SPOUDAI Journal of Economics and Business, Vol. 63 (2013), Issue 3-4, pp. 181-190

University of Piraeus

SPOUDAI
Journal of Economics and Business
Σπουδαί
http://spoudai.unipi.gr

Efficient AIS Data Processing for Environmentally Safe Shipping

Marios Vodas[a], Nikos Pelekis[b], Yannis Theodoridis[c], Cyril Ray[d], Vangelis Karkaletsis[e], Sergios Petridis[f], Anastasia Miliou[g]

[a]*University of Piraeus, Department of Informatics, 80, M. Karaoli & A. Dimitriou Str, 185 34 Piraeus, Greece, Email: mvodas@unipi.gr*
[b]*University of Piraeus, Department of Statistics and Insurance Science, 80, M. Karaoli & A. Dimitriou Str, 185 34 Piraeus, Greece, npelekis@unipi.gr*
[c]*University of Piraeus, Department of Informatics, 80, M. Karaoli & A. Dimitriou Str, 185 34 Piraeus, Greece, Email: ytheod@unipi.gr*
[d]*IRENav – Naval Academy Research Lab, BCRM de Brest, CC 600 - Lanveoc, F-29240 BREST Cedex 9, France, Email: cyril.ray@ecole-navale.fr*
[e]*Institute of Informatics and Telecommunications, NCSR "Demokritos", Agia Paraskevi Attikis, P.O. Box 60228, 153 10 Athens, Greece, Email: vangelis@iit.demokritos.gr*
[f]*Institute of Informatics and Telecommunications, NCSR "Demokritos", Agia Paraskevi Attikis, P.O. Box 60228, 153 10 Athens, Greece, Email: petridis@iit.demokritos.gr*
[g]*Archipelago – Institute of Marine Conservation, Pythagorio, P.O. Box 42, 831 03 Samos, Greece, Email: a.miliou@archipelago.gr*

Abstract

Reducing ship accidents at sea is important to all economic, environmental, and cultural sectors of Greece. Despite an increase in traffic and national monitoring, ships formulate routes according to their best judgment risking an accident. In this study we take a dataset spanning in 3 years from the AIS (Automatic Identification System) network, which is transmitting in public a ship's identity and location with an interval of seconds, and we load it in a trajectory database supported by the Hermes Moving Objects Database (MOD) system. Presented analysis begins by extracting statistics for the dataset, both general (number of ships and position reports) as well as safety related ones. Simple queries on the dataset illustrate the capabilities of Hermes and allow to gain insight on how the ships move in the Greek Seas. Analysis of movement based on an Origin-Destination matrix between interesting areas in the Greek territory is presented. One of the newest challenges that emerged during this process is that the amount of the positioning data is becoming more and more massive. As a conclusion, a preliminary review of possible solutions to this challenge along with others such as dealing with the noise in AIS data is mentioned and we also briefly discuss the need for interdisciplinary cooperation.

JEL Classification: Z00.
Keywords: AIS; Data; Environment; Analysis; Hermes MOD; Aegean Sea.

1. Introduction

The maritime environment has a huge impact on the economy and our everyday lives. Beyond, being a space where numerous marine species live, the sea is also a place

where human activities (sailing, cruising, fishing, goods transportation...) evolve and increase drastically. For instance, oil spill clean-ups of the Greek Seas can cost over 1 billion Euros, whereas accidents involving water soluble cargos would result in irrevocable changes, especially to ecosystem. Reducing the number of accidents at sea is therefore important, with respect to economic, environmental, and cultural sectors of Greece. Despite an increase in traffic and efforts at the European level, there are no national-level monitoring policies and ships formulate routes according to their best judgment, which at many cases involves crossing sensitive biodiversity areas or narrow passages, thus risking the effects of strong winds and congestion which could lead to vessel groundings or collisions. However, the Automatic Identification System (AIS, IALA 2004), that has been recently made a mandatory standard on cargo, passenger and fishing ships, brings opportunities. AIS is embedded on ships, transmitting in public their identity and location (among other information) with an interval of seconds, depending on the ship's behaviour[1].

