Maritime Traffic Data Mining Using R

Melita HADZAGIC*, Marie-Odette ST-HILAIRE†*, Sean WEBB‡ and Elisa SHAHBAZIAN†

*Université de Montréal, OODA Technologies Inc., Montréal, QC, Canada
Email: hadzagic@crm.umontreal.ca
†OODA Technologies Inc., Montréal, QC, Canada
Email: melita.hadzagic,marie-odette.st-hilaire, elisa.shahbazian@ooda.ca
‡Defence R&D Canada Atlantic, Dartmouth, NS, Canada
Email: sean.webb@drdc-rrdc.gc.ca

Abstract—Human operators trying to establish individual or collective maritime situational awareness often find themselves overloaded by huge amounts of information obtained from multiple and possibly dissimilar sources. This paper explores potential use of open source data mining tools, in particular R software, to enable discovery of maritime traffic patterns. It also presents an assessment of R software as a data mining tool using spatio-temporal maritime traffic data such as from the Automatic Identification System (AIS), and includes scenarios of potential interest to the maritime environment.

Keywords: maritime traffic data, AIS, data mining, R

I. INTRODUCTION

Human operators trying to establish individual or collective maritime situational awareness often find themselves overloaded by huge amounts of information obtained from multiple and possibly dissimilar sources. The requirement of the International Maritime Organization (IMO) to install the Automatic Identification System (AIS) on-board ships, together with the use of other self-reporting systems based on Global Positioning System (GPS)-quality navigation information [1], contribute to the overabundance of information. This information is potentially of great value and importance but typically the resources are not fully exploited. In such circumstances, there is a challenging issue of how to extract/discover valuable knowledge from the available large volumes of maritime traffic information usually stored in a database (DB).

The process of Knowledge Discovery from Databases (KDD) has been defined in [2] as an interactive and iterative nontrivial process which includes planning, data integration, selection of target data, data cleaning and pre-processing, data reduction and transformation, selection of suitable data mining techniques to support the discovery process, and evaluation, presentation and interpretation of results, a subset of which may be considered as new knowledge. Within the overall KDD process, data mining is viewed as the sub-process concerned with the discovery of hidden information. Applying data mining techniques to large sets of marine traffic data to extract knowledge will facilitate vessel traffic analysis and management as well as improved decision-making in maritime domain. Since maritime traffic data differs from the data commonly mined in business domains [3], the selection of appropriate data mining tools is crucial for meaningful knowledge extraction.

This paper presents our research efforts towards exploration of data mining technologies applicable to maritime situational awareness problems. The fusion of the mined information is the topic of a separate research effort. Here, the focus is on open source data mining tools, while the data are restricted to spatio-temporal maritime traffic data, such as AIS data. The paper provides a review of R as a data mining tool with potential use by maritime analysts, and includes example applications relevant to maritime traffic analysis. The assessment is performed using data stored in a DB containing spatio-temporal data.

The paper is organized as follows. Section II provides an introductory theoretical background on data mining, while Section III summarizes the most recent results concerning AIS data mining reported in the literature. In Section IV, mining for typical patterns in maritime traffic AIS data using R was described. In Section V the concluding remarks are presented.

II. DATA MINING

The most well-known definition of data mining, given by Frawley et al. (1991), in [4], defines data mining as a set of mechanisms and techniques, realized in software, used to extract implicit, previously unknown (or hidden) and potentially useful information from large databases. The word "hidden" in this definition is important; Standard Query Language (SQL) style querying, however sophisticated, is not considered as data mining. In addition, the term "information" should be interpreted in its widest sense. By the early 1990s, data mining was commonly recognized in computer science as a subprocess within the KDD. In the modern context of data mining, the term Knowledge Discovery in Data (as KDD) would be more apt, as we are no longer preoccupied solely by databases but also by non-tabular data (e.g. matrices, graphs, tensors, network data, etc.) [3]. Nowadays, there are plenty of data mining techniques for tabular data available, which are extensively performed by many commercial enterprises and researchers using software such as RapidMiner†, KNIME‡ or others, on standard desktop machines.

