has not been implemented here.

For many classes of vessels, ships move between ports. Ships exhibit movement patterns inconsistent with this goal are suspect. However, there are other nodes which require inclusion, such as ballast water exchange points, like one located 100nm offshore of California (Figure 13), and canals, which provide a means of ships to move between otherwise distant locations. However, this remains a powerful validation technique: because the high resolution data is clustered around the shores with strong coverage at the ports, the data contains a disproportionate look at the vessels in transit between this constrained set of locations.

As in many large datasets, the distribution of observations per ship follows an approximate power-law distribution. Using the raw number of observations received in our 15 month window provides another filter. The peak in the kernel density estimation (Figure 2) is seen around 10^4 observations, with a clear drop-off after 10^5 which is consistent with the theoretical maximum (one observation in every sample) of about 6.8×10^5.

Figure 2: Number of observations per vessel, log scaled.

Another important point is that while ship attributes provide us some details as to their nature, categories (such as tanker or cargo) are not based purely on attributes, but also include movement patterns. This is a distinguishing characteristic between the data, as different classes of vessel show distinct movement patterns, which is useful both in validation, and as an output product, since vessel classes have differing effects and interactions. Because this data is stored in a spatial database, I link derived representation back to the records that formed it. While this is a common feature of spatial databases, it is an important trait since data are often provided as simple 'end products' without the provenance to understand its limitations.

2.4 Data Representations

After filtering and validating the data as described, I build data representations which bring us closer to answering some of the problems raised in the introduction. Frequently there is no single optimal representation of data, but instead a set of representations which, when matched to particular uses, can be combined to provide insight.

Here, I look at how maintaining both discrete object and continuous fields representations (Figure 3) allows us to address questions about shipping, including its ecological effects. While the point data alone is too simple for making predictions related to these phenomena, incorporating too much complexity risks making computation infeasible (De Smith *et al.*, 2007).

Figure 3: GIS abstractions, this work's specific additions in blue. Adopted from Bivand (2011)

2.5 Ship Types

An open problem within the maritime community is: what designates a ship? Ships can be classified on a variety of dimensions, including propulsion form, hull material, function and many others. Ontologies have been in development to solve this problem (Vries *et al.*, 2009), but here I focus on use: what is the primary activity the vessel is engaged in? Starting with the initial classes provided by our data sources, I collapsed them into nine major functional groups (Table 1). Because the full attributes of each vessel are retained, and frequently include multiple type labels, it is possible to break this down further for future analyses, but these broad classes served well for classifying distinct movement patterns.

Identifiers – what ship is this – similarly poses a problem. By incorporating information from many different identifiers and focusing on those less changeable, such as IMO number and radio callsign, we increase the chances of valid matches. But additional attributes such as the Maritime Mobile Service Identity (MMSI), and vessel name are often equally useful, particularly because they are broadcast alongside each AIS record. Vessel operators may fail to maintain this information, so it becomes important to cross validate the attributes to gain reasonable match rates.

Type	Vessels *(AIS)*	Vessels *(VOS)*	fleet size	coverage (%)	observations (M)
Cargo	25214	5838	33392[1]	75.6	665.45
Tanker	9758	2375	14068[1]	69.4	264.42
Passenger	4007	777	6370[1]	62.9	142.16
Support	9954	735	25234[1]	39.4	298.02
High-speed	404	81	1178	34.3	2.52
Fishing	11186	349	51200[3]	21.8	6.87
Pleasure	20727	661	800,000[2]	2.59	267.48
Other	9507	1400	–	–	4.75
Authority	656	55	–	–	1.44

Table 1: Summary of analyzed ship data

1. Equasis (2012).
2. Westwood *et al.* (2001).
3. FAO (2012).

2.5.1 Raw point data

Records are sampled at 600 second intervals (10 minutes). For most ship classes, this sample rate retains the movement patterns, and is a similar sample rate to studies using aggregation to remove noise, like (Vries *et al.*, 2009), which uses piecewise aggregate approximation with w = 30 (300 seconds) to "strike a good balance between capturing the general movement and ignoring noise".