Nowadays, many systems (including ethatic ones) have been constructed around the world with low cost antennas in order to capture and record such data. These positioning data are either published on the Web, showing real-time ship location, and in some cases are stored in relational databases, where they more or less remain unexploited. The idea is that such a massive spatio-temporal dataset can be used towards the following objectives (the list can be extended):

- *Improving safety:* analyse and understand past conflicts (when two ships fail to meet minimum safety distance); analyse the accuracy of data provided by base stations; etc.
- *Optimizing traffic at sea:* calculate metrics from the traffic (e.g. traffic density, mean distance between ships, number of trajectories that are close to the 'optimal' departure – arrival path); devise new sea routes to handle traffic increase; measure the activity of each ship (e.g. number of intermediate stops); etc.
- *Considering the environment:* compare trajectories with environmental considerations (fuel consumption, noise pollution, vertical profile comparison); study the environmental impact of ship routes on population of marine life; etc.

Along this direction, the recently launched AMINESS project (the acronym stands for *Analysis of Marine Information for Environmentally Safe Shipping)* aims at contributing in the safety, management and monitoring of the sea environment and the Greek Seas, in particular. Based on historical and real-time maritime data, including information for ships' position and speed, weather forecasting and areas of interest (in land or sea), the project objective is the development of a platform that envisages to (a) suggest vessel and environmentally optimal safe route planning, (b) deliver real-time alerts for ships that violate environmental security regulations, and (c) support policy recommendations. Through this platform, the project aims to help reducing the risk of a ship accident and, consequently, contribute in the protection of the Greek Seas.

1. Beyond AIS, which is the most widely spread network, there are other systems such as the EU's Long Range Identification System (LRIT) that can provide positioning data for ships using more secure communication protocols to ensure the integrity of data.

2. Motivation

AIS data are massively available through internet sources (e.g. marinetraffic.com, mariweb.gr, aishub.net, vesseltracker.com) in the form of streams of messages, so called AIS sentences, and nearly at no cost. Nowadays, position reports in European waters reach almost 5 million per day corresponding to 17,000 ships (EMSA, 2012). Keeping in mind that the number of transceivers is continuously increasing and the quality of data is improving it is clear that the effective amount of data we have to process is ever increasing.

The amount of data gathered, together with their spatial/temporal distribution and their origin, provide further challenges related to distributed and scalable reasoning, retrieval of data from the different sources considered, "feeding" the knowledge-intensive tasks of interest. Issues related to the provenance and trustworthiness of incoming data sources, which are of major importance to the safe shipping domain, should be also considered, mainly focusing on the less studied intertwined nature of the sources (Zhao *et al.*, 2004).

The raw format for AIS data in relational database terms is a table of positions along with the timestamp each position was observed by the base stations. Querying such schema has certain drawbacks in terms of computing cost and interpolation. For instance, consider a ship reporting positions around and inside an area of interest. It is very difficult and complex to determine if a ship entered or left this area using only the point-based model. Therefore, trajectory features are required to allow us to inherently solve most issues that arise from movement of objects, such as interpolation and distance computation. Furthermore, by using this model we can reduce the computational cost by using simplification methods to limit the number of points in the trajectory.

The ways to define a trajectory vary significantly and are subject to the domain they are applied (Spaccapietra *et al*., 2008). Our focus in this study is mainly on answering questions regarding the movement of ships based on spatiotemporal predicates. Therefore, we model a trajectory as a sequence of sampled time-stamped locations (p, t) where p is a 2D point (x, y) and t is the recording timestamp of p. We can choose from two alternatives for interpolating the position of an object between two sampled points: the first and most common one is to assume constant speed linear interpolation while the second one is to consider constant acceleration. Although, the second option is closer to a real-world model we adopt the first option, since it is easier to be implemented and suitable for our data (a ship would not change its speed as rapidly as a car for example).

