Unsupervised extraction of maritime patterns of life from Automatic Identification System data

Nicola Forti, Leonardo M. Millefiori, Paolo Braca
Research Department
NATO STO Centre for Maritime Research and Experimentation (CMRE)
La Spezia, Italy
{nicola.forti,leonardo.millefiori,paolo.braca}@cmre.nato.int

***Abstract*—This paper presents an unsupervised approach to extract maritime Patterns of Life (PoL) from historical Automatic Identification System (AIS) data based on a low-dimensional synthetic representation of ship routes. Recent advances in long-term vessel motion modeling through Ornstein-Uhlenbeck mean-reverting stochastic processes make it possible to encode knowledge about maritime traffic via a compact graph-based model where waypoints are graph vertices and the connections between them, i.e., the navigational legs, are graph edges. The resulting directed graph ultimately leads to the detection and statistical characterization of recurrent maritime traffic patterns. The proposed methodology has been tested on two extensive AIS datasets, collected for the North Sea and Ionian Sea operational trials of H2020-EU MARISA (Maritime Integrated Surveillance Awareness), to demonstrate the effectiveness and computational efficiency on real-world applications.**

***Index Terms*—Automatic Identification System, maritime situational awareness, Pattern-of-Life, Ornstein-Uhlenbeck process, change detection, DBSCAN clustering, maritime traffic graph**

I. Introduction

Extracting recurrent ship mobility patterns is crucial to improve Maritime Situational Awareness (MSA), which aims at understanding behaviors and activities that have an impact on the maritime environment. Ship traffic monitoring represents one of the biggest challenges in terms of law enforcement, search and rescue, environmental protection and resource management and, in recent years, has led to intensive research activities in order to exploit new methodologies in support of maritime surveillance. The concept of Pattern-of-Life (PoL) is commonly used in the context of activity-based intelligence which aims at understanding complex behaviors showing some regularity. In the maritime domain, this term is used to describe normal patterns of behavior of ships, i.e. recurrent maritime traffic patterns and summary statistics on the volume and type of vessels that, if well-characterized, can be beneficial for MSA applications such as anomaly detection. Nowadays, recently developed satellite and terrestrial networks of cooperative self-reporting ship location systems, such as the Automatic Identification System, provide ever-increasing volumes of maritime traffic data that can be used to enhance the general awareness of vessel pattern-of-life activities in both coastal and open waters.

This work was partially funded by the European Union's Horizon 2020 project MARISA - Maritime Integrated Surveillance Awareness.

As established by the International Maritime Organization's (IMO) Convention for the Safety of Life at Sea (SOLAS) [1], Automatic Identification System (AIS) must be on board all vessels with gross tonnage of 300 or more, and passenger ships of any size. This way AIS has become the major source, by coverage and volume of data, of maritime traffic monitoring, as each AIS transmitting vessel will report its identity (MMSI number), position, speed over ground (SOG), course over ground (COG), and other relevant information. On the one hand, this vast amount of information is becoming increasingly intractable to human operators, and calls for a high degree of automation in maritime route extraction and synthetic representation in order to convert data into usable knowledge for operational authorities and policy-makers. On the other hand, the availability of big maritime data opens up new possibilities for creating new types of analyses and extracting new information at one's disposal for MSA applications including anomaly detection [2]–[4], knowledge-based tracking and classification, prediction of long-term vessel motion and behaviors, threat assessment, etc.

Previous work on extracting maritime patterns focused on modeling the normal traffic activities in support of anomaly detection. In this context, [5] designs motion anomaly detectors based on the patterns extracted from AIS data in the framework of adaptive kernel density estimation. A machine learning framework performing the tasks of clustering, classification and outlier detection has been developed in [6] to represent vessel traffic and detect anomalous behaviors. Other proposed methods include associative learning procedures based on biological principles to determine abnormal behaviors and predict future vessels positions [7], and a two-level representation of maritime traffic [8] tested in the Baltic Sea region. Another well-known approach is for maritime traffic characterization is Traffic Route Extraction for Anomaly Detection (TREAD) [9]. TREAD generates a dictionary of historical vessel PoL represented by waypoint and route objects, containing dynamic and static AIS properties. A key property of TREAD is that waypoints and routes can be merged, split, removed, created as new data is collected. This comes at the expense of significantly increased computational costs due to a complex early-stage analysis of flag states, directions, velocities, destinations, and the required integration with databases.