The spatial databases' historical focus was on accuracy (Goodchild and Gopal, 1989). This showed the value in retaining raw observations, both by minimizing lossy transformations, and providing multiple models of the same data to retain consistency with reality.

2.5.2 Tracks

Tracks convert discretely sampled time points into a interpolated continuous phenomena. The uncertainty surrounding the interpolation can be described by a time prism (Figure 4), which bounds the area based on the maximum velocity of the object at motion. Here, I first constrain the track segments to have movement above 0.5 Kt/hr, below which vessel control is difficult, and an upper bound of 40 Kt/hr, derived empirically from our ship speed distributions (See Figure 8). Next I require observations have coordinates bounded by the coordinate space, and exclude any beyond the surface of the earth. I further require that to be considered a track, the above criteria must be met for at least one segment of that vessels transit. The ships are then constrained to move along the shortest distance on a geoid, or great circles, using GeographicLib (Karney, 2012). This information is reused in computing the speed along each segment.

Figure 4: A simple space–time prism, as in (Miller, 2005); original from (Wu and Miller, 2002)

2.5.3 Field-based models

Initially, kernel density estimation was used to produce field based estimates. However, based on the data volume, it proved more useful to provide models directly from discretized tracks, which are used in the two following datasets.

Ship density by type The primary output of this work is a field-based density model of ship movement. This view is useful in a wide variety of contexts, from extercises in marine spatial planning (MSP) to detecting conflict zones between resource users, and the simple density estimation in Halpern *et al.* (2008) remains a highly requested product.

Each vessel track was rasterized to both an 90 arcsecond grid (~5.5km at the equator) and an equal area grid in the Hobo Dyer projection (Figure 5). The latter case insures that the density function is computed on grid cells representing the same area for each cell, unlike the geographic grid where area varies by latitude. A vessel is counted only once for each cell it passes through, as the focus here os on overall movement patterns, and this criteria helps de-emphasize vessels with limited movement. Each raster vessel track was combined using simple map algebra to produce density maps for both the AIS and VOS data, for each of our vessel classes.

For each cell the output density value s calculated as a standardized equal-weighted addition between the two input datasets, AIS and VOS:

$$s = \frac{R_{AIS}}{max(R_{AIS})} + \frac{R_{VOS}}{max(R_{VOS})} \tag{1}$$

Speed density estimations of ships by ship type Ship speed plays a critical role in determining the survivability of collisions for many marine species (Vanderlaan *et al.*, 2009). Speed also directly relates to the emissions profile of a vessel, and it has been shown that speed reduction alone can reduce 50-80% of greenhouse gas emissions Lack *et al.* (2011). Conversely, decreased speeds require more vessels to ship the same volume of goods or passengers, though companies such as Maersk are mitigating this by moving to significantly larger container ships.

Average speed per cell was calculated as sum speed over all observations $R_{AIS} \cup R_{VOS} = n$, and dividing it by the total number of observations, but only for those locations where a sufficient density s is present:

$$\bar{s} = \begin{cases} \frac{\sum_{i=0}^{n} s}{x} & s \geq 10 \\ 0 & s < 10 \end{cases} \tag{2}$$

Figure 5: Cargo density grid, generated by combining AIS and VOS records.

2.6 Record Linkage

By combining authoritative data from a variety of sources (Table 2), we can reconcile our observations, greatly improving the quality of resulting ship movement models. Though we do include two authoritative datasets, the sources are inconsistent, and require an initial step of cross-linking records. This approach was initially developed with medical records, and more recently developed as the record linkage field in computer science (Christen, 2012).