This need for capturing, storing, and querying complete histories of objects' movement has promoted the investigation of Moving Object Databases (MOD) (Güting *et al.*, 2000), which are based on the modelling of moving extensions of spatial types, such as points, lines, regions, etc. A state-of-the-art realization of the above model in Object-Relational DBMS is the so-called *Hermes* (Pelekis *et al.*, 2008; Pelekis *et al*., 2013). Hermes provides an extensive spatiotemporal functionality over MODs, including aspects from online data feeding and efficient trajectory storage to advanced query processing and data mining operations (Pelekis and Theodoridis, 2013).

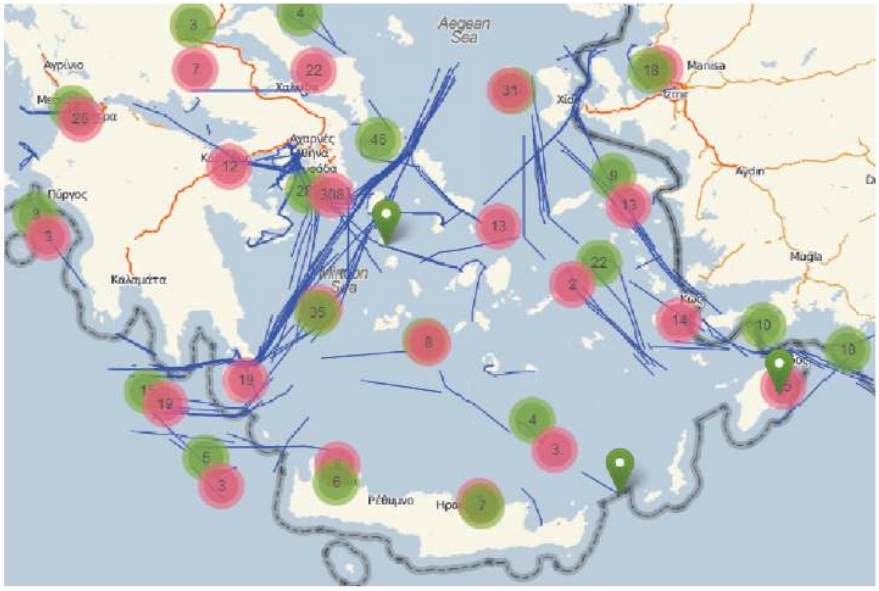
For instance, application of analytical methods, tailored for spatiotemporal data, are

essential to gain insight in ship movement behaviour (Devogele *et al.,* 2013). Novel, scalable, and application domain aware warehouse designs and algorithms are required to effectively handle large volumes of data and to provide answers to new types of questions concerning environment protection. Hermes already incorporates clustering algorithms in that direction with the prospect of expanding these capabilities.

3. Efficient AIS data processing

This section introduces the first, to our knowledge, real world experience with order of million AIS sentences. The database in its original form is almost 2 TB in size and contains records of raw AIS sentences from April 2007 to June 2010[2]. However, for the purposes of this study, we extracted a 3 days subset which contains approximately 3 million GPS records from 933 ships, spans in a 3 days period, from 31/12/2009 to 2/1/2010, and covers an area of 496736 km^2 in a rectangle bounded by coordinates from (21, 35) to (29, 39) expressed in WGS84 (i.e. long, lat). Fig. 1 presents a sample of the ship trajectories as blue lines. Green numbered circles indicate that there were a certain number of ships (the number is in the center of the circle) that started their trip around that area. Red numbered circles have a similar meaning only regarding the end of the trip.

