Anomaly detection in Multivariate Temporal Data for Vessels
Abnormal Behaviour Detection

Rui Maia
Cláudia Antunes

Follow this and additional works at: https://aisel.aisnet.org/capsi2019

This material is brought to you by the Portugal (CAPSI) at AIS Electronic Library (AISeL). It has been accepted for inclusion in CAPSI 2019 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Anomaly detection in Multivariate Temporal Data for Vessels Abnormal Behaviour Detection

Rui Maia, Instituto Superior Técnico, Portugal, rui.maia@tecnico.ulisboa.pt, INOV – Inovação. rui.maia@inov.pt
Cláudia Antunes, Instituto Superior Técnico, Portugal, claudia.antunes@tecnico.ulisboa.pt

Abstract

The growing number of deployed data mining systems leverage the interest in temporal data anomaly detection. From cyber-security or finance to heart-diseases detection, unexpected data often incorporate critical information that must be analysed. Data anomalies have long been studied from an univariate perspective where only one data dimension changes over time. Few works have been dedicated to multivariate anomaly detection. In this work we provide a comprehensive and structured analysis of the main definitions, state-of-art methods and approaches focusing multivariate temporal data anomaly detection. Our research focus on dealing with variable length data series with millions of samples and multiple feature categories, either static or dynamic, real or categorical valued. We describe a case-study in the maritime domain investigating the unusual spatio-temporal behaviour of commercial vessels and experiment over two open datasets and one got from the MARISA H2020 Project.

Keywords: Anomaly; Multivariate; Time-Series; Outlier; Detection

1. INTRODUCTION

Anomaly detection applications aim for the identification of unusual observations or series of observations considered to be crucial for the understanding of specific events. From bio-informatics to finance or engineering, anomalies have long been studied (Fox, 1972). Most research works are based on univariate approaches analysing each data dimension independently from others. This perspective still prevails nowadays (Shokoohi-Yekta, Hu, Jin, Wang, & Keogh, 2017). Real world temporal data is by the contrary of complex nature, being typically noisy, non-stationary (statistical distribution might change with time) or having different scales, for example. Most statistical or machine learning approaches have only narrow support for such dynamic and multi-faceted challenge (Gupta, Gao, Aggarwal, & Han, 2014).

1 The Maritime Integrated Surveillance and Awareness (MARISA) Project is funded by the European Comission through the European Union’s Horizon 2020 research and innovation programme under grant agreement N. 740698.
Anomaly detection approaches are heavily dependent on the size and characteristics of the datasets (Chandola, Banerjee, & Kumar, 2009b) (Aggarwal, 2015). However, other multivariate temporal data complexities should be taken into account when defining this research problem:

- It is frequently assumed that the major part of datasets observations behaves normally, and just a small subset are anomalies. Although this may not always be the case. Domain knowledge is typically involved supporting preliminary assumptions about observation and dimension values distribution.
- Most time dependent datasets are unlabelled meaning that observations are not labelled as normal or abnormal. Being so, most of the available datasets do not directly support supervised learning approaches rather they support unsupervised ones.
- Anomaly detection methods have to deal with resource intensive computational tasks specifically regarding contexts where data is being continuously captured by multiple sensors.
- Temporal data may require online processing methods considering that anomaly detection is frequently associated to critical systems faults, life support systems or cyber threat detection, for example.
- Differences between observation novelties, anomalies and noise can be hard to define or distinguish either conceptually and computationally.

Regarding maritime context, specifically the vessel domain, anomalies may indicate irregular vessel activity. Therefore, it is relevant to anticipate such behaviours – the anomalies - to support informed decision making and to improve the maritime situational awareness. The Maritime Integrated Surveillance Awareness (MARISA) H2020 Project is aiming for this challenge involving multiple industrial, academic and military partners, and is one of the dataset sources of this research, supporting also the validation phase through the involvement of military domain specialists.