Building on a compact synthetic representation of maritime

traffic routes proposed in [10], this paper presents and validates through large-scale real-world applications an unsupervised approach to automatically and efficiently extract maritime patterns of life from large volumes of historical AIS data. In particular, based on the fact that i) the majority of ship mobility is actually very regular since routes are usually computed based on fuel consumption, and ii) vessels in deep waters rarely perform maneuvers, the idea is to reduce ship trajectories into a sequence of waypoints (spatial regions where ships regularly stop or change their velocity) and navigational legs (whereon ships show a non-maneuvering behavior). This allows us to extract a compact, low-dimensional graph-based model of maritime traffic from raw AIS data where waypoints are represented by graph vertices and the connections between them, i.e., the navigational legs, are represented by directed graph edges.

Recent developments in long-term vessel motion modeling [11], have shown that the motion of ships along each trajectory can be effectively described by a piecewise Ornstein-Uhlenbeck (OU) mean-reverting stochastic process where model parameters switch in correspondence of waypoints. By adopting this strategy to encode recurrent maritime traffic patterns, unsupervised procedures have been developed to automatically extract knowledge via change detection tools aimed at finding relevant waypoints [12], and parameter estimation techniques to infer the dynamic behavior of ships in each navigational leg [13]. In order to discover significant waypoint areas, change points corresponding to the same geographical region are grouped into waypoint clusters, progressively simplified and merged using an incremental DBSCAN (i.e., Density-Based Spatial Clustering of Applications with Noise) procedure [14]. The resulting knowledge is shaped in a compact form via waypoints and navigational legs, generated and updated from the sequence of input AIS messages. The creation of such a maritime traffic graph ultimately leads to the detection and statistical characterization of maritime traffic patterns of life.

The effectiveness and computational efficiency of the proposed maritime patterns extraction tools have been investigated within the H2020-EU Project MARISA, where NATO-STO CMRE contributed to developing data fusion services for maritime surveillance including the ship routes extraction module.

II. Graph-based representation of maritime traffic

A. Dynamic model of long-term vessel motion

In order to be able to extract information about maritime traffic from the available AIS data, the key ingredient is a suitable mathematical model to describe the long-term motion of ships in open seas. The analysis of real-world AIS data shows that a significant portion of commercial maritime traffic, given that ships often try to optimize fuel consumption, is characterized by infrequent maneuvers.

Recently, it has been demonstrated in [11] that, by modeling the dynamics of non-maneuvering ships by means of the Ornstein-Uhlenbeck (OU) mean-reverting stochastic process, we can reduce by several orders of magnitude the uncertainty region of the long-term predicted position with respect to traditional state-of-the-art models (such as the nearly-constant velocity model). The main difference between the OU process and other conventional dynamic models is the presence of a feedback loop, which ensures that the velocity of the target does not diverge with time, but is instead bounded around a finite value, representing the desired (cruise) velocity of the ship.

Let the kinematic state of a ship be denoted by $x(t) = [p(t), \dot{p}(t)]$, where $p(t)$ and $\dot{p}(t)$ are the position and, respectively, velocity of a vessel in a two-dimensional Cartesian coordinate system. Then, the ship dynamics can be described by the following stochastic differential equation:

$$\dot{x}(t) = Ax(t) + Bu + D\dot{w}(t), \tag{1}$$

where $u = [u_x, u_y]^T$ is the long-run mean velocity and $w(t)$ is a standard 2-D Wiener process. The matrices A, B and D are defined as

$$A = \begin{bmatrix} 0_2 & I_2 \\ 0_2 & -\Lambda \end{bmatrix}, \quad B = \begin{bmatrix} 0_2 \\ \Lambda \end{bmatrix}, \quad D = \begin{bmatrix} 0_2 \\ \Omega \end{bmatrix}, \tag{2}$$