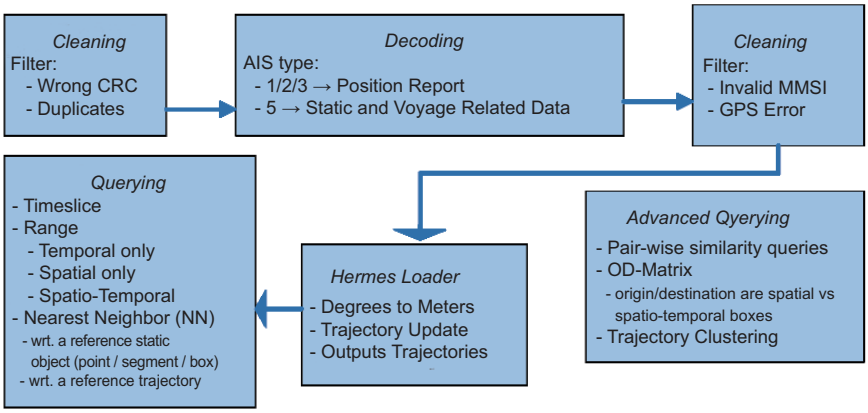
Figure 1. "IMIS 3 Days" visualization



2. The "IMIS 3 Days" dataset is publicly available at ChoroChronos.org online datasets and algorithms repository (http://www.chorochronos.org/?q=node/8).

Fig. 2 graphically presents the steps in our data processing methodology. These several steps are explained in detail throughout the next subsections.

Figure 2. Processing methodology



3.1 Loading the AIS data in a MOD and extracting useful statistics

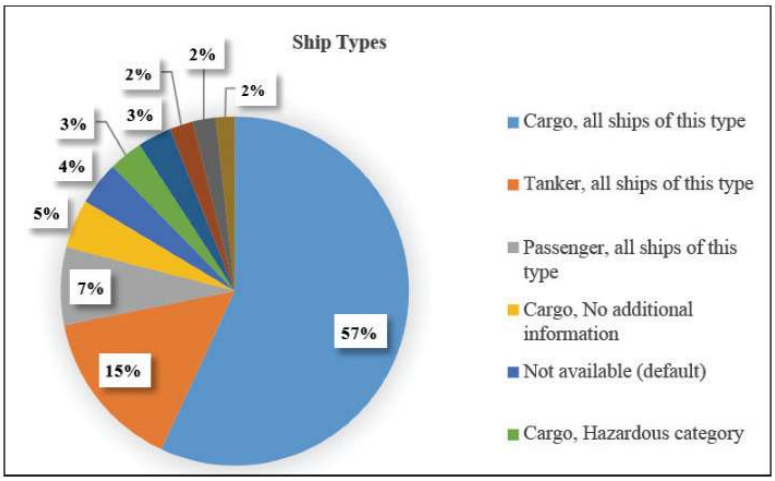
Examination of the dataset showed the first challenges in data cleaning. Each AIS sentence is received multiple times from different stations. Another problem is that some sentences are incorrect or incomplete therefore cannot be decoded. It is expected to confront these problems in every raw AIS dataset (historical or streaming). In this case study, where the database is historical, cleaning duplicated and bad messages is straightforward and even in a centralized system the computational cost is acceptable. In order to speed-up this operation we partitioned data into multiple tables using the AIS checksums to make the cleaning process more effective. More ways can be used for partitioning of course but our experience in this experiment showed that choosing checksums as the partition key results in relatively same size partitions.

To extract the position reports from the AIS sentences we used Hermes' built-in AIS decoder and applied simple filters on the positions and the unique ship identifiers (Maritime Mobile Service Identity-MMSI) to eradicate invalid ship identifiers and GPS errors. After that, the *Loader* module of Hermes is used to construct (valid) ship trajectories out of the decoded dataset. An index structure on the trajectories is also built to make queries with spatio-temporal predicates more efficient.

Understanding the dataset is prerequisite for its efficient and effective exploration. Typical statistical analysis, include the number of ships and positions recorded, the spatial and temporal bounds of the dataset, as well as ship types and flags we present in the following paragraphs.