2. ANOMALY DETECTION GENERIC FRAMEWORK

The main goal of an anomaly detection method is to identify a typically small group of abnormal sequences of observations considering as a preliminary assumption that most of the samples are normal (Gupta et al., 2014). For that, most anomaly detection methods can be divided in two main steps: (1) model building and (2) anomaly identification. Regarding the (1) first step, a model can

---

2 Marisa - Maritime Integrated Surveillance Awareness is a project funded by the European Comission. This project has received funding from the European Union’s Horizon 2020, research and innovation programme, under grant agreement N. 740698.
be trained and built using previously labelled data - enabling *supervised* learning approaches - although the chance is that it will be built upon unlabelled or only partially labelled data, allowing either *unsupervised* or *semi-supervised* approaches. The (2) second step is heavily dependent on the choice of the approach type (made in the previous step).

Approaches are also characterized by their ability to focus on the complete sequences (or series) or sub-sequences of multivariate temporal observations. Anomaly detection procedures can therefore need to split data onto multiple sub-sequences also found as *windows*, *motifs* or *fragments*.

Besides the complexity of anomaly detection in multivariate time dependent contexts, additional challenges must be dealt with when defining research approaches: multivariate temporal data including a spatial dimension is commonly captured using different and distributed sensors which affects series sizes or rates. For example, sensors installed in vessels or cars, their properties and capabilities, their data generation methods or transmission failures, all introduce additional dynamic complexities that should to be taken into account when dealing with anomaly detection.

Several research fields have been dedicated to anomaly detection, from statistics to machine learning, sharing high level approach steps. Most of them can be mapped into a three steps meta architecture: (1) *data preparation*, to clean the dataset or to normalize observations and series; (2) *model construction* describing the normal or expected behaviour of the series, and (3) finally the *anomaly detection* process considering unseen observations. Figure 1 illustrates this generic architecture.

![Figure 1 – Multivariate Time-Series anomaly detection generic framework.](image-url)
3. DATA PREPARATION IN MARITIME VESSEL DOMAIN

The anomaly detection process starts with input data preparation, frequently including normalization and cleaning. Regarding spatio-temporal data series - the scope of this work - positions measured in Latitude and Longitude have decimal representations such as 38.898556, which is typically normalized to enable geographic positions comparison and enhance processing efficiency. More specifically, positions frequently rely on Global Positioning System (GPS) having a variable accuracy depending on multiple factors such as satellites geometry, signal blockage or atmospheric conditions. Considering that 1 degree of Latitude is equivalent to 111,111 meters (at Equator), then the sixth decimal place in one degree represents approximately 11 centimetres. There is room for lowering values precision to the fourth decimal digit while maintaining real accuracy.

Another context where data normalization takes its place is on the distributed sensors scenario. Communication failures (Gupta et al., 2014) affect samples existence or quality, attracting researchers to data cleaning aiming for removing erroneous observations as they can affect model training depending on the recurrence and type of failures.

After the preparation phase, a model representing the normal behaviour is acquired based on clean data. This will be finally used in classification or scoring phase over test data - usually a subset of the main dataset - allowing each unseen sample to be marked as normal or abnormal (or either scored). Approaches that calculate scores typically compare these against a threshold that represents a boundary between a normal or abnormal behaviour.

Univariate and multivariate anomaly detection are intrinsically connected, pulling authors to adapt univariate based approaches to the multivariate case, for example calculating anomaly scores of each different feature or dimension, and finally computing the final anomaly score for the multivariate series. Other authors prefer to test only a subset of multivariate series features. Both perspectives are limited and suggest that multivariate temporal data anomaly detection has a full set of open challenges.

4. MACHINE LEARNING MODES

While anomaly detection approaches are heavily shaped by input dataset properties, the existence of labelled data - where a target variable or set of variables is associated with each input sample - defines if the model can be trained and built in supervised, unsupervised or semi-supervised mode. A machine learning approach extending these perspectives is reinforcement learning, although, this will not be covered in this study since it was not possible to find published papers on multivariate anomaly detection in the maritime domain supported by this approach. Figure 2 illustrates the three approach categories analysed in this work.
Figure 2 – Approach categories and model learning differences.

- **Unsupervised Learning** - Approaches based on unsupervised learning do not require labelled data to define, train and build their models. To overcome the lack of information regarding which samples are abnormal, these approaches frequently assume that the majority of samples are normal, usually requiring the definition of normal and abnormal samples proportion or percentage. Unsupervised approaches mostly use statistical, density or distance-based calculations to define what is the normal behaviour of the series, and what is abnormal.