Source	Code	Records	Cross-linked	Attributes
Digital Seas	DS	212166	68002	name, IMO, MMSI, callsign, type, width, length
FCC[1] ULS[2]	FCC	319964	24531	name, MMSI, callsign, class, gross gonnage, length
ITU[3] MARS[4]	ITU	372183	75928	name, IMO, MMSI, callsign, class, owner, gross tonnage
VesselTracker	VT	126534	83372	name, IMO, MMSI, callsign, class, length

Table 2: Ship data sources.
1. Federal Communications Commission
2. Universal Licensing System
3. International Telecommunication Union
4. Maritime mobile Access and Retrieval System

I built a probabilistic model which evaluated all possible pairwise combinations be-

tween source records. By using the methodology of record linkage, a set of rules was developed to map records between the six possible source pairs. Each pair was evaluated for common, consistent attributes, and compared against these columns. The software package used, (FRIL, Jurczyk *et al.*, 2008), provides an Expectation Maximization algorithm to iteratively optimize the column weighting, but due to the large number of records, this proved ineffective. Instead, samples were examined, and the criteria were set by manual tuning both the weightings and acceptance levels to match a training dataset of valid linkages (Table 3).

Source A	Source B	acceptance level	column	distance metric	weight
DS	ITU	92	callsign	Jaro-Winkler[1]	50
			MMSI	equal	40
			name	Jaro-Winkler	40
DS	VT	85	callsign	Jaro-Winkler	60
			IMO	Jaro-Winkler	20
			name	Jaro-Winkler	20
FCC	ITU	85	callsign	Jaro-Winkler	95
			name	Jaro-Winkler	5
FCC	VT	95	callsign	Jaro-Winkler	66
			name	Jaro-Winkler	5
			MMSI	Jaro-Winkler	24
			length	equal	5
VT	ITU	80	callsign	Jaro-Winkler	20
			MMSI	Edit Distance	30
			name	Jaro-Winkler	10
			IMO	Jaro-Winkler	40
DS	FCC	95	callsign	equal	99
			name	Jaro-Winkler	1

Table 3: Ship record linkage methods used.
[1] Jaro-Winkler distance (Winkler, 1990): length $l = 4$ and scaling factor $p = 0.1$

For most attributes, either the equal fields (both values are the same) or the Jaro-Winkler distance metric were used. Jaro-Winkler has useful properties for this data: it is effective on both numeric and textual data, and is particularly good in picking up the kinds of errors inherent in user provided data sources such as those used in this study. A study of string comparison metrics found it to be both efficient, and effective, with a high match rate on diverse data (Cohen *et al.*, 2003). The equations used in the

Jaro-Winker distance are described in (Appendix Appendix C).

Once each combination between two sources was finished running in FRIL, a second rule-based method was developed to capture valid pairings initially missed. If vessels had equality in any two attributes of the set $MMSI, callsign, name$, if callsign and vessel length matched, or if the IMO number provided was a valid seven digit number, then the pair was marked as linked. This was tested against a number of pairs manually, and successfully caught many of the initially missing yet valid linkages.

2.6.1 Linkage Validation

After pairwise linkage, matched records (Table 4) were further cross-linked, to account for vessels appearing in multiple sources. The number of links per ship averaged 3.5 ($\mu = 3.49, \sigma = 0.828$), though this distribution is skewed by a handful of vessels which were linked to many vessels incorrectly. To correct these problems and measure the effectiveness of the record linkage steps, I performed validation using the techniques which follow.

Detecting errors in this data is particularly problematic, as the records are highly correlated by nature – often, IMO number, callsign and name differ by only one character for two ships in the same fleet. We want to keep these kinds of vessels separate, while simultaneously finding small differences due to entry error, which proved difficult. I developed a validation score to remove over-aggressive links between the data, and allow us a second quality check which is tunable to threshold the data, depending on the kinds of error we are willing to accept.

To start with, I evaluated the bad joins present in the data, and the specific traits that were common across the sample. These included:

1. >6 linked records
2. >1 radio callsign
3. >1 MMSI
4. clear name mismatches
5. vessel assigned to multiple incompatible classes

Based on this, a set of rules was developed to assign a validation score and probable vessel class based on the inputs. Again Jaro-Winkler was used to compare attribute matches for both ship name and radio callsign, with $1 - d_w$ being added to the validation score. For attributes that had a single value, the attribute score was increased by one, otherwise vessels which had more than two MMSIs or five linked records had one point removed from their validation score. Lastly, if a ship was identified as being in multiple

incompatible classes, one point was subtracted for each additional class. Our final scores for all vessels is shown in (Figure 6).