In Fig. 3 we show the types of ships that travel in the Greek Seas; the chart makes clear that the majority of ships are cargo and tankers and that quite often they carry hazardous materials. This is an indicator that supports the idea of monitoring the movement of ships and creating safe routes on sea.

Figure 3. Ship types in Greek seas



Regarding the flags of vessels, Table 1 presents the most frequent ones. It is worth to be mentioned that 4 out of the top-6 flags in the dataset are flags of convenience.

Table 1. Ship flags in the Greek Seas

Country	Number of ships	Flag of Convenience
Greece	263	No
Panama (Republic of)	112	Yes
Turkey	96	No
Malta	76	Yes
Liberia (Republic of)	32	Yes
Vincent and the Grenadines	29	Yes

3.2 Querying the AIS database, focusing on spatio-temporal attributes

Taking geographic information into account is of paramount interest in processing an AIS database. Apart from the information already embedding in AIS signals, complementary information can be obtained from official maritime vector charts (eg. type S-57) that contain different kind of objects useful for spatial analysis (coast lines, restricted areas ...) or even added manually, according analysis objectives. In our study, we have integrated geographic data, such as points or regions of interest (ports, lighthouses), coastal boundaries, etc. Thereafter, we were able to support queries, such as:

- Find the ships that were located closer than 1/2 n.m. from the port of Patras' lighthouse (Fig. 4);

- Find the ships that crossed the Euboea - Andros narrow passage (Fig. 5); etc.

Hermes provides the ability to build a 3D-Rtree index on the spatio-temporal segments of the trajectories in order to efficiently process queries in this category (Theodoridis, 1996). Basically, the 3D-Rtree has an operator class associated with it so that whenever we use an operator from that class we can take advantage of the index. An example of such operators would be the "within distance", which traverses the tree and returns only the segments that fulfill the distance threshold we provide as input. This operator was used in the query shown in Fig. 4. Another example would be the "intersects" operator used in the query shown in Fig. 5. This query finds the enter and leave points of a ship in the specific area and uses the "intersects" operator to prune on the number of segments it has to process in order to find which ones are entering or leaving that area.

Figure 4. Ships located closer than ½ n.m. from the lighthouse port of Patras'

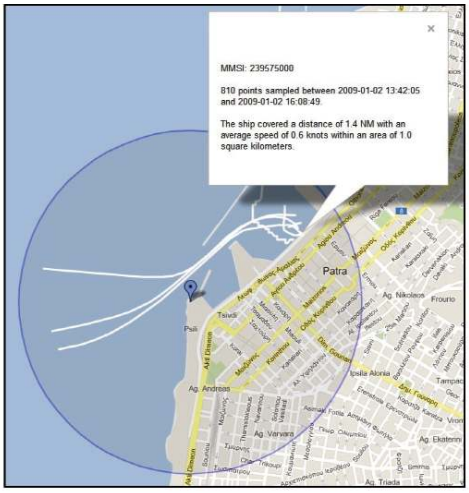
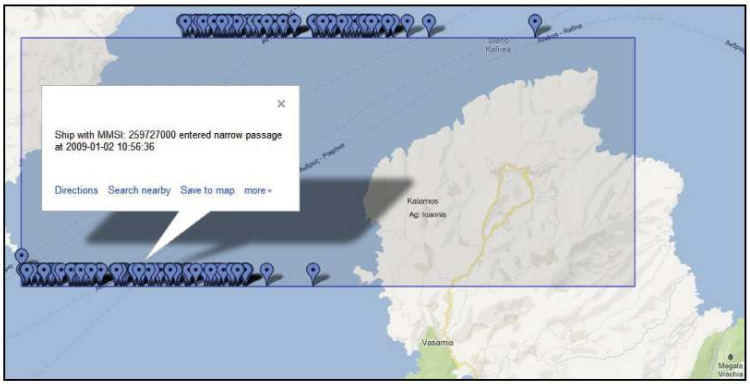


