Continuous detection of black holes for moving objects at sea

Loïc Salmon, Cyril Ray, Christophe Claramunt

Continuous detection of black holes at sea

Loïc Salmon, Cyril Ray, Claramunt Christophe

Continuous detection of Black Holes for moving objects at sea

Loic Salmon
Naval Academy Research Institute
29240 BREST Cedex 9 -France
loic.salmon@ecole-navale.fr

Cyril Ray
Naval Academy Research Institute
29240 BREST Cedex 9 -France
cyril.ray@ecole-navale.fr

Christophe Claramunt
Naval Academy Research Institute
29240 BREST Cedex 9 -France
christophe.claramunt@ecole-navale.fr

ABSTRACT

The main objectives of moving objects queries are to search for objects that either lie in some specific areas (i.e., range queries) or are close to one specific location (i.e., kNN queries). Such queries have been previously studied considering either offline database processes using some index techniques or online approaches where incoming data are processed to answer those queries "on the fly". The research presented in this paper considers hybrid queries applied to historical data as well as streaming data. When considering the specific context of the maritime domain and moving objects at sea, a key issue is to make a difference between covered and non covered areas (i.e., regions from where AIS positioning signals are either received or not received). This leads us to introduce the concept of "*Black Holes*" query where the objective is to identify regions respectively covered and non covered, this providing useful insights for maritime authorities in charge of the regulation of maritime transportation.

CCS Concepts

•Information systems → Online analytical processing; Geographic information systems; Data streams;

Keywords

Hybrid processing; Moving objects; Maritime monitoring; Black Holes

1. INTRODUCTION

Over the past few years, the fast emergence and proliferation of position-based sensors and devices produce very large volumes of data that need to be continuously analyzed. The maritime field is one amongst many domains impacted by this phenomenon with the rapid development of vessel positioning systems such as Automatic Identification Systems (AIS), Satellite AIS, Vessel Monitoring System (VMS) or Long Range Identification System (LRIT) that overall contribute to the real-time availability of large traffic data sets at sea. In particular, maritime transportation is a domain of increasingly intense traffic as the number of vessels navigating worldwide is in constant augmentation. Practically, ships are fitted out with almost real-time position report systems whose objective is to remotely identify and locate vessels (e.g., AIS real-time position reports). However, the current geographical coverage of these positioning systems is not complete. In fact, there still exists many areas from where position signals emitted from ships cannot be received as no antennas cover these regions, this being typically the case for AIS. In the remaining part of this paper we consider these specific areas as *Black Holes* and as regions were typically AIS data emitted by ships cannot be received. Figure 1 illustrates this notion of *Black Holes*, this map has been derived by extraction and visualization of AIS positions in the Aegan Sea. A regular grid has been applied to this map where at the bottom cell B can be identified as *Black Hole* because some vessels are going from cell A to cell C whereas no signal has been received from cell B.

The concept of Black Hole and the related notions of coverage or non coverage of positions are closely related and of crucial importance at sea. Indeed, the monitoring and analysis of mobilities at sea is particularly important for safety and security reasons. Mobility and behavior analysis could be used to detect illegal or criminal activities, risks at sea (flow of illicit products, illegal immigration, overfishing, pollution by hazardous materials, piracy, accidents , etc.), and more generally any violation to maritime regulations [11].

However, the issue is far from being straightforward. First *Black Holes* are likely to have fuzzy and fluctuating boundaries that change over time as weather and atmospheric conditions are changing and as those have non negligible impacts on AIS signal transmission and reception. Moreover, AIS behavior with irregular positions reports and limited coverage reinforce the difficulties to search for those regions. In fact such concept of *Black Hole* should be studied at different times in order to identify which areas are covered and uncovered. At a given time, one should ideally identify these uncovered regions in order to identify possible anomalies and suspicious behaviors, that is, difference between vessels that have their AIS emitter switched off to those that cross those specific regions with not enough signal emission, or even to

detect fake messages [12].