- **Semi-supervised Learning** - In this learning mode, the train dataset only contains normal samples. The obtained model is not contaminated by abnormal information and consequently can detect an abnormal sample whenever the deviation from the base model is above a certain threshold. These are typically found as "one-class" models. One of the major issues of semi-supervised approaches is that it does not distinguish between novelty and anomaly, where the former is a new sample behaviour that should influence model base definition (taking it to an update). By the contrary, an anomaly is a sample associated with a faulty or not acceptable behaviour that should be identified but not influence the base model definition.

- **Supervised Learning** - Training a model in supervised mode requires a labelled dataset for normal and abnormal classes. The model is then used to predict if an unseen input sample belongs to either normal or abnormal class. The training of such a model involves a major issue: the number of abnormal samples in a dataset is typically significantly lower than the
normal ones, which turns it difficult to build a model that can classify accurately both normal and abnormal samples. Moreover, supervised approaches, such as decision trees, do not deal well with these highly unbalanced datasets. To overcome the lack of abnormal labelled samples in datasets, researchers tend to generate new abnormal ones, which is most of the times an ad-hoc process, highly dependent from the application domain.

5. Anomaly Definition and Types

The definition of anomaly - as a class - has been revisited since 1969, when Frank Grubbs (Grubbs, 1969) identified it as unusual behaviours of isolated samples or groups of samples. Anomalies can also be frequently found as outliers, exceptions, faults or discordant observations, depending on the research field and application domain. (Chandola et al., 2009b) defined anomaly as a pattern that do not follows an expected behaviour. This generic definition can be intuitively illustrated in one, two- or three-dimensional spaces, as illustrated in Figures 3. Anomalies occurring in larger dimensional spaces will escape human visualization abilities, resulting in human intuition loss.

Figure 3 – Two- and three-dimensional anomalies identified by red crosses.

(Agarwal & Yu, 2001) defined anomaly as an observation which is very different from the dataset behaviour considering a specific measure calculation. This work adopts the base definitions from (Chandola et al., 2009a) and (Aggarwal & Yu, 2001), considering their complementarity.

5.1. Anomaly Types and Maritime Traffic Practical Examples

The anomaly detection overview by (Chandola et al., 2009a) identified three main types of anomalies: Observation, Context, Sequence and Collective.

- **Observation Anomaly** - An observation is considered abnormal when one or more-dimensional values - categorical or real valued - are outside the expected intervals or sets (as illustrated in Figure 3). This is the simplest type of multivariate temporal anomaly and
derives directly from the univariate concept. Depending on the identified abnormal values this can be a *unidimensional* or *multidimensional anomaly*.

**Practical Example** - A vessel track may be considered abnormal if the registered speed is very high compared to the average vessel speed, for that specific type of vessel. If all other parameters are considered as normal, this would be an *unidimensional observation anomaly*.

- **Contextual Anomaly** - An observation can be considered abnormal despite all dimension values are according the expected intervals or sets. (Song, Wu, Jermaine, & Ranka, 2007) introduced a statistical conditional anomaly detection approach by defining each observation as a pair of attribute sets \((x,y)\) where \(x\) is the set of contextual dimension values and \(y\) is the set of indicator dimensions values (see Figure 4). An observation is anomalous if \(y\) values are not expected considering \(x\) values. Typically, multivariate temporal data dimensions are not classified as Contextual or Indicator, and this classification is not easy to define (Chandola et al., 2009a) as it strongly depends on the problem domain.

![Figure 4 – Conditional Anomaly Detection as presented by (Song et al., 2007)](image)

**Practical Example** - Contextual dimensions can be vessel latitude and longitude in a spatial dataset or transmission time in a time dependent domain. The vessel speed can be the dimensional indicator. This anomaly type relies on the separation between contextual and indicator dimensions which in turn essentially depends on the application domain.

- **Sequence Anomaly** - A sequence of observations is considered abnormal when, even if each observation dimension values are normal, the relation between observations is unexpected as illustrated on Figure 5.