Most data mining techniques are heuristic, tailored to discover patterns of a generic type such as classes, associations, rules, clusters (the "large patterns") and outliers (the "small patterns") [5]. Since these techniques are heuristic, there is no single optimal algorithm for discovering patterns of a given type; different techniques highlight different aspects of the information space implied by the database at the expense of other characteristics.

†www.rapid-i.com
‡www.knime.org
Various pattern discovery techniques can be described more in detail as follows:

(i) **Rule pattern extraction/identification.** Rule pattern (or only “pattern”) recognition has been one of the primary goals of data mining (e.g. identifying purchasing/sales patterns, trends in temporal or longitudinal data, etc.). The association rule, as first proposed by Agrawal et al. (1993), [6] in the context of super market basket analysis, is the most well-known algorithm for rules mining, while the current most popular algorithm is the frequent pattern growth [7] by Han (2000).

(ii) **Data clustering.** Clustering is grouping of data into categories, a technique also used in machine learning. Coenen (2011), [3], also states that there is no “best” clustering algorithm applicable to all data; instead, for reasons that are not entirely clear, some algorithms work better on some data sets than others. The K-means algorithm [8] by MacQueen (1967), is one of the most commonly used algorithms if the number of clusters is known. Other clustering algorithms include classifications according to some proximity threshold, such as K-Nearest Neighbor (K-NN) [9] by Hastie and Tibshirani (1996), hierarchical clustering where the data are iteratively partitioned to form a set of clusters, such as BIRCH [10] by Zhang et al. (1996), and model-based clustering where a model is hypothesized for each of the clusters. Assuming that the data are generated by a mixture of underlying distributions, the idea is to optimize the fit between the data and the model, e.g. COBWEB algorithm [11] by Fisher (1987).

(iii) **Classification/Categorization** Classification involves building classifiers of data so as to categorize the data into classes. Unlike clustering, it requires pre-labeled training data from which the classifiers can be built. As such, classification in data mining is sometimes referred to as supervised learning while clustering is considered unsupervised learning. A typical classification algorithm in application to KDD is a decision tree classifier, such as the one described by Rastogi and Shim (1998) in [12].

(iv) **Outlier/Singularity/Anomaly detection/identification** This class of data mining algorithms makes possible the identification of rare events and exceptional cases, usually referred to as “small patterns”. For example, in vessel behavior analysis, this means identifying anomalies. This usually requires the development of a model representing normal vessel behavior. The anomalous behavior is then identified by the degree to which a vessel’s motion does not conform to that model normalcy. The subject of motion behavior analysis, however, is not unique to maritime surveillance. It has initially been studied in computer vision in the context of traffic monitoring, human activity monitoring, and unusual event detection. In the literature, there are various machine learning techniques used to generate normalcy models for analyzing vessel behavior and helping in the detection of anomalies from AIS data using Bayesian networks [13], kernel density estimation [14], Gaussian mixture models [15], support vector machines [16], neural networks, etc.

It is important to emphasize that a data mining process is interactive, iterative and exploratory. A standard data mining task represents itself as a KDD process, as defined in [2]. Specifically, it includes the following steps [17]:

1. **Setting the target:** Understanding the domain in which data are to be mined in terms of clearly describing the objectives, and listing the possible assumptions and anticipated desired results.
2. **Establishing the target data set:** Choosing the initial data set to be analyzed, e.g. AIS data set.
3. **Data pre-processing:** Using effective or readily available approaches to process noisy or incomplete data, e.g. decoding and amending AIS into DB or GIS processing of spatial information.
4. **Data cleaning and transformation:** Deleting or adding some attributes using standardization and/or conversion methods.
5. **Data mining:** Applying the most appropriate data mining algorithms to optimally process data by
   5.1 **Choosing the data mining task.** This involves selecting the generic type of pattern sought through data mining; this is the language for expressing facts in the database. Generic pattern types include classes, associations, rules, clusters, outliers and trends.
   5.2 **Choosing the data mining technique** for discovering patterns of the generic type selected in the previous step. Since data mining algorithms are often heuristic (due to scalability requirements), there are typically several techniques available for a given pattern type, with different techniques concentrating on different properties or possible relationships among the data objects.
   5.3 **Data mining.** Applying the data mining technique to search for interesting patterns.
6. **Explanation and evaluation:** Searching for useful and interesting information. If none, repeat the previous steps with other attributes and samples.
7. **Action:** If mined knowledge is found useful then it is integrated and applied to solve the appropriate problem, supporting decision making.