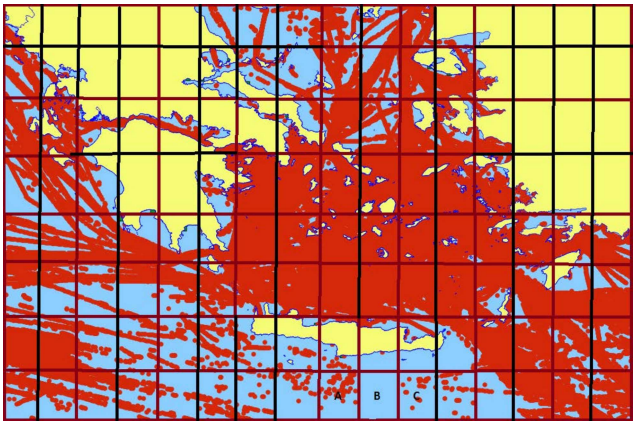

Figure 1: Example of a *Black Hole* in the Aegan Sea

In order to identify those *Black Holes*, historical and streaming data are likely to provide useful inputs to apprehend appropriately the fluctuations of the *Black Holes* boundaries. In fact, the way a given region might be considered as a *Black Holes* not only depends on the weather conditions, but also on the analysis of historical data to provide a sufficient coverage of the data in space and time. For instance, regions from where some data has been received before can become a *Black Hole* if weather conditions are not favorable. On the other hand, if only streaming data is considered, empty regions might not denote non covered regions when maritime traffic is absent in that region. Accordingly, lack of positioning data needs to be interpreted as either a non coverage of a region or a region where no vessel has been through. This motivates our search for an integrated approach whose objective is to derive a global picture of a given maritime region and then of those *Black Holes* at different time periods and thus potentially detecting some maritime traffic anomalies.

The remainder of this paper is organized as follows. Section 2 provides an overview of the problem to address and introduces a series of concepts used by our modelling approach. Section 3 describes the main principles of a hybrid approach for moving object handling and more specifically *Black Hole* detection. Section 4 gives the general method for the *Black Hole* identification and the processes involved at each step (i.e. offline, online and hybrid). Section 5 presents the experimentation and implementation results. Section 6 discusses the interest of the *Black Holes* approach and current limitations. Finally, Section 7 summarizes and concludes the paper.

2. PROBLEM STATEMENT

This section introduces the concept of *Black Holes* and a series of definitions that support this notion. The aim of the method explored in this paper is to characterize some spatial areas as covered or uncovered with a precision depending on the amount of data received for the considered area and regardless of performance issues. The problem of coverage considered illustrates the need for an approach that requires to consider both knowledge extracted from data stored in database and streaming data to take into account the evolving and dynamic context (i.e. actual coverage depends on signal propagation and thus weather conditions). This coverage problem has not been investigated yet in the maritime domain to the best of our knowledge. Related works have explored density analysis techniques on positions (as kernel density estimation [13]) and this not allowing the implicit detection of *Black Holes* in real-time.

Let us consider a large series of positions coming from AIS data representing a set of n moving objects on a maritime region S $\subset \mathbb{R}^2$. Those moving objects can be represented as series of tuples of the form *(x,y,t)* where (x,y) denotes a location p $\in S$ (cf. Definition 1) and $t \in T$ (cf. Definition 2) is the timestamp of the position recorded.

Definition 1 : Spatial Domain. *Let P denote the spatial domain that contains all possible pairs of values (x,y) with real cartesian coordinates (x,y)* $\in \mathbb{R}^2$.

Let us consider S a subregion of $\mathbb{R}^2$ such as S $\subset$ P. A segment *sg* is defined by a pair (a,b) with (a,b) $\in$ S.

Definition 2 : Time Domain. *Let T denote the Time domain as an ordered infinite set of time instants* $t_i \in T$.

A time interval $[t_1, t_2] \subset$ T consists of all distinct time instants of T between t_1 and t_2.

Definition 3 : Trajectory. *A trajectory Traj is defined as an ordered list of timestamped positions where positions are defined as points from the Spatial Domain P and timestamped as time instants of the Time Domain T. A trajectory is denoted as Traj :* $Id \in \mathbb{N} \rightarrow List(p_i \in S, t_i \in T)$ *where* p_i *is the position of the moving object at the timestamp* t_i *with* $i \in \mathbb{N}$ *and Id gives the identifier of the trajectory. Each trajectory materializes a polyline given by the ordered list of segments derived from its successive positions. More formally a trajectory Traj :* $Id \in \mathbb{N} \rightarrow List(p_i \in S, t_i \in T)$ *can be mapped as Poly :* $Id \in \mathbb{N} \rightarrow List(s_i \in S^2, \tau_i \subset T)$ *where each* s_i *is a pair of two successive positions and* τ_i *is the corresponding time interval between those two successive positions.*

In order to model the concept of *Black Hole*, let us introduce a uniform grid G, splitted into regular disjoint cells $C_{i,j}$ of fixed size $size_{cell}$ to determine the number of positions by cells, and finally identify those that are not covered.