Figure 5. Ships that crossed Euboea – Andros narrow passage



3.3 Advanced processing of vessel routes

The queries belonging this category are more complex and cost more than the previous ones. Yet, they are very interesting and provide deeper insight into the data. Some of the complex operations that we are able to support are:

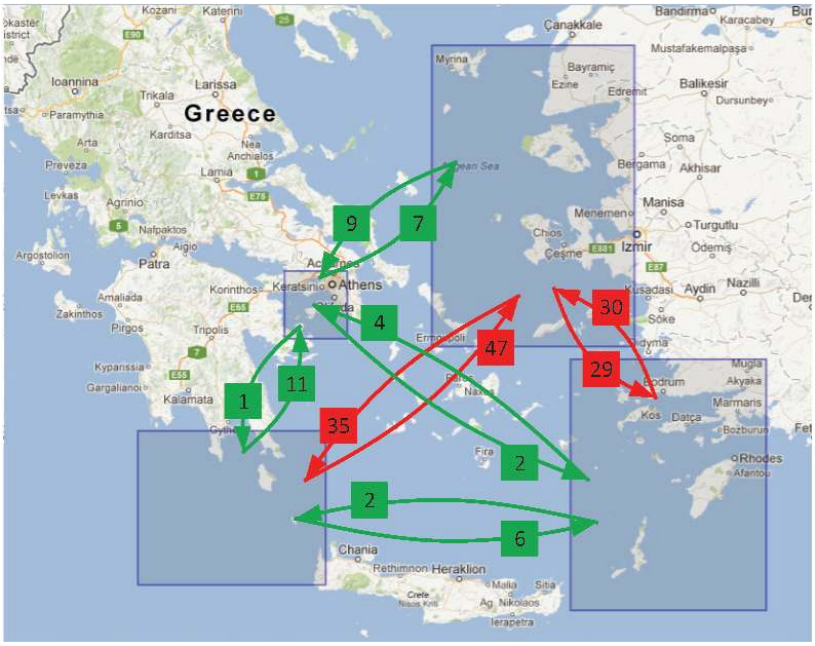
• Building the Origin-Destination (OD) matrix illustrating the number of ship trajectories between areas of particular interest (ports, sea zones, etc.);

• Searching for clusters (and outliers) over a set of ship trajectories using state-of-the-art algorithms;

• Finding the typical routes and comparing ship trajectories to them; etc.

As shown by the OD – Matrix illustrated in Fig. 6 the two most popular pairs of Sea Zones are North Aegean – Dodecanese (west) and North Aegean – Ionian / Cretan (east); meaning that North Aegean is in a congested crossroad for travelling ships. Taking into consideration the results from the statistical analysis from subsection 3.1, where we found that cargo ships are the majority and that there is a large number of flags of convenience, we can conclude that national monitoring and the formation of safe routes are necessary in order to prevent an accident that could cause irreversible damage both to the ecosystem and tourism the economic driving force of Greece.

Figure 6. Origin-Destination Matrix between 4 main entrance and exit areas in Aegean Sea



4. Discussion

Even though our experience with the 3 days subset of the AIS dataset gives us a lot of insight on the data, there is much to be done regarding the entire 3 years AIS dataset. In order to query such huge dataset we may have to develop new methods to partition it to multiple server instances so that we can take advantage of parallel query execution techniques. Of course, partitioning the data is possible but not straightforward when we try to optimize queries. Currently, we are working on a schema according to Map-Reduce paradigm (Dean, 2008).

In cases where data is streamed in real-time the approach to first gather it, then clean it, and finally store it has to be revised because accessing the history of each object is a costly operation. We can avoid it by keeping history in-memory and propose methods that use it to calculate the possibility that the position report is clean or noisy. Notably, we may better understand the parameters of the message by comparing them to the objects within the extents of the short history, thus obtaining a better accuracy on identifying noise.