where 0_2 and I_2 are the 2-by-2 null and identity matrices, respectively, $\Lambda \in \mathbb{R}^{2\times 2}$ quantifies the mean-reversion effect, while Ω represents the process noise. If Λ has positive and distinct eigenvalues, then it can be written as $\Lambda = \bar{G}\Gamma\bar{G}^{-1}$, where $\Gamma = \mathrm{diag}(\gamma)$. The target state evolution is given by the first moment of the solution of (1), which takes the form

$$\begin{aligned} x_k = {} & G\tilde{\Phi}(t_k - t_{k-1}, \gamma)G^{-1}x(t_{k-1}) \\ & + G\tilde{\Psi}(t_k - t_{k-1}, \gamma)\bar{G}^{-1}u + w_k, \end{aligned} \tag{3}$$

where $G = I_2 \otimes \bar{G}$ and $\otimes$ denotes the Kronecker product. The full expressions of $\tilde{\Phi}(t, \gamma)$, $\tilde{\Psi}(t, \gamma)$ can be found in [11], [13].

Although (3) is best suited to capture the motion of non-maneuvering ships whose long-run mean velocity does not change over time, this model can be easily adapted to represent linear piecewise trajectories. These are useful to describe maritime traffic, which has been shown to be very regular, as most ships tend to navigate by following a sequence of waypoints. Hence, ships trajectories can be compactly represented by piecewise mean-reverting stochastic processes where each *segment* (or navigational leg) is characterized by a different long-run mean velocity, piecewise-constant over time, and by a set of *waypoints*, each corresponding to geographical regions where changes in velocity are likely to happen. Those regions will represent the nodes of the resulting maritime traffic graph.

The key challenge is that, in practical maritime traffic surveillance, waypoints and navigational legs are not known a priori. This motivates the development of unsupervised algorithms that, given the available historical AIS data, automatically extract a graph-based model of maritime traffic patterns by using change detection tools and clustering methods to find relevant waypoints, and parameter estimation techniques to compute the mean velocity in each navigational leg.

B. Detection of navigational change points

Based on the Ornstein-Uhlenbeck dynamic model for accurate long-term ship prediction, efficient statistical procedures [10], [12] can be developed to automatically identify specific geospatial waypoints where the parameters of the underlying OU process tend to change. In particular, if we assume that the long-run mean velocity of the process can abruptly change at any time instant, then the following sequential detection procedure based on Page's test [15] can be used to estimate the piecewise long-run velocity (see [12] for details):

1) First, the velocity of a ship is estimated using a sample mean estimator on n velocity samples;
2) Second, two Cumulative Sum (CUSUM) statistics are initialized in parallel to detect possible positive and negative changes of velocity;
3) Finally, if one of the two statistics exceeds a given threshold τ, then a change point is detected.

It is worth noting that the change detection capability depends on the value of τ, since low values will lead to a large number of detections, whereas high values will either prevent detection or worsen the performance in terms of average run length.

In other words, the above change detection procedure relies on the estimation of the long-run mean velocity parameter of the OU process, before and after a change. The CUSUM is computed in step 2) and, if a deviation from the desired threshold is observed, a change of the long-run mean velocity is declared. In this work, we considered four different types of change points: ports, navigational waypoints, entry and exit points.

C. Clustering of navigational waypoints

To find significant waypoint areas, standard clustering techniques can be used in order to group together multiple change points into a lower number of distinct waypoint clusters. Here, the DBSCAN (Density-Based Spatial Clustering Of Applications With Noise) algorithm [14] is adopted to associate detected change points to waypoint clusters based on a Mahalanobis distance that takes into account not only position, but also velocity. DBSCAN is a well-known data clustering algorithm, commonly used in data mining and machine learning. The main pros of using DBSCAN with respect to other state-of-the-art clustering methods, are its capability to identify an arbitrary number of clusters (no prior number of clusters needs to be predefined) of arbitrary shape, as well as its robustness to noisy data which makes it well-suited for our application. The DBSCAN algorithm basically requires to set only two parameters: i) the minimum distance ϵ, such that if the distance between two points is lower or equal to ϵ, the two points will be clustered together; ii) the minimum number η of points needed to form a dense cluster.