Definition 4 : Grid. *Let G denote a regular grid with G=Set(*$C_{i,j}$*) and each disjoint cell* $C_{i,j}$ *is of fixed size* $size_{cell}$*, defined by* $f : p \in S \rightarrow C_{i,j}$ *such as* $i = \lfloor x/size_{cell} \rfloor$ *and* $j = \lfloor y/size_{cell} \rfloor$*, with* $C_{i,j}$ *the cell related to the* i^{th} *abscissa and the* j^{th} *ordinate for* $i \in 1..n$, $j \in 1..n$.

The size of the cells is evaluated based on the following norm [1] which describes the duration between two records of a vessel considering its actual speed and heading. The necessary size of the tiles is chosen to prevent vessels from emitting two successive positions in two non adjacent cells (i.e., outside non covered regions). In order to do so, the worst case has been considered where a vessel is close to the boundary of a cell and moves with a right heading and the highest speed noted V_{max} and $T_{t->t+1}$ is the time interval between two position records. In fact, there is a need to consider a fixed size greater than $size_{cell}$. To minimize

the "answer loss problem" described in [7] that occurs for a regular grid, smallest cells as possible are chosen. Indeed, selecting thinnest cells increases computation costs, but while larger cells decreases computation cost identification of some specific events can be compromised. Therefore, G is defined with a chosen size cell denoted $size_{cell}$, and with $size_{cell} = V_{max} * T_{t->t+1}$.
Let us introduce a few additional notations related to the grid previously defined. For each cell $C_{i,j}$ related to the i^{th} abscissa and the j^{th} ordinate for i ∈ 1..n, j ∈ 1..n, then the following measures are introduced :

- $Npos_{i,j}(\tau)$ gives the number of distinct positions that have been recorded in the cell $C_{i,j}$ during the period time $\tau \subset T$, more formally $Npos_{i,j}(\tau) = \sum_{k=1}^{nbpositions} l_k /$

$$l_k = \begin{cases} 1, & \text{if } p_k \in C_{i,j} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

  where *nbpositions* is the number of positions recorded over τ.

- $Nvessel_{i,j}$ τ gives the number of distinct vessels that have been recorded in the cell $C_{i,j}$ during the period time $\tau \subset T$, more formally $Nvessel_{i,j}$ $(\tau) = \sum_{Id=1}^{nbvessels} t_{Id}$ during τ /

$$t_{Id} = \begin{cases} 1, & \text{if } \exists (p_k, t_k) \in Traj(Id)/p_k \in C_{i,j} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

  where *nbvessels* is the number of distinct vessels recorded over τ.

- $Ncross_{i,j}(\tau)$ gives the number of distinct vessels that have crossed the cell $C_{i,j}$ during the period time $\tau \subset T$, that means the number of vessel trajectories (modeled as a polyline) that intersect the corresponding cell $C_{i,j}$, more formally $Ncross_{i,j}(\tau)$= $\sum_{Id=1}^{nbvessels} t_{Id}$ during τ /

$$t_{Id} = \begin{cases} 1, & \text{if } Poly(Id) \cap C_{i,j} \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

  where *nbvessels* is the number of distinct vessels recorded over τ.

  It is immediate to note that, $Ncross_{i,j}(\tau) > Nvessel_{i,j}(\tau)$.

Definition 5 : Empty cells set. *An empty cells set over an interval of time τ is given by the cells that haven't been crossed by some trajectories. The empty cells set noted $E(\tau)$ is defined by $E(\tau)$=$\{C_{i,j} \in G \mid Ncross_{i,j}(\tau) = 0\}$.*

Definition 6 : Crossed cells set. *A crossed cell set is given by the cells that have been crossed by some vessels over a time interval τ. More formally, the Crossed cells set denoted $\bar{E}(\tau)$ is given by $\bar{E}(\tau)$=$\{C_{i,j} \in G \mid Ncross_{i,j}(\tau) > 0\}$. We have G=$E(\tau) + \bar{E}(\tau)$*

Definition 7 : Black Hole cells set. *A Black Hole cells set denoted $BH(\tau)$ is given by the cells that are not covered for the τ time period considered. It is composed of two subsets: the formally identified Black Hole cells and the supposed Black Holes cells.*