Each step requires expertise from a domain expert, a data analyst and a data miner. The results from data mining are usually presented in the form of concepts, rules, regularities, patterns, constraints and visualization.

### III. AIS DATA MINING

AIS data obtained from AIS transponders on-board ship/aircraft have an important role in littoral state monitoring. As one of the most important self-reporting maritime systems, which has been made compulsory by the IMO for most commercial ships, its use can be extended to sharing data between Vessel Traffic Service (VTS) and appropriate national administrations, gathering information on the presence and patterns of traffic, planning aids to navigation, legal evidence and accident investigation, search and rescue, risk analysis and
Generating statistics. AIS messages are automatically broadcast with a reporting frequency directly proportional to the speed of the vessel.

AIS information include:

- ship static information, programmed into the unit at commissioning,
- voyage related data, entered manually by the master through a password protected routine, and
- dynamic positioning data, derived from interfaces with the ship’s GPS and other sensors.

In this Section, we summarize the most recent reported research addressing the problem of knowledge discovery from AIS data.

AIS data are usually noisy and inconsistent (conflicting), therefore appropriate choices for data cleaning and pre-processing, and (motion) pattern mining algorithms are equally important in the discovery of meaningful information. While out-of-range errors and outliers can be processed/removed using some of the pre-processing methods described in the previous Section, detecting inconsistencies still remains a challenging issue. In [18], we have described an efficient method for verifying the consistency of the AIS sources. The method is based on calculating the consistency score using data from multiple sources such as websites and Maritime Domain Awareness (MDA)-related DBs, pertaining to a specific ship, thus also allowing to evaluate the source reliability.

A common AIS data mining task involves extraction and definition of motion patterns. In case of detecting anomalies in ship motion, an anomaly detection algorithm is subsequently applied. Definition of motion patterns can be quite complex if one deals with a network of origins/destinations with multiple connecting paths.

Assuming that the motion trajectories have already been extracted, Ristić et al. (2008) in [14] define a motion pattern by kinematic and attribute information, with only one compulsory attribute - its origin. Other useful attributes, if available, can be the vessel type, season of the year, etc. The kinematic information includes the ship location (in two-dimensions) and velocity (also in two-dimensions), while the origin of a motion pattern is defined by the location-velocity vector and its associated uncertainty ellipsoid. Ristić et al. also suggest that it is very useful for a pattern to contain the elapsed time information in the form of the time interval since the vessel was at the origin of the pattern, because the complexity about the topology of the network of paths is then eliminated. However, no information on how the motion patterns have been extracted is provided.

AIS reports are timestamped, but some traditional data mining techniques lose the time stamp and represent a navigation trajectory as a set, rather than a sequence, of consecutive latitude-longitude vessel positions. To include the time stamp information in the process of maritime traffic modeling, Bruno and Appice (2011) in [19] applied a multi-relational data mining method called SPADA where relational patterns (i.e. patterns which may involve several relations at once) and association rules are discovered from a relational DB in which data are stored. The vessel data and AIS data are modeled as distinct relational data tables (one for each data type) which helps distinguish between the reference objects of analysis (vessel data) and the task-relevant objects (AIS data), and also helps represent their interactions. The modeled interactions also include the total temporal order over AIS reports for the same vessel and interesting associations between a vessel (reference objects) and a navigation trajectory. Each navigation trajectory represents a spatio-temporal pattern obtained by tracing the subsequent AIS reports (task-relevant objects) of vessels.

In [20] Oo et al. (2004) present a fuzzy inference based model for identifying congested zones by investigating the current varying maritime traffic speed. The model is based on the improved DBSCAN [5] algorithm by using the neighborhood three models. In [13] Mascaro et al. (2010), after the pre-processing that involves cleaning and separating the AIS data into tracks, applied the machine learner CaMML on AIS data combined with additional real world data, and used this to produce the two networks time scales, in the form of the time series and track summary models. By adding some real world attributes the normalcy model is improved, while weather variables proved to have no impact on vessel behavior.