Noise might not always be due to technical errors, but to a human deliberate action to avoid surveillance in a sensitive area. Online active monitoring and reporting for the objects per area of interest could help find those actions that most probably will be repetitive e.g. follow a pattern. This is a characteristic case where integrating heterogeneous data is catalytic in increasing efficiency and quality of produced knowledge. At this point, mining both offline and online should be an appropriate direction for outlier detection.

Moreover, transparent inclusion of semantic trajectory representation is nowadays a key part of mobility analysis. During the last years, methods for annotating a trajectory with semantic information have been proposed in order to understand moving object's activity (Yan, 2011). However, there is still the need for methods that will allow us to work with data originating from sources that are heterogeneous and noisy like AIS data. These methods must be enhanced with real-time capabilities, that would help many mobile applications offer richer (i.e., context-aware) services.

5. Conclusion

Maritime environment represents an increasing potential in terms of modelling, management, and understanding of mobility data. Recently, several real-time positioning systems, such as the AIS system, have been developed for keeping track of vessel movements. Analysis of such positioning data needs cooperation and input from experts in environmental and marine domains. Indeed, knowing the everyday actions of ships and their patterns help to model them and put them to efficient use. Case-based analysis with experts also provides a better understanding of the problems and challenges and guides researchers towards efficient data mining process.

This research aims at exploring new ways to analyse and visualize the information gathered and publish it to stakeholders and the public. Automatically processing huge amounts of such data in real time is a very challenging task which, when addressed, will give the ability to the public and stakeholders to report environmental wrongdoing of ships.

Acknowledgements

This research was partially supported by AMINESS project funded by the Greek government (www.aminess.eu). Cyril Ray was supported by a Short Term Scientific Mission performed at the University of Piraeus by the COST Action IC0903 on "Knowledge Discovery from Moving Objects" (http://www.move-cost.info). IMIS Hellas (www.imishellas.gr) kindly provided the AIS dataset for research purposes.

References

Dean, J. and Ghemawat, S., 2008. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51(1):107-113.

Devogele, T., Etienne, L. and Ray, C., 2013. Maritime Monitoring. In C. Renso, S. Spaccapietra and E. Zimanyi (Eds.), Mobility Data: Modeling, Management, and Understanding (pp. 224-243). Cambridge University Press.

EMSA, 2012. European Maritime Safety Agency, Annual Report 2011, 114 pages, June 2012.

Güting, R.H., Böhlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N.A., Schneider, M. and Vazirgiannis, M., 2000. A foundation for representing and querying moving objects. ACM Transactions on Database Systems, 25(1):1-42.

IALA, 2004. The Automatic Identification System (AIS), Volume 1, Part I, Operational Issues, 131 pages.

Pelekis, N., Frentzos, E., Giatrakos, N. and Theodoridis, Y., 2008. HERMES: aggregative LBS via a trajectory DB engine, In Proceedings of ACM SIGMOD.

Pelekis, N., Frentzos, E., Giatrakos, N. and Theodoridis, Y., 2013. Supporting movement in ORDBMS – the "Hermes" MOD engine. Int'l Journal of Knowledge-based Organizations (IJKBO), in press.

Pelekis, N. and Theodoridis, Y., 2013. Mobility Data Management and Exploration. Springer.

Spaccapietra, S., Parent, C., Damiani, M., Macedo, J., Porto, F. and Vangenot, C., 2008. A conceptual view on trajectories. Data & Knowledge Engineering, 65(1):126-146.

Theodoridis, Y., Vazirgiannis, M. and Sellis, T., 1996. Spatio-Temporal Indexing for Large Multimedia Applications. In Proceedings of ICMCS, page 0441.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S. and Aberer, K., 2012, SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. In Proceedings of EDBT, 259-270.

Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D. and Greenwood, M., 2004. Using semantic Web technologies for representing E-science provenance. In Proceedings of ISWC, 92-106.