The clustering of navigational waypoints is based on the following Mahalanobis distance in a four-dimensional feature space as similarity measure between nodes i and j:

$$d_{ij} = (\phi_i - \phi_j)^T \Sigma^{-1} (\phi_i - \phi_j)$$

where Σ is a weighting matrix. The feature vector of change point i is defined as $\phi_i = [p_i, \theta_i]^T$ where p_i denotes the position detected through Page's test, while $\theta_i = [\theta_i^{in}, \theta_i^{out}]^T$ contains the velocity angle before and after the change. Intuitively, DBSCAN will look for regions where navigational change points are very close in the feature space and will identify outliers as points lying in low-density regions. In our approach, a large volume of historical AIS data is used to generate the traffic graph. However, as the amount of available AIS data grows to massive scales, computational techniques are needed to process, manage and store data efficiently. In this regard, we developed a fast and scalable implementation of DBSCAN that computes only distances between change points that are *nearest neighbors*, i.e. such that the distance metric between those data points is less than or equal to a predefined radius. This allows us to avoid computing the full large-scale matrix of all-to-all distances.

III. Extraction of maritime patterns

Starting from a collection of raw AIS messages, a ship trajectory can be represented by means of a set of waypoints and a piecewise constant profile for the long-run mean velocity. This allows us to extract a low-dimensional representation of maritime traffic that can be encoded through a directed weighted graph, whose nodes represent navigational waypoints, while edges represent navigational legs. Each edge is assigned a weight, proportional to the number of vessels which transitioned from the source to the target node of that specific link. By using this graph-based model, we can automatically extract maritime traffic patterns from AIS data.

A. Merging and pruning procedures

In real-world applications the procedure presented in Section II has to necessarily deal with all the underlying non-idealities. Although some of them are taken into account and compensated in the pre-processing stages described in [12], residual sources of noise and errors may rise after the clustering step. Moreover, when clustering algorithms such as DBSCAN are applied to large-scale real-world datasets involving a huge number of data points, the output generated by such unsupervised classification will usually need some post-processing.

In particular, pruning and merging techniques can be used to progressively improve and simplify the overall maritime traffic graph by reducing the number of graph entities (i.e., nodes and edges), and thus encode knowledge about ships' patterns using a lower-dimensional representation. The number of graph edges can be reduced by eliminating those links characterized by low weights that are least likely to represent recurrent maritime patterns, and by merging closely-spaced edges (connecting close source-target nodes) into one, as they are more efficiently represented by a single route. Moreover, edges falling over land are removed by a land avoidance logic based on bathymetry data. The number of waypoints clusters is also reduced by removing unreachable nodes that might originate from temporal and spatial gaps in the AIS

data. This can be the effect of a ship that either turned off the AIS transceiver or simply exited the coverage area. In addition, statistically close clusters (i.e., that fall within a given Mahalanobis distance) are also grouped together into a single waypoint node.

B. Graph structure and attributes

Using graph formalism, the structure of maritime traffic in the area of interest during a reference time interval can be represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N} = \{1, 2, ..., N\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. In particular, it is supposed that (i, j) belongs to $\mathcal{E}$ if and only if there are ship tracks with two consecutive waypoints i and j, where node i is the predecessor of node j. In this way, the adjacency matrix of the graph $\mathcal{A}$ can be directly constructed from the raw ship tracks data (i.e., time-ordered lists of AIS messages) by simply identifying the transitions (and associated direction) of ships from a generic pair of nodes. We also assume that $(i, i) \notin \mathcal{E}$ for any $i \in \mathcal{N}$, so that diagonal elements of $\mathcal{A}$ are set to zero, i.e. $\mathcal{G}$ is a simple directed graph with no self-loops.