In [21], a data mining platform based on AIS data (ADMP) has been proposed by Tang and Shao (2009), and which produces the eigenvalue of marine traffic using clustering and statistics. The ADMP offers basic data support for data mining, marine traffic flow forecast and development and programming of marine traffic engineering. In [22], Zhu F. (2011), applies Agrawal’s association rule [6] in mining ship trajectory patterns. Another approach to mining vessel traffic flow data has been proposed in [23] by Zheng et al. (2008). It uses K-Means clustering algorithm [8] from WEKA data mining tool\(^3\) to extract multi-factor related regulations according to which clusters were generated. The considered factors include hour, direction, tonnage and ship type.

Tsou (2010) in [24], presented a framework similar to Business Intelligence (BI) discovery, and applied data mining techniques used in that framework to the processing and analysis of AIS data. The difference between the two applications is that information received by AIS includes spatial and temporal features. In order to distinguish these features and to extract the relevant knowledge, the AIS received data are first decoded and converted to a readable format. The DB management system is used for storage and management, while a GIS is used to analyze and process the information, converting text data to meaningful spatial and temporal data. This data are then warehoused, while the GIS and BI analysis tools are subsequently used to perform visual, spatial and temporal data mining to discover the status and regularities of maritime traffic flow. ArcGIS is used to perform visualization data mining, and SQL Server 2005 BI module’s association rule mining and sequential pattern mining algorithms to perform analysis. The association rules for interpreting AIS data as retail market data can be found in [24].

Although not a typical data mining technique, a query based approach based on ESRI ArcGIS technology to analyzing AIS data for improved maritime awareness, has been proposed by Ou and Zhu J. (2008) in [25]. The relevant

\(^3\)www.weka.net.nz
statistics are generated querying the AIS DB which consists of a main target table containing set of attributes collected by AIS and three look-up tables of interest.

IV. AIS DATA MINING WITH R

As one aspect of our efforts in the exploration of data mining technologies for improved maritime domain awareness, in this Section we present the evaluation of R software in application to mining of typical patterns in maritime traffic AIS data.

R is an open source software environment for statistical computing and graphics, having large extensibility through user-created packages available from Comprehensive R Archive Network (CRAN), [26]. There is a vast collection of R packages and functions, specifically made for data mining. These include packages for clustering (fpc, cluster, pvlust, mlust), which contain common clustering algorithms such as k-means, hierarchical clustering, DBSCAN as well as packages for plotting cluster solutions. There are also classification packages (rpart, party, tree, etc.), which contain decision tree, regression and survival analysis algorithms, while the association rules and frequent itemsets packages (arules, drm) contain algorithms for finding frequent itemset, association rules (e.g. APRIORI algorithm). Other available packages include sequential patterns package (arulesSequences), time series, statistics, and graphics as well as various data manipulation packages; and the interface to WEKA mining tool. There exist several Graphical User Interfaces (GUIs) to interface with R, among which Rattle is specifically designed for data mining.

The work flow of data mining with R typically comprises importing the required packages, importing data, transforming data to a convenient format for analysis, using the data mining/statistical functions and/or Rattle, and visualizing, validating and exporting results.

For this study, the data have been loaded from Comma Separated Values (CSV) file formats and directly from the DB which is managed by PostgreSQL. Importing data from a CSV file (or a text file) requires no package installation, while importing data directly from the DB requires installing the packages DBI and RPostgreSQL from CRAN website.

The output of the data import is a data frame, an R data object that is a type of table in which the rows are observations and the columns are variables. For the data manipulation and pre-processing, R allows operations on array-based objects which significantly simplifies all array-based data object manipulation by imposing no requirements on looping on array dimensions.