These two subsets are formally defined as follows:

Definition 8 : Formally identified Black Hole cells. *Given a time interval τ, the Black Hole cells can be given as all cells $C_{i,j}/Nvessel_{i,j}(\tau) = 0$ and $Ncross_{i,j}(\tau) > 0$ (i.e., the set of cells that don't have any vessel position recorded but more than a given threshold θ of trajectories that cross that cell for a given interval of time τ.)*

Definition 9 : Supposed Black Hole cells. *Given a time interval τ, the supposed Black Hole cells are given by the cells that belong to the empty cells set and are not covered during τ. (i.e all $C_{i,j} \in BH(\tau) \cap E(\tau)$)*

Definition 10 : Covered cells set. *A covered Cells set is given by the cells that are covered for the τ time period considered. The Covered cells set noted $\bar{BH}(\tau)$ is the complement of $BH(\tau)$ such as G=$BH(\tau) + \bar{BH}(\tau)$. Similarly, to $BH(\tau)$, $\bar{BH}$ (τ) is composed of two subsets called Formally identified covered cells and Supposed covered cells.*

More formally these two subsets are defined as follows:
Definition 11 : Formally identified covered cells. *Given a time interval τ and a threshold θ, the Covered cells can be defined as all cells $C_{i,j} such as Nvessel_{i,j}(\tau) > 0$ (i.e., the set of cells that have recorded more than one vessel position for a given interval of time τ.)*

Definition 12 : Supposed covered cells. *Given a time interval τ, the supposed covered cells are given by the cells that belong to the empty cells set because no vessels have been through this area over the last τ period. (i.e all $C_{i,j}$ ∈ $\bar{BH}(\tau) \cap E(\tau)$)*

Considering the actual time period, the following table covers all cases that can happen according to the previous definitions :

	Crossed cells	Empty cells
Covered cells	Formally identified covered cells	Supposed covered cells
Black Hole cells	Formally identified Black Hole cells	Supposed Black Hole cells

Table 1: Cell taxonomy

In order to populate the concept of *Black Hole*, a distinction is made between cells where no positions are recorded, because no moving objects can record their location, and cells that have not been crossed during a given time period τ. In both cases, no positions should have been recorded into those regions (i.e. $C_{i,j} \in E(\tau)$). Indeed, uncovered cells might denote in some cases cells that have not been crossed by some moving object during a time period as well as some covered ones. However, the objective is to derive *Black Hole*s that denote regions where no positions can be recorded because such regions are really uncovered for the time period considered. The challenge here is that some behaviors or events should be inferred but with a lack of information (i.e., mainly for empty cells). This entails the need for historical data to complete the information provided by streaming data, but this being not always sufficient enough.
Overall, identifying less populated cells is of high interest during an offline process. Also, while considering historical data, the amount of data considered is relatively

high, with some regions that might have been covered or uncovered during a short period of time. In some cases, regions that are usually uncovered can record some positions. From the offline process, let us extract what we call the Candidate *Black Holes* set noted Set_{off}(cdtBH) corresponding to the less populated cells. As defined in [7] we maintain a data structure called Density Histogram DH(τ) that counts for each cell $C_{i,j}$ the number of positions $Npos_{i,j}(\tau)$ recorded during the time period $\tau \subset T$. Finally, we have : DH(τ)=[$C_{i,j}$,$Npos_{i,j}(\tau)$] sorted by ascendant order of $Npos_{i,j}(\tau)$.

Considering the number of positions $Npos_{i,j}(\tau)$, Set($C(N < k,\tau)$) (respectively Set($C(N < k,\tau)$)) denote the set of cells that have recorded less (respectively more) than k positions during a given time interval $\tau \subset T$, more formally :

- Set($C(N < k,\tau)$) = $\{C_{i,j} \in G \ / \ Npos_{i,j}(\tau) < k\}$.
- Set($C(N > k,\tau)$) = $\{C_{i,j} \in G \ / \ Npos_{i,j}(\tau) > k\}$.
- Set($C(N = k,\tau)$)= $\{C_{i,j} \in G \ / \ Npos_{i,j}(\tau) = k\}$.