Figure 6: Validation scores, bin width=0.3

These scores were then used to select valid vessels, and to assign vessels to the nine classes. Only scores exceeding zero were used for the movement models. Finally, though all these steps give us good self-consistency, I wanted to test against even better sources. While some of our input datasets are authoritative, the best available data remains commercial. A a 1% sample of the records were compared to those provided in Equasis (Equasis, 2012), which includes validated records from the commercial fleet. This showed that, after cross-linking, the validation score showed good correpsondence with true vessels, though only comparisons to commercial classes of vessels were possible.

3 Results

3.1 Record Linkage

Record linkage matched 30-50% of each record source pair (Table 4), and additional validation boosted this even further. That authoritative data is inconsistent points toward needing better unified and public identifiers than are currently exist.

Source *A*	Source *B*	matched records	% possible matched
DS	FCC	3481	50.35[1]
DS	ITU	41380	30.73
DS	VT	72286	53.68
FCC	ITU	27874	50.58[1]
FCC	VT	5282	53.23[1]
VT	ITU	54727	43.25

Table 4: Ship record linkage results summary.
1. FCC data is US only, % possible is of US-only data from these sources.

3.2 Geographic Validation

The land-sea mask developed in Section 2.2.2 was compared against each observation: is the point contained within a water body? However, this simple validation technique wasn't sufficient to resolve many observations, which lie at the edge of the two classes when docked and moving near shore (Figure 12). Due to this and other limitations, the observations 'on land' were retained for the rest of the analysis, but perhaps a better approach would be to apply a local density estimation on the vessels, and determine a mask based on thresholding the data. Instead, most of the on land observations were filtered by using our track generation rules, which by restricting vessel movements to a possible range, generally remove individual erroneous observations, but a more robust technique would be useful for systematic errors.

3.3 Ship results

3.3.1 Speed

Ship speed is an important predictor for a variety of questions, but is challenging to represent spatially, as it is sensitive to the observation frequency in a particular location, along with accurate distance and time measurements. To get a general sense of ship

Figure 7: Boxplot of validated ship speeds.

speed, the speed distribution of each vessel class was examined, both through simple boxplots (Figure 7), and through a more nuanced kernel density estimation (Figure 8). The density estimation shows that for many classes, distinct speed patterns are clearly visible. For example, support vessels, which include tugs and barges move much more slowly than high-speed transport vessels. Other classes are less distinct in speed signature: cargo and tanker vessels have surprisingly similar average speeds, though by breaking this down spatially (Figure 9), we can see a more nuanced story.

Figure 8: Distribution of validated ship movement speeds, computed using kernel density comparisons, as described in (Bowman and Azzalini, 1997).

Figure 9: Average ship speed examples. Scale shows 10-20 Kts/hr range for all vessels. Of note is the large difference in speeds between the opposite shores of North America.

3.3.2 Density

The density maps show the dramatically different movement structure of the major vessel classes: cargo, tankers and passenger ships all exhibit important differences in their movement.

Figure 10: Ship movement densities. first row: cargo, tankers, support. second row: pleasure, passenger, fishing. third row: other, high-speed, authority.

3.3.3 Spatial Autocorrelation

Geographic features tend to be clustered, exhibiting spatial autocorrelation. Here, we use Moran's I to compute global autocorrelation statistics for our density rasters:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

where n is the number of cells, and w_{ij} is the spatial weight.

4 Discussion

Shipping is a major user of the ocean, but little is known about its distribution and effects. Here I attempted to build the first validated and global models of ship movement, to better enable us to manage the ocean effectively. The shipping companies acknowledge the importance of managing the ocean holistically, but lack the scientific knowledge and tools to do so effectively. By incorporating ecological information alongside logistical efficiencies, it should be possible to improve the system robustness.