Although designed to support data mining tasks, Rattle proved to have limited data mining capabilities in terms of GUI, supporting only a subset of data mining algorithms and fewer options for data mining output analysis (e.g. it is not possible to transform the output of a clustering algorithm to use it as the input to another algorithm). Therefore, only the native command line interpreter has been used in the pre-processing (i.e. removing outliers and out-of-range-values) and data mining steps, while Rattle has been used for visualization and export.

Next, we describe the AIS data mining tasks, using R to find the patterns of interest to maritime traffic analysis.

A. Mining Invalid Observations

With large amounts of AIS data stored in a DB, a common question of interest defined as a data mining task is:

- If an out-of-range error (value) is identified in one field, is it possible to determine if other fields are likely to also contain an out-of-range error?

For example, when a vessel heading field is identified as invalid, if there is a greater likelihood that other fields such as latitude or speed, are also invalid?

1) Data Set: The considered data set contained AIS data extracted from the DB, which covers the period from 3-10 November, 2011, reported by the exactEarth monitoring system\(^2\). The data attributes which have been considered for the analysis are: Maritime Mobile Service Identity (MMSI), latitude, longitude, IMO number, ship dimensions (to bow, port, stern and starboard), ship type, Estimated Time of Arrival (ETA), draught, rate of turn, speed over ground, course over ground and heading.

This data set contains a total of 4 321 563 complete reports\(^3\). The DB also contains a data quality flag identifying reports with out-of-range attribute values. This field was used to extract the data from the DB using R commands, totaling 75 601 reports in the data frame. Otherwise, out-of-range values for AIS data can be determined explicitly, by verifying if the value of the field is in the specified range according to the AIS specifications, [27]. The specifications describe for each field the range of possible values, (e.g. heading, available from messages of type 1, 2, 3, 18 and 19, takes values between 0 and 359 and has the value 511 when is not available).

2) Data Mining Approach: The idea here is to explore the co-frequency of occurrence of invalid values. For each pair of fields \((A, B)\), the frequency of invalid value occurrence for field A when field B has an invalid value, \(f(A|B)\), is computed. \(f(A|B)\) represents the conditional sum of all out-of-range values for field A across all reports in the data set, normalized by the number of invalid values for field B. It is computed as,

\[
f(A|B) = \sum_{i=1}^{N} \delta(A_i) \delta(B_i),
\]

where \(N\) is the number of reports in the data set and

\[
\delta(A_i) = \begin{cases} 
1, & \text{report } i \text{ as an invalid value in field } A, \\
0, & \text{otherwise.}
\end{cases}
\]

The data have to be transformed in order to obtain the data frame populated by \(\delta(X_i)\) values as defined in (2), i.e. each field with out-of-range value/invalid value has to be replaced with 1 and each field with a valid value with 0. This is done by encoding the out-of-range and invalid values specifications in R.

\(^2\)http://www.exactearth.com/

\(^3\)In the DB, each report corresponds to an AIS message, where a complete report refers to an AIS message that has been successfully parsed.
Although well suited for this data mining task where the out-of-range values are known, this approach may not be considered as a typical data mining technique. In the case where out-of-range values are unknown, e.g. for a new data type with unknown specifications, outlier detection algorithms can be used. Based on the assumption that invalid values are outliers, i.e. observations significantly distant from other data, these methods detect the out-of-range values for each fields. Once these values are identified, the same methodology as described above can be used. The R packages containing algorithms for outlier detection, outliers and extReves, the latter having a GUI, proved to be robust for the data set used.

3) Results: Table I summarizes the results for the fields where invalid values were found in the data, showing their corresponding value of $f(A|B)$. It is read from rows to columns, e.g. $f(\text{Heading}|\text{Lon}) = 96.39\%$ signifies that when an invalid longitude value is observed in the data set, 96.39\% of the time it is also observed in the heading field. Table II contains the total count and proportion of invalid values in the complete data set for each field.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Lon</th>
<th>Heading</th>
<th>Course</th>
<th>Ship Type</th>
<th>MMSI</th>
<th>IMO</th>
<th>ETA Hour</th>
<th>ETA Month</th>
<th>IMO Month</th>
<th>ETA Hour</th>
<th>ETA Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lon</td>
<td>-</td>
<td>2.46</td>
<td>7.94</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heading</td>
<td>96.39</td>
<td>23.81</td>
<td>65.14</td>
<td>0.29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Course</td>
<td>0.56</td>
<td>0.04</td>
<td>14.14</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ship Type</td>
<td>0.56</td>
<td>2.62</td>
<td>6.35</td>
<td>0.56</td>
<td>2.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MMSI</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IMO</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETA Hour</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETA Month</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE II. TOTAL COUNT AND PROPORTION OF INVALID VALUES IN THE COMPLETE DATA SET FOR EACH FIELD IN WHICH INVALID VALUES WERE FOUND.