The offline candidate Black Hole set or Set_{off}(cdtBH) for brevity, is obtained from an offline process. Given a percentage p and time period τ, Set_{off}(p,τ)= Set(C (N<q, τ)) where q is the upper number of recorded positions in the p percent less populated cells. Those less dense regions are more likely to be *Black Hole* regions and should be examined while geostreaming positions are processed "on the fly" by the online part. For each cell $C_{i,j}$, let us define an associated metric called Density Value and denoted $d(i,j)(\tau)$. The Density function denoted d is defined as follows d: $Npos_{i,j}(\tau) \in \mathbb{N} \rightarrow$ v $\in$[0..1] where v is the real value associated to the q^{th} quantile to which $C_{i,j}$ belongs regarding the ordered $Npos_{i,j}(\tau)$ values. Thus, if $C_{i,j} \in Set_{off}$(p,τ) then v = 0.

Let us introduce the final concept of online Black Hole set involved during the online process. The online candidate Black Hole set or Set_{on}(cdtBH) for brevity, is obtained during the online process. Set_{on}(cdtBH)= $\{C_{i,j} \in G \mid Nvessel_{i,j}(\tau)=0 \}$ where τ corresponds to the more recent time interval. Similarly to the offline part, a data structure is maintained on streaming data and called Cover Histogram. This Cover Histogram denoted Cover(G,τ), contains the $Nvessel_{i,j}(\tau)$ and $Ncross_{i,j}(\tau)$ values associated to each cell $C_{i,j}$ for the time interval τ considered. More formally we have Cover($C_{i,j}$,τ)=($Nvessel_{i,j}(\tau)$,$Ncross_{i,j}(\tau)$).

Concerning time notations, $t_{current}$ is defined as the timesptamp corresponding to the current time whereas t_{old} corresponds to the timestamp of the older record stored in database. τ_{Past} refers to the time interval concerning historical data from the first timestamp of the database to the last timesptamp of historical data (the time interval concerning the whole data except of the sliding window of range ω and slide β), more formally τ_{Past} = [t_{old}, $t_{current-\omega}$]. $\tau_{current}$ refers to time interval considered for the whole sliding window that means $\tau_{current}$ = [$t_{current-\omega}$, $t_{current}$]. Finally, we have τ_{β} that refers to time interval between $t_{current}$ and $t_{current-\beta}$, more formally τ_{β} = [$t_{current-\beta}$, $t_{current}$] .

3. PROCESSING PRINCIPLES

This section introduces the concept of a hybrid approach for moving object processing in real-time (i.e. combining knowledge from data stored in database with streaming data) and more specifically the guideline process associated to *Black Holes* identification.

3.1 General approach

The concept of *Black Hole* motivates the development of a hybrid processing in order to consider moving objects queries. Those regions have fuzzy and fluctuating boundaries that require to consider both streaming and historical data to identify precisely their locations for every considered time period τ. To find those regions without any coverage, let us apply the approach first described in [14] which provides a hybrid processing to deal with moving objects as shown in (cf. Figure 2) and derived from the lambda architecture [9].

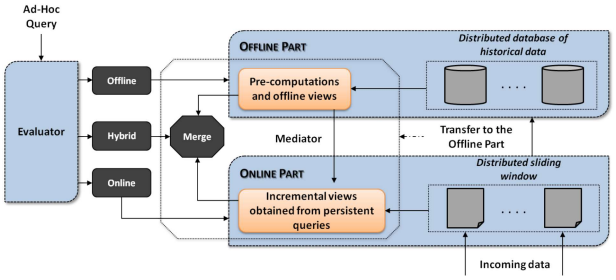

Figure 2: Architectural principles

Online processing is performed on a distributed sliding window whose size can be changed according to the amount of data collected in real-time. Online views on continuous queries are updated and incremented while incoming data stream is processed. When a user would like to execute a query for which no synthesized summary exists, necessary data to handle the query are accessible via the sliding window. When the temporal interval of the sliding window is exceeded, data is transferred to the historical database to perform distributed processing offline. In order to have a reactive system, summaries are performed on historical data and updated upon arrival in the database, and then transferred to online part to provide real-time answers.

Furthermore, two entities (i.e. the Evaluator and the Mediator) have the role (cf. Figure 2) of identifying the data to extract and process, and to manage interactions between the historical database and real-time processing system. In other words, these are the major components of this hybrid system which allows to merge the online and offline parts, and to answer the query using the minimum data as possible in accordance to user's requirements and are described in more details in [14].