Marine protected areas (MPAs) have been shown to be effective (Halpern and Warner, 2002), but the multidimensional nature of ocean use is pointing toward dynamic MPAs, which may rely on providing users, such as ship operators, with real-time information about the state of the environment. Marine spatial planning is proving a promising avenue for brining stakeholders together (Merrifield *et al.*, 2012). This new form of planning has greater data requirements, which in the ocean can be simplified down to three major areas of use: fisheries management, transportation management, and energy management (Lubchenco and Laura E. Petes, 2010). Transportation in the ocean is the least studied of the three, and here we have shown how volunteered geographic information methods, along with volunteered observations, can provide us a way to tackle the data-poor problem.

5 Conclusion

The future will include global AIS coverage Jones (2012); Carson-Jackson (2012), but methods to validate and integrate this data into a scientifically useful system don't yet exist. Here we've shown that by using VGI methods, we can incorporate many data dimensions and correct for multiple forms of error present in the data.

The approach outlined here is in many ways just the beginning: logical next steps would be deriving true cost-path movement models for the ships, determining the predominant direction within each cell, and producing an abstract network model from the data.

Appendix A Figures

Figure 11: Raw AIS observations, November 2011. Note the observations located in the Hoggar Mountains in Algeria.

Type	Sub-type	Vessels
cargo	cargo ship	30355
	merchant	2919
	bulk carrier	1338
	container ship	669
	general cargo	608
	vehicle carrier	56
tanker	tankship	10460
	tanker	8731
	oil tanker	1312
	liquefied gas carrier	112

Type	Sub-type	Vessels
	chemical carrier	39
other	merchant	17460
	other ship	9421
	unspecified	4580
	motor boat	4231
	inland waterways	3825
	sloop	2799
	reserved for future use	2164
	all other activities	1345
	reserved for regional use	712
support	pusher/tug	4392
	tug	7885
	towing vessel	2747
	supply vessel	1501
	service vessels	1388
	trawler	1056
	dredger	995
	vessel engaged in dredging or underwater operation	776
	pilot vessel	685
pleasure	pleasure/leisure	22013
	pleasure	14573
	pleasure craft	7290
	sailing vessel	5527
	yacht	8214
fishing	fishing vessel	11110
	fishing boat	8454
	fishing industry	5767
	fishing	3213
passenger	passenger ship	6478
	ferry	704
high-speed	high-speed craft	658
	high speed craft	520
authority	sar-vessel	699
	search and rescue vessel	609
	rescue vessel	102

Table 6: Detailed ship class breakdown. Classes come from observationsgT

Figure 12: Long Beach Harbor, California. Points shown here in purple are 'on land', but most of these on-land observations are actually parts of the harbor.

Figure 13: AIS observations, Southern California Bight. Nov 2011–Dec 2012. Note the ballast water exchange point lower left.

Figure 14: Approximate AIS coverage (green), global ports (red). Cylindrical Equal-Area projection

Figure 15: "World Shipping Lanes" map produced by the Central Intelligence Agency, 1973.

Attribute	Accuracy
Location (fixed from GPS signal)	$\simeq$10 meter accuracy)
Timestamp (on broadcast)	$\simeq$100 ms *radio transmission & processing latency*
Name	
Call Sign	
Maritime Mobile Service Identity (MMSI)	
Heading	
Speed	
Destination	*rarely valid*

Table 5: AIS broadcast attributes. Update frequency depends on ship speed, but varies between a minimum of a record every 2 seconds for quickly moving vessels, to once per 3 minutes for moored vessels. Additional attributes are available, but infrequently used.

Appendix B Source Code

All code used to generate the project is available at https://github.com/scw/ais-kml-parser. Due to it's length, I've opted to omit it here.

Appendix C Record Linkage

The Jaro-Winkler formula is defined in two

The Jaro distance:

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{4}$$

d_j Jaro Distance m = number of matching patterns t = number of transposed characters |s1| = length of first string |s2| = length of second string

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1 \tag{5}$$

range = max(length(a), length(b)) / 2 - 1

Jaro-Winkler distance:

$$d_w = d_j + (\ell p(1 - d_j)) \tag{6}$$

References

Bivand R (2011). "Introduction to representing spatial objects in R."