B. Mining Ship Trajectories

The following patterns are set as data mining targets:

- **Identifying all ships that transited from port A to port B, for a time period** $T$, and
- **If a ship is spotted in port A, in what port is it the most likely to be transiting to?**

under the assumption that if a ship is observed in port A at an earlier time than it is observed in port B then there is a route between port A and port B.

1) Data Set: The reports data set consists of AIS data extracted from the DB, covering the period from 3-10 November, 2011, limited to the bounding box defined by latitude values between $35^\circ N$ and $55^\circ N$ and longitude between $80^\circ W$ and $48^\circ W$ (i.e. the Atlantic coast from Newfoundland and Labrador, Canada down to Virginia, USA), totaling 93 000 complete reports. The AIS attributes contact ID, MMSI, longitude, latitude, report time stamp, and data quality were considered.

2) Data Pre-processing and Transformation: In order to know if a ship has visited a port, it is required to compute the distance between the port’s location and the ship’s reported positions. Computing the distance has been done by using a predefined grid covering the region of interest, where each position (for ports and contacts) is associated with a grid cell. If a ship has transited to a port, it thus has visited the cell associated with that port. Using grid cells is a common way to reduce complexity and speed up computation. The grid and map positions to cells were created using PostGIS. This operation was performed on the DB, with cell size $0.25^\circ \times 0.25^\circ$, after which the data were imported as two data sets using R: one containing the contacts with the associated cell IDs, and the other containing the ports, their position and cell IDs. It is also possible to build and manipulate geographical grid using R’s package gdistance. Eighteen ports were included in the data set. Since only MMSIs and cell IDs are required for this analysis, other fields as well as all duplicate entries are removed from the data frame. A duplicate entry is defined by a ship report in the same cell for more than one consecutive time. Only the unique cells visited by each ship are considered. The resulting data frame is shown in Table III.

<table>
<thead>
<tr>
<th>MMSI</th>
<th>cell id</th>
</tr>
</thead>
<tbody>
<tr>
<td>211205790</td>
<td>1665</td>
</tr>
<tr>
<td>210938000</td>
<td>2649</td>
</tr>
<tr>
<td>210938000</td>
<td>2957</td>
</tr>
<tr>
<td>210938000</td>
<td>2853</td>
</tr>
<tr>
<td>210938000</td>
<td>2748</td>
</tr>
<tr>
<td>212226800</td>
<td>1046</td>
</tr>
<tr>
<td>212226800</td>
<td>1048</td>
</tr>
</tbody>
</table>

TABLE III. EXCERPT OF THE DATA FRAME USED FOR THE PORT TRANSITING ANALYSIS.

3) Data Mining Approach: The first data mining task concerns identifying all ships that transited from port X to port Y, for a time period $T$, i.e. all ships that passed through the grid cell associated with port X and through the grid cell associated with port Y, after time $T$.

Consider the data frame containing only contacts from a given period $T$ and ordered by MMSI and time (as illustrated in Table III). The filtering on time and ordering can be done on DB using SQL and also with R using the Date class. The R Date class allows the transformation from a string to a Date object. Then the filtering on e.g. time can be done as

```r
> mydata_filtered
  <- mydata[mydata$mydate %in% as.Date(c(‘2012-01-05’, ‘2012-01-09’))]
```

Obtaining only MMSIs associated with the cells corresponding to ports X and Y, while ensuring ordering in time, i.e. that port X was visited before Y, can be done in the following way:

```r
> visitors <- subset(reports, reports$cell_id=="1597"
  | reports$cell_id=="2538")
> visitors[ which(visitors$cell_id=="2538"
  | visitors$cell_id=="1597")
```
In this case, the data set contains the reports’ information as shown in Table III, while the port of Montreal (cell ID = 1597) and the port of Quebec (cell ID = 2538) are considered. From the given data set, this produces the result that 26 ships transited from Montreal to Quebec and 22 vice versa. Only ships with two or more contacts are considered. There are 1029 such ships found in the data set, after the data manipulation.