The role of the *Mediator* is to manage the data flows between components online and offline, preserve and store the associated views and merge them to answer hybrid queries. The *Evaluator* analyzes the input query and tries to infer the type of request, (i.e., online, offline or hybrid) to guide, based on the identified type of query, recovery of data and information needed in the architecture. It transmits the desired data to the *Mediator* to deal with, then the *Mediator* is responsible to answer the query taking, combining or performing processing on the sliding time window or on the archive following the query type.

3.2 Guideline process for identifying Black Holes

The detection *Black Holes* process developed in this paper follows this hybrid framework. Regarding the offline part, less populated cells are identified using a similar process as the one described in [4] where the authors aim to find dense regions for moving objects. We maintain a *Density Histogram* noted DH(τ_{Past}) that counts the number of different positions recorded by cells for historical data corresponding to the time interval τ_{Past}, this acts as a filtering process to identify the offline *Candidate Black Holes* set (or Set_{off}(cdtBH)). With this offline *Candidate Black Holes* set, one can extract the current *Black Holes* set (or BH($\tau_{Current}$)) giving some additional information to the online process. During the online process, online synopses are built for each time period $\tau_{Current}$ and combined with the offline synopsis in a way to give more weight to recent data that have arrived into the system and finally extract the current *Black Hole* cells (cf. figure 3).

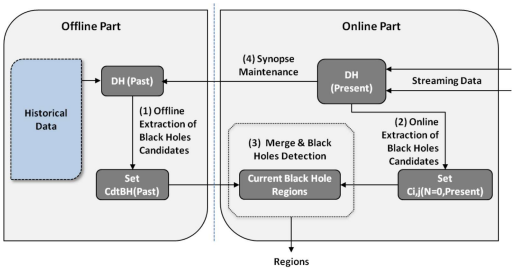

Figure 3: Guideline Process for Black Holes detection

Several steps are involved in this *Black Holes* identification process as follows:

- *First Step : Offline Candidate Black Holes Extraction.* The process to identify the less populated cells is applied to historical data. A quite similar method to the one used in [4] for dense cells finding, can be used to search for less populated cells. However, problems such as answer loss or ambiguity can happen when considering cells of fixed size to define and observe densities [10]. Those issues are not addressed in this paper that mainly focuses on the hybrid processing rather than on the design of a new state of the art technique to deal with historical data.

- *Second Step : Online Candidate Black Holes Extraction.* During this step, the objective is to categorize the different cells of the grid and identify the formally identified Black Holes cells, as well as the formally identified covered cells and the empty cells (cf. Table 2). This then gives a picture of the candidate *Black Hole* set concerning the online part similarly to the offline part. The issues concern the granularity aspects involved by the sliding windows variables (slide and range considered) as well as the data structure related to the online density (or coverage) [8].

- *Third Step : Merge and Black Holes Detection.* The merge and identification of the real *Black Holes* set requires a specific technique to obtain them from both the candidate *Black Holes* set extracted from the offline part with positional records received "on the fly" [3]. One of the main issues is to infer some knowledge from empty cells by combining both streaming position and knowledge from historical data.

- *Fourth Step : Synopse Maintenance.* The continuous update of the candidate *Black Holes* set, derived from the offline part and maintenance of synopses over the different time periods, and that need to be updated continuously as old streaming data are transferred to the offline part [14].

In order to identify those *Black Holes*, a uniform grid G is applied (cf. Section 2). Every cells of the grid are evaluated considering the different measures previously introduced; that is, the number of positions recorded inside the cell for the offline part and both the number of distinct vessels that have recorded their positions in the cell, and the number of vessels that have been through a cell this during the online process to determine current *Black Holes* cells. We have chosen fixed size cells in spite of limitations given in [10], as it is more convenient and appropriate to compare synopses extracted from the offline part with actual data received in real-time using regular cells rather than an index. Indeed, using indexes should have been required to apply two indexes at least each related to both online and offline parts, with probably a deep index for the offline part and an unbalanced one for the online one that are thus not easily comparable. For some cells, an exact differentiation between uncovered and covered cells does not work, for several reasons : Firstly, uncertainty concerning AIS data and the fact that some users in their boats can switch off their emitter; secondly, because there is a need to distinguish covered cells from *Black Hole* cells among *Empty cells*. Finally, without additional data as meteorological data, it is still difficult to determine exactly changes in the coverage regions. In some cases, some cells might not be identified as covered or not, this reflecting some uncertainty in the delimitation of those regions we intend to find.