In the context of H2020-EU MARISA, waypoints nodes and edges have been represented as vector features with different attributes directly extrapolated from the available AIS messages or computed during the graph extraction phase. The attributes of each node entity include: waypoint identifier, type, geographical location, traffic data category, statistics about the inward/outward speed of ships, and number of vessels that passed through the waypoint. Edges attributes sum up properties of source and target nodes. The resulting geospatial information layer, that can be used and shared to generate maps and display attributes of features, serves as a baseline reference of maritime traffic for different MSA applications.

IV. Real-world applications: North Sea and Ionian Sea

A. North Sea

The unsupervised approach presented in this paper has been applied to an extensive dataset of more than 5.5 millions of AIS messages broadcast by commercial cargo ships, and collected by a worldwide network of satellite and terrestrial receivers. The available AIS data was recorded from April to June 2018 in the area of the North Sea, spanning more than $8 \times 10^5\ km^2$, approximately from $0°$ to $10°$ longitude and from $45°$ to $60°$ latitude.

The parameters of the Ornstein-Uhlenbeck process are set as follows: $\Gamma = \gamma I_2$, $\Omega = \omega I_2$ with $\gamma = 3 \times 10^{-3}$ and $\omega = 14.1 \times 10^{-3}$. The clustering parameters are chosen as $\epsilon = 0.14$ and $\eta = 4$. After aggregating the available AIS data into 19009 tracks, the change detection routine described in Section II-B detects 162662 navigational change points. Those waypoints get clustered by DBSCAN into 2286 graph nodes connected through 5504 graph edges.

Finally, pruning and merging procedures are executed to improve the final output of the proposed maritime traffic extraction scheme. A qualitative representation of the resulting graph is shown in Fig. 1, where the main commercial routes and the paths connecting the major ports in the area have been correctly identified and extracted.

Fig. 1. Maritime traffic graph extracted from the North Sea AIS dataset, April-June 2018.

B. Ionian Sea

The proposed approach for unsupervised maritime patterns extraction has been also tested on an AIS dataset collected in the Ionian Sea area from June to August 2018. The extent of the collected AIS data is shown in Fig. 2, where a density map of the spatial distribution of AIS messages in the area of interest is reported. Each pixel in the figure covers a 6-by-6 nmi (one-tenth-degree) square on the ground and its color is proportional to the logarithm of the number of recorded AIS messages whose reported positions fall within its footprint.

The results of ship routes extraction on the Ionian Sea dataset are shown in Fig. 3. As we can see from the comparison against the density map in Fig. 2, the proposed unsupervised approach succeed in correctly capturing the main navigational waypoints and the high-density traffic routes from the available AIS data.

Conclusion and future work

In this paper we presented and validated through real-world applications an unsupervised approach to automatically extract a synthetic representation of maritime traffic routes from large volumes of historical AIS data. The key novelty of this method is the introduction of a compact graph-based model of maritime traffic, which in turn relies on a novel statistical modeling of ship motion. The ability to extract meaningful patterns from large amounts of noisy data is a

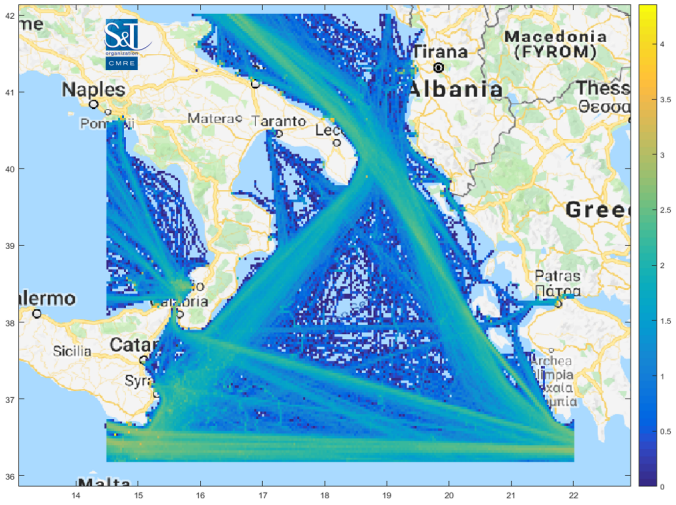

Fig. 2. Density map of AIS messages collected from June to August 2018 in the Ionian Sea region.