The second data mining task concerns identifying most probable links between ports from the data set, i.e. discovering if a ship visits port X how likely it is that it will also transit to port Y.

The data mining approach here is to discover route behaviors by discovering the high traffic density grid cells. This is similar to discovering consumers’ behavior and discovering popular products in order to create a marketing strategy. Each customer’s basic information is equivalent to the static message, the transaction data to route related message; ship’s MMSI to customer’s number; and every product list in a transaction equivalent to grid cells passed along the route. The most popular approach to discovering hidden patterns in these situations is association rule mining. Having AIS data and route related messages, we can assume that $I = \{i \subset 1, i \subset 2 \cdots i \subset m\}$ represent all grid cells in the area, called itemsets, and every route $R$ is a subset of all itemsets, i.e. $R \subset I$. The association rule $Ship$ with some kind of profile who passes region $A$ will also pass region $B$ can then be formally written as:

$$A \rightarrow B [support, confidence]$$

$$\text{support} = P(AB)$$

$$\text{confidence} = P(AB)/P(B)$$

where $A \subset I, B \subset I$ and $A \cap B = 0$.

A classic algorithm for association rule mining is APRIORI [6]. It is a level-wise, breadth-first algorithm which counts transactions to find frequent itemsets and then derive association rules from frequent itemsets. An implementation of it is the function `apriori()` in R’s package `arules`. Another algorithm for association rule mining is the ECLAT algorithm [28], which finds frequent itemsets with equivalence classes, depth-first search and set intersection instead of counting. It is implemented as function `eclat()` in the same package.

In this paper, we demonstrate association rule mining with `apriori()` for which the default settings are: $\text{supp} = 0.1$, which is the minimum support of rules; $\text{conf} = 0.8$, which is the minimum confidence of rules; and $\text{maxlen} = 10$, which is the maximum length of rules. The smaller the support and confidence values, the higher the number of produced rules.

The use of `arules` association algorithm requires additional data manipulation due to its input requirements in the form of an incidence (or frequency) matrix. In our case, it is a matrix having MMSIs as rows and cell IDs as columns, (e.g. there is a 1 in the matrix cell (967191190, 1597) if the ship with MMSI 967191190 has a contact within the cell of ID 1597, and 0 otherwise). Since all duplicates have been removed, the matrix is only populated by zeros and ones.

With a confidence of 0.1, the support of 0.01 and rules of maximum length, the total of 2 827 rules are produced. From this set of rules, only rules that imply the list of ports of interest need to be considered.

Results are summarized in Table IV, ordered by lift and confidence. The first rule means that if a ship is spotted in the port of Philadelphia, PA and is transiting to another port, then the subsequent port is most likely to be Marcus Hook, PA. The `arulesViz` package allows you to visualize rules in various forms (e.g matrix, graph, etc.). Figure 1 is a graph representation of the rules listed in Table IV, where Marcus Hook corresponds to cell ID 742, Philadelphia to cell ID 847, New York to cell ID 1370, New Haven to cell ID 1788, Montreal to cell ID 1597, Quebec to cell ID 2538, Baltimore to cell ID 419, and Hampton to cell ID 324. Note that if the minimum number of contacts for each ship increases, the rules change. For example, if we consider only ships with at least 20 contacts, which decreases the total number of ships for the analysis to 153, the rules become:

<table>
<thead>
<tr>
<th>If</th>
<th>Then</th>
<th>Lift</th>
<th>Confidence</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philadelphia, PA</td>
<td>Marcus Hook, PA</td>
<td>9.615</td>
<td>0.585</td>
<td>0.03109</td>
</tr>
<tr>
<td>Marcus Hook, PA</td>
<td>Philadelphia, PA</td>
<td>9.615</td>
<td>0.511</td>
<td>0.03109</td>
</tr>
<tr>
<td>Quebec, QC</td>
<td>Montreal, QC</td>
<td>5.147</td>
<td>0.333</td>
<td>0.0117</td>
</tr>
<tr>
<td>Montreal, QC</td>
<td>Quebec, QC</td>
<td>5.147</td>
<td>0.180</td>
<td>0.0117</td>
</tr>
<tr>
<td>New Haven, CT</td>
<td>New York, NY</td>
<td>4.536</td>
<td>0.652</td>
<td>0.0194</td>
</tr>
<tr>
<td>New York, NY</td>
<td>New Haven, CT</td>
<td>4.536</td>
<td>0.138</td>
<td>0.0194</td>
</tr>
<tr>
<td>Baltimore, MD</td>
<td>Hampton, VA</td>
<td>2.853</td>
<td>0.243</td>
<td>0.0129</td>
</tr>
<tr>
<td>Hampton, VA</td>
<td>Baltimore, MD</td>
<td>2.853</td>
<td>0.152</td>
<td>0.0129</td>
</tr>
<tr>
<td>Philadelphia, PA</td>
<td>New York, NY</td>
<td>1.866</td>
<td>0.268</td>
<td>0.0142</td>
</tr>
</tbody>
</table>

**TABLE IV. PORT TRANSITING RULES AS COMPUTED BY APRIORI ALGORITHM.**
Canadian University of Technology, the Department of National Defence, and the International Telecommunication Union for their valuable remarks.

the scenarios and Francine Desharnais, both from the DRDC-Atlantic, for their valuable remarks.

the work was provided by the DRDC Agility Fund. The funding for this work was supported by Defence R&D Canada-Atlantic, under the contract No. W7707-11537. Funding for the work was provided by the DRDC Agility Fund. The authors would also like to thank Anthony Isenor for proposing the scenarios and Francine Desharnais, both from the DRDC-Atlantic, for their valuable remarks.

V. CONCLUSION

In this paper, AIS data are used in a maritime traffic data mining analysis that reveals typical patterns relevant to maritime environment. The out-of-range errors and ship trajectories were selected as relevant patterns of interest. The assessment shows that the software application R has generally good data mining capabilities as well as promising spatio-temporal data mining capabilities. Result validation must be performed by the user by using visualization, statistical tests or external validated results.

R allows connecting to and retrieving data sets from major DBMS. However, there is no SQL GUI. All interactions with a RDBMS have to be in explicitly written in SQL. R is able to handle large data sets and extensive queries having no memory issues. Rattle showed limitations that may be critical for a custom data mining process in both considered data mining examples.

Possible future work regarding ship trajectories mining could make use of many R functions and packages available for time series decomposition and forecasting. Since there are no R functions or packages for time series classification and clustering implemented yet, to do time series classification, in [29] it is suggested to extract and build features first, and then apply existing classification techniques, such as SVM, k-NN, neural networks, regression and decision trees, to the feature set. For time series clustering, one needs to define distance or similarity metrics, and then use existing clustering techniques, such as k-means or hierarchical clustering in order to find the clustering structure.

Finally, most maritime traffic operators would not have the time, background, or resources to work through the complete KDD/DM process if they had to start from the beginning each time they require situational awareness. It is expected that once a domain expert, data analyst, and data miner have iterated through the KDD process they will have discovered some analysis that are repeatable. For an optimal maritime traffic data analysis, it is desirable to have an application with a list of pre-defined analysis and specific variables that the operator can change through a wrapper GUI such as time-period, geographic area, or class of ship. An operator-oriented application would also compensate for the GUI limitations of R/Rattle.

ACKNOWLEDGMENT

This work was supported by Defence R&D Canada-Atlantic, under the contract No. W7707-11537. Funding for the work was provided by the DRDC Agility Fund. The authors would also like to thank Anthony Isenor for proposing the scenarios and Francine Desharnais, both from the DRDC-Atlantic, for their valuable remarks.

REFERENCES