Bowman AW, Azzalini A (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Brohan P, Allan R, Freeman J, Waple A, Wheeler D, Wilkinson C, Woodruff S (2009). "Marine observations of old weather." *Bulletin of the American Meteorological Society*, **90**(2), 219–230.

Canning D, Fay M (1993). "The effects of transportation networks on economic growth."

Carson-Jackson J (2012). "Satellite AIS–Developing Technology or Existing Capability?" *Journal of Navigation*, **65**(02), 303–321.

Christen P (2012). *Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer, Berlin; New York. ISBN 9783642311642 3642311644.

Cogan J (2009). "Representation and Power in International Organization: The Operational Constitution and Its Critics." *American Journal of International Law*, **103**, 209.

Cohen W, Ravikumar P, Fienberg S, *et al.* (2003). "A comparison of string distance metrics for name-matching tasks." In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pp. 73–78.

Corbett JJ, Winebrake JJ, Green EH, Kasibhatla P, Eyring V, Lauer A (2007). "Mortality from ship emissions: a global assessment." *Environmental science & technology*, **41**(24), 8512–8. ISSN 0013-936X.

Costello C, Ovando D, Hilborn R, Gaines S, Deschenes O, Lester S (2012). "Status and Solutions for the Worlds Unassessed Fisheries." *Science*, **338**(6106), 517–520.

Currier KE, Goodchild MF, Li L, Hurt I, Mooney P (2012). "The Social Construction of Feature Types in Volunteered Geographic Information." Unpublished Draft.

Davenport J, Davenport JL (2006). *The ecology of transportation: managing mobility for the environment*. Springer, Dordrecht, Netherlands.

De Smith M, Goodchild M, Longley P (2007). *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Publishing.

Elwood S, Goodchild M, Sui D (2011). "Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice."

Equasis (2012). "The world merchant fleet in 2011." *Technical report*, Equasis. URL http://www.equasis.org/EquasisWeb/restricted/PublicStatistic?fs=ShipSearch.

FAO (2012). "Vishing Vessels." URL http://www.fao.org/fishery/topic/1616/en.

Federal Geographic Data Committee (2008). "Geographic Information Framework Data Content Standard Part 0: Base Document. FGDC-STD-014.0-2008." URL http://www.fgdc.gov/standards/projects/FGDC-standards-projects/framework-data-standard/GI_FrameworkDataStandard_Part0_Base.pdf.

Fletcher J (2010). "JCOMM Voluntary Observing Ship Scheme." URL http://www.bom.gov.au/jcomm/vos/vos.html.

Freeman JE (2010). "NCDC: Volunteer Observing Ship Climate Project." URL http://www.ncdc.noaa.gov/oa/climate/vosclim/vosclim.html.

Fujiwara M, Caswell H (2001). "Demography of the endangered North Atlantic right whale." *Nature*, **414**(6863), 537–41. ISSN 0028-0836. doi:10.1038/35107054.

Goodchild M (1992). "Geographical information science." *International Journal of Geographical Information Systems*, **6**(1), 31–45.

Goodchild M (1999). "Cartographic futures on a digital earth." *The Map Reader*, pp. 141–146.

Goodchild M (2007). "Citizens as sensors: the world of volunteered geography." *GeoJournal*, **69**(4), 211–221.

Goodchild M, Gopal S (1989). *The accuracy of spatial databases*. CRC.

Goodchild M, Li L (2012). "Assuring the quality of volunteered geographic information." *Spatial Statistics*, pp. 110–120.

Guimera R, Mossa S, Turtschi A, Amaral L (2005). "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles." *Proceedings of the National Academy of Sciences*, **102**(22), 7794–7799.

Halpern B, Warner R (2002). "Marine reserves have rapid and lasting effects." *Ecology letters*, **5**(3), 361–366.