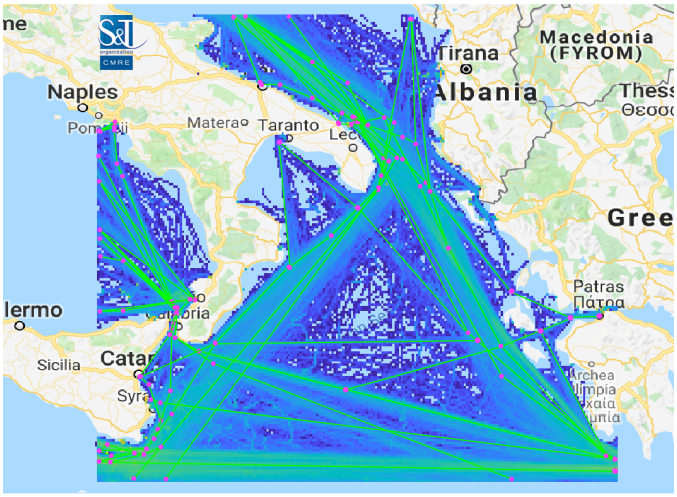

Fig. 3. Maritime traffic graph extracted from raw Ionian Sea AIS data, June-August 2018.

critical capability for operators, as the overwhelming amount of ship mobility data does not allow for an effective human inspection of the maritime traffic. With respect to other existent solutions, the graph-based representation of the traffic opens up to new opportunities for traffic analysis that could potentially have operational relevance, such as traffic network analysis.

Future work will focus on exploiting the knowledge extracted about maritime traffic to improve ship routes prediction, as well as on the introduction of novel criteria to assess the performance of maritime patterns extraction methods.

References

[1] IMO, *International Convention for the Safety of Life at Sea (SOLAS)*. 1974.

[2] N. Forti, L. M. Millefiori, and P. Braca, "Hybrid Bernoulli filtering for detection and tracking of anomalous path deviations," in *21st International Conference on Information Fusion*, pp. 1178–1184, 2018.

[3] N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Anomaly detection and tracking based on mean–reverting processes with unknown parameters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8449–8453, 2019.

[4] E. d'Afflisio, P. Braca, L. M. Millefiori, and P. Willett, "Detecting anomalous deviations from standard maritime routes using the Ornstein-Uhlenbeck process," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6474–6487, 2018.

[5] B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction," in *11th International Conference on Information Fusion*, pp. 1–7, 2008.

[6] G. K. D. de Vries and M. van Someren, "Machine learning for vessel trajectories using compression, alignments and domain knowledge," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13426 – 13439, 2012.

[7] N. A. Bomberger, B. J. Rhodes, M. Seibert, and A. M. Waxman, "Associative learning of vessel motion patterns for maritime situation awareness," in *2006 9th International Conference on Information Fusion*, pp. 1–8, 2006.

[8] V. Fernandez Arguedas, G. Pallotta, and M. Vespe, "Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 722–732, 2018.

[9] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 6, pp. 2218–2245, 2013.

[10] P. Coscia, P. Braca, L. M. Millefiori, F. A. N. Palmieri, and P. Willett, "Multiple Ornstein-Uhlenbeck processes for maritime traffic graph representation," *IEEE Transactions on Aerospace and Electronic Systems*, 2018.

[11] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, "Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 5, pp. 2313–2330, 2016.

[12] L. M. Millefiori, P. Braca, and G. Arcieri, "Scalable distributed change detection and its application to maritime traffic," in *IEEE International Conference on Big Data*, pp. 1650–1657, 2017.

[13] L. M. Millefiori, P. Braca, and P. Willett, "Consistent estimation of randomly sampled Ornstein-Uhlenbeck process long-run mean for long-term target state prediction," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1562–1566, 2016.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Intenational Conference on Data Mining and Knowledge Discovery*, vol. 4, pp. 226–231, 1996.

[15] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1-2, pp. 100–115, 1954.