![Figure 5 – Vessel speed sequence anomaly example](image)
Practical Example - None of the vessel speed observations in Figure 5 can be considered abnormal by itself. This anomaly depends on the existence of a dimension that relates observations, for example time, in multivariate temporal data. (Chandola et al., 2009a) referring (Song et al., 2007) definition, noticed that sequence anomalies can be observed as context anomalies if the context is defined by one or more time related features.

- **Collective Anomaly** - Searching for collective anomalies involves exploring the hidden structure of data for abnormal relations between observations or sequences of observations. Each multivariate series behaviour is not necessarily an individual, neither sequence nor contextual anomaly. In fact, there will be no series anomalies, but a change in the dataset behaviour, considering multiples series in the same temporal reference. It is common to find collective anomalies being defined as previously sequence or contextual anomalies, but this would exclude collective anomalies as defined in this work, i.e. from the collective behaviour perspective.

Practical Example - Two vessels behaving normally from observation and sequence perspectives, can change the relation between them. If they agree to exchange illegal products during their normal missions, they can still maintain normal track parameters (such as speed or GPS position, for example) and they will also behave normally according to their seasonal or temporal context, the admitted regions and the expected route. Although, these vessels might have a change in the relation between their temporal series normal behaviour, which might indicate they are orchestrating irregular activities.

5.2. **Anomaly Detection Outputs**

In multivariate temporal data, like in the univariate case, an anomaly detection process tries to identify the unexpected observation or sequences of observations. (Chandola et al., 2009a) identified two output types: labelling and scoring.

- **Label** - Anomaly detection approaches classify one observation or sequence as normal or anomaly. It is a binary classification process taking generally less computational complexity (Ahmed, Mahmood, & Hu, 2016) than the score calculation.

- **Score** - An anomaly score is calculated for each observation or sequence. This value can be used to rank observations and sequences according to their odds of being abnormal while it can be compared against a threshold that distinguishes between normal and abnormal values.
6. Preliminary Analysis

This research's preliminary analysis on vessels abnormal behaviour detection is observed on three different levels:

- **Point-wise anomaly detection** - Targets abnormal multivariate samples detection. Each sample is considered as non-related to any other, even in the case of having the same origin (ex: sensor, vessel, radar...).

- **Univariate series anomaly detection** - Focuses on experimenting different methods to point out sample sequences where the analysed dimension shows abnormal behaviour when compared with the base dataset. Regarding vessels, each temporal spatio-dimension - Latitude and Longitude - Speed or Heading, should be taken into account for traffic analysis. One of the dimensions can grow exponentially diverging from the normal behaviour while all other remain under normal parameters.

- **Multivariate series anomaly detection** - Deals with abnormal behaviour not detected in point-wise or univariate analysis. Still considering the vessels case-study, it may have registered normal individual samples and series, regarding each dimension separately. Although abnormal data may be detected when analysing all multiple dimensions, i.e., all series and the mutual relations. At this level, feature relations may be analysed using features like covariance and interference.

6.1. Datasets

Preliminary experiments were run over two commercial vessels transmission datasets made freely available by national authorities of Australia and United States of America (EUA). Both datasets register Automatic Identification System (AIS) transmissions distributed in two main areas (Figure 6 illustrates samples geographical distribution). The experiments were also run over the MARISA\(^3\) H2020 project dataset which is not publicly available at this moment.

---

\(^3\) Marisa - Maritime Integrated Surveillance Awareness is a project funded by the European Comission. This project has received funding from the European Union’s Horizon 2020, research and innovation programme, under grant agreement N. 740698.
A transmission is a multidimensional valued observation containing both real-valued continuous features and categorical features. Each one is associated to a moment in time. There is no information about abnormal samples or series. The characteristics of both datasets are summarized in the following Table.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Samples</th>
<th>Dimensions</th>
<th>Real-Valued Dimensions</th>
<th>Categorical Dimensions</th>
<th>Vessels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Australian Bight</td>
<td>79,082</td>
<td>11</td>
<td>Latitude, Longitude, Course over Ground (COG), Speed over Ground (SOG), Timestamp, Length, Beam, Draught</td>
<td>Vessel Id, Type, Sub-Type</td>
<td>27</td>
</tr>
<tr>
<td>USA (Zone 10)</td>
<td>279,773</td>
<td>8</td>
<td>Latitude, Longitude, Course over Ground (COURSE), Speed over Ground (SPPED), Timestamp</td>
<td>Vessel Id, Type, Status</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 1 – Characterization of datasets experimented in preliminary analysis