Halpern BS, Longo C, Hardy D, McLeod KL, Samhouri JF, Katona SK, Kleisner K, Lester SE, O'Leary J, Ranelletti M, Rosenberg Aa, Scarborough C, Selig ER, Best BD, Brumbaugh DR, Chapin FS, Crowder LB, Daly KL, Doney SC, Elfes C, Fogarty MJ, Gaines

SD, Jacobsen KI, Karrer LB, Leslie HM, Neeley E, Pauly D, Polasky S, Ris B, St Martin K, Stone GS, Sumaila UR, Zeller D (2012). "An index to assess the health and benefits of the global ocean." *Nature*, **488**(7413), 615–20. ISSN 1476-4687. doi:10.1038/nature11397.

Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, Bruno JF, Casey KS, Ebert C, Fox HE, Fujita R, Heinemann D, Lenihan HS, Madin EMP, Perry MT, Selig ER, Spalding M, Steneck R, Watson R (2008). "A global map of human impact on marine ecosystems." *Science New York NY*, **319**(5865), 948–952.

IALA AISM (2004). "1028 On the Automatic Identification System (AIS) Volume 1, Part I, Operational Issues."

International Maritime Organization (1980). "International convention for the Safety of Life at Sea (SOLAS), Chapter V Safety of Naviation, Regulation 19."

International Maritime Organization (2009). "Prevention of air pollution from ships second IMO GHG study 2009." *Technical report*, International Maritime Organization.

ITU (2010). "Recommendation ITU-R M.1371-4." *Technical report*.

Jensen A, Silber G, Calambokidis J (2004). *Large whale ship strike database*. US Department of Commerce, National Oceanic and Atmospheric Administration.

Jones MT (2012). "Keynote: 2012 Joint Warfighting Conference." `http://www.youtube.com/watch?v=Dk3IIOimGwE`. URL `http://www.youtube.com/watch?v=Dk3IIOimGwE`.

Jurczyk P, Lu J, Xiong L, Cragan J, Correa A (2008). "FRIL: A tool for comparative record linkage." In *AMIA Annual Symposium Proceedings*, volume 2008, p. 440. American Medical Informatics Association.

Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010). "The complex network of global cargo ship movements." *Journal of the Royal Society, Interface / the Royal Society*, **7**(January), 1093–1103. ISSN 1742-5662. doi:10.1098/rsif.2009.0495.

Karney C (2012). "Algorithms for geodesics." *Journal of Geodesy*, pp. 1–13.

Lack D, Cappa C, Langridge J, Bahreini R, Buffaloe G, Brock C, Cerully K, Coffman D, Hayden K, Holloway J, *et al.* (2011). "Impact of Fuel Quality Regulation and Speed Reductions on Shipping Emissions: Implications for Climate and Air Quality." *Environmental science & technology*, **45**(20), 9052–9060.

Lagueux K, Zani M, Knowlton A, Kraus S (2011). "Response by vessel operators to protection measures for right whales Eubalaena glacialis in the southeast US calving ground." *Endangered Species Research*, **14**(1), 69–77. ISSN 1863-5407. doi: 10.3354/esr00335.

Lloyd's Register-Fairplay Ltd (2010). "World shipping encyclopaedia." URL `http://www.ihsfairplay.com/Maritime_data/WSE/WSE.html?product=WSE`.

Lubchenco J, Laura E Petes (2010). "The interconnected Biosphere: Science at the oceans tipping Points." *Oceanography*, **23**(2), 115.

McKenna MF, Katz SL, Condit C, Walbridge S (2012). "Response of Commercial Ships to a Voluntary Speed Reduction Measure: Are Voluntary Strategies Adequate for Mitigating Ship-Strike Risk." *Coastal Management*, **40**(October), 634–650.

Merrifield M, McClintock W, Burt C, Fox E, Serpa P, Steinback C, Gleason M (2012). "MarineMap: a web-based platform for collaborative marine protected area planning." *Ocean & Coastal Management*.

Miller H, Bridwell S (2008). "A field-based theory for time geography." *Annals of the Association of American Geographers*, **99**(1), 49–75.

Miller H, Han J (2009). *Geographic data mining and knowledge discovery*. CRC.

Miller HJ (2005). "A measurement theory for time geography." *Geographical analysis*, **37**(1), 17–45. ISSN 0016-7363. doi:10.1111/j.1538-4632.2005.00575.x.

Mueller H (2011). "Sailwx.info." URL `http://www.sailwx.info`.

Open Geospatial Consortium Inc (2008). *OGC KML 07-147r2*, 2.2.0 edition.

Open Geospatial Consortium Inc (2010). "Simple features specification for SQL."

PostgreSQL Global Development Group (2012). "PostgreSQL 8.4.14." `http://www.postgresql.org`.

Rabus B, Eineder M, Roth A, Bamler R (2003). "The shuttle radar topography missiona new class of digital elevation models acquired by spaceborne radar." *ISPRS Journal of Photogrammetry and Remote Sensing*, **57**(4), 241–262.

Ramsey P, *et al.* (2005). *PostGIS manual*.

Raymond ES, Schwehr K (2012). "The Toils of AIS: A Case Study in Application Protocol Design and Analysis." Unpublished Draft, URL https://gitorious.org/toils-of-ais/toils-of-ais.

Rodrigue J, Comtois C, Slack B (2009). *The geography of transport systems*. 2nd edition. Taylor & Francis.

Slater J, Garvey G, Johnston C, Haase J, Heady B, Kroenung G, Little J (2006). "The SRTM data finishing process and products." *Photogrammetric engineering and remote sensing*, **72**(3), 237–247.

Tetreault B (2002). "Use of the Automatic Identification System (AIS) for Maritime Domain Awareness (MDA)." *Proceedings of OCEANS 2005 MTS/IEEE*, pp. 1–5. doi: 10.1109/OCEANS.2005.1639983.

Tun M, Chambers G, Tan T, Ly T (2007). "Maritime port intelligence using AIS data." *Recent advances in security technology*, p. 33.

Univeristy of the Aegean (2010). "Live Ships Map - AIS." URL http://www.marinetraffic.com/ais/.

Vanderlaan A, Corbett J, Green S, Callahan J, Wang C, Kenney R, Taggart C, Firestone J (2009). "Probability and mitigation of vessel encounters with North Atlantic right whales." *Endangered Species Research*, **6**(March), 273–285. ISSN 1863-5407. doi:10.3354/esr00176. URL http://www.int-res.com/abstracts/esr/v6/n3/p273-285/.

Vries G, Malaisé V, Someren M, Adriaans P, Schreiber G (2009). "Semi-automatic ontology extension in the maritime domain." In *Proceedings of the twentieth Belgian-Dutch Conference on Artificial Intelligence*, pp. 265–272. University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

Wang C, Corbett J, Firestone J (2007). "Modeling energy use and emissions from North American shipping: application of the ship traffic, energy, and environment model." *Environmental science & technology*, **41**(9), 3226–32. ISSN 0013-936X.

Wessel P, Smith W (1996). "A global, self-consistent, hierarchical, high-resolution shoreline database." *Journal of Geophysical Research*, **101**, 8741–8743.

Westwood J, Parsons B, Rowley W (2001). "Global ocean markets." *Oceanography*, **14**(3), 83.

Winkler W (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." Proceedings of the Section on Survey Research Methods (American Statistical Association).

Woodruff S, Worley S, Lubker S, Ji Z, Eric Freeman J, Berry D, Brohan P, Kent E, Reynolds R, Smith S, *et al.* (2010). "ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive." *International Journal of Climatology*, **31**(7), 951–967.

Worm B, Hilborn R, Baum J, Branch T, Collie J, Costello C, Fogarty M, Fulton E, Hutchings J, Jennings S, *et al.* (2009). "Rebuilding global fisheries." *science*, **325**(5940), 578–585.

Wright D, Goodchild MF (1997). "Data from the deep: Implications for the GIS community." *International Journal of Geographical Information Science*, **11**(6), 523–528.

Wu Y, Miller H (2002). "Computational tools for measuring space-time accessibility within transportation networks with dynamic flow." *Journal of Transportation and Statistics*, **4**(2/3), 1–14.