### 6.2. Preliminary Results

Considering the three levels of analysis introduced in Preliminary Analysis - multivariate observation, univariate temporal data series and multivariate temporal data series - an experimental framework was developed in order to test and compare different reference approaches. The preliminary study experimented the following approaches for multivariate observations:
- One-Class Support Vector Machines (SVMs) (Manevitz & Yousef, 2001) (Ma & Perkins, 2003) support two-class unsupervised classification even if class distributions are heavily imbalanced, which is the case of our domain.

- Local Outlier Factor (LOF) (Breunig, Kriegel, Ng, & Sander, 2000) is a method for finding anomalous observations by measuring the local deviation of a sample considering its neighbours. This method has been also successfully applied to time-series streams (Pokrajac, Lazarevic, & Latecki, 2007).

- k-Nearest Neighbours (k-NN) (Altman, 1992) is a foundation for different learning methods. k-NN iterates through all dataset positions to find a predefined number $k$ of neighbour’s type (outlier or inlier). For our preliminary experiments we used $k=10$.

- K-Means (MacQueen, 1967) is an unsupervised clustering learning algorithm applied to classification. The procedure associates each dataset sample to a cluster (one of $k$ predefined number of clusters). In outlier detection domain, two clusters will represent normal and abnormal samples.

Regarding univariate time-series experimentation, we started by using Dynamic Time Warping (DTW) and a Long Short-Term Memory (LSTM) Neural Network using some of the available features.

Dynamic Time-Warping (DTW) (Sakoe & Chiba, 1978) is a technique that finds the optimal alignment between two given time-dependent data series. By measuring the distance between time-series, and being tolerant to sample-time changes, DTW is applied to detect deformations with time-dependent data.

![Figure 7](image)

*Figure 7 – Ground truth is presented on the top left map. Others show DTW using different parameters.*
Figure 7 shows similar results found applying a DTW based approach for Latitude, or Longitude. An LSTM on LON parameter got significantly better results due to its ability for preserving vessel movement history.

![Figure 8](image)

Figure 8 – Ground truth is presented on the left map. The other shows LSTM results.

7. **Preliminary Analysis and Ongoing Research**

Real world multivariate temporal datasets frequently include different data types and scales, frequencies, noise or dynamic statistical distributions. These differences may be due to different sensors such as different sensor hardware and software versions in different vehicles or different contexts, such as the satellites used for GPS positions acquisition, weather conditions or seasonality effects. From categorical or real-valued variables, all can be affected, independently or in their correlations (Figure 9).

![Figure 9](image)

Figure 9 – Distribution of track sizes per vessel on the Australia AIS Dataset

The limitations of current state-of-the-art approaches have been identified, but still few anomaly detection research works focus on multivariate temporal data series, without significant need of preparation, such as data cleaning, imputation or feature engineering.

We experimented over some of the current state-of-the-art anomaly detection approaches, mainly focusing on density based and neural networks.

Density approaches on one hand, such as the ones based on Nearest Neighbours, and Neural Networks on the other, keeping an historical memory of input data – such as a Long Short-Term
Memory or Gated Recurrent Unit – based networks, may take to a complimentary or ensemble approach to the identified challenge.

The Long Short-Term Memory Neural Network had positive results confirming that it can learn the relation between different feature dimensions while preserving information on vessel movements, speed and course.

Next steps of this research will involve experiments over bigger datasets and evaluation of statistical significance of the results. It should be also considered a new experimentation phase over a different domain such as the Car Traffic.

ACKNOWLEDGEMENTS

The Maritime Integrated Surveillance Awareness (Marisa) H2020 European Project aims the development of new behavior analysis and anomaly detection methods in order to anticipate possible irregular activities, support Search and Rescue (SAR) operations and other civil and military operations. This project is funded by the European Commission through the European Union’s Horizon 2020 research and innovation programme under grant agreement N. 740698.

REFERENCES