4. BLACK HOLE DETECTION

The concept of *Black Hole* requires for its implementation extraction of historical data and confrontation to streaming data. The detection process itself can be organized in three steps (cf. Figure 3) described respectively in section 4.1, section 4.2 and section 4.3.

4.1 Offline process for Black Hole detection

Extraction of historical data has been already addressed in related work for identifying dense regions from moving objects, but the problem here is the reverse one [7]. First, a summary of cells which are candidate black holes denoted *Black Holes* (or Set_{off}(cdtBH)) should be extracted from historical data. In order to extract Set_{off}(cdtBH) an uniform grid G is considered (cf Section 2) and for each cell the number of positions $Npos_{i,j}(\tau_{Past})$ recorded into the corresponding cell $C_{i,j}$ is computed. Let us consider that a cell $C_{i,j}$ belongs to the candidate *Black Holes* set if and only if the count of distinct positions recorded in $C_{i,j} < \theta$ where θ is threshold value. θ is computed as the upper limit of the lowest percentile q amongst all $Npos_{i,j})(\tau_{Past})$ with q a percentage given by the user (cf. algorithm 1). A Density Value $d_{i,j}(\tau_{Past})$ is generated for each cell, this value is used during the merge part as well as the offline candidate *Black Hole* set to identify current *Black Holes*.

Algorithm 1: Calculate the offline candidate Black Hole set

Data: [p] Positions, Grid G [$C_{i,j}$], Int θ, Float q
Result: $Setoff$(cdtBH)
initialization;
foreach *grid cell* $C_{i,j}$ *in* G **do**
  $Npos_{i,j}(\tau_{Past}) \Leftarrow \sum_{k=1}^{nbpositions}(v_k)$
  where v_k=1 if $p_k \in C_{i,j}$ otherwise 0
  DH((τ_{Past})) $\Leftarrow$ DH((τ_{Past})) :: ($C_{i,j}$, $Npos_{i,j}(\tau_{Past})$)
end
$\theta \Leftarrow Npos_{i,j}$ such that $Npos_{i,j} = DH\tau_{Past}[q * Length(G)]$
`// θ is the upper Npos value of the first quantile q;`

foreach *grid cell* $C_{i,j}$ **do**
  compute $d_{i,j}(\tau_{Past})$
  if $Npos_{i,j}(\tau_{Past}) < \theta$ **then**
    Set_{off} (cdtBH) $\Leftarrow Set_{off}(cdtBH) \cup C_{i,j}$
end

Let us first consider the extracted set of offline candidate black holes Set_{off}(cdtBH) and secondly the density value associated to each cell $C_{i,j}$ of the grid. Such candidate black holes should be compared to positions coming from the streaming data over a given time interval $\tau_{current}$.

4.2 Online process for identifying Black Holes

Let us value the Cover Histogram introduced in section 2 and denoted Cover(G,$\tau_{current}$) which is continuously updated by the incoming streaming data. For each grid cell $C_{i,j}$, $Nvessel_{i,j}(\tau_{current})$ and $Ncross_{i,j}(\tau_{current})$ measures associated to each cell are computed. Thus, let us denote Cover($C_{i,j}$,$\tau_{current}$)=($Nvessel_{i,j}$,$Ncross_{i,j}$). A sliding window with a range ω and a slide β is used during the process. For every (ω/β) time period, a Cover Histogram is determined and thus the candidate *Black Hole* online extraction is computed at every β time period considering a whole time period ω by aggregating the different Cover Histograms relative to each pane of period (ω/β) taking care of the fact that each vessel is counted once only. To determine the online candidate black hole set Set_{on}(cdtBH) the set($C(Nvessel_{i,j}$ ($\tau_{current}$)$< \theta$) is computed. Thanks to the value of the temporal range, the cells that recorded a few positions for such short time period are considered as not covered. This leads us to consider candidate *Black Hole* cells as all cells that have been less than θ recorded vessel positions where θ is a threshold fixed similarly to the offline part but considering $Nvessel_{i,j}$ ($\tau_{current}$) instead.

Algorithm 2 computes the set of candidate *Black Holes* related to the online part. In order to obtain the final figures of the *Black Holes* during the last ω time period, one have to consider additional information obtained from the online process and also historical data in order to make a difference between covered cells and *Black Hole* cells among *Empty* cells.

Algorithm 2: Calculate the online candidate Black Hole set

Data: [v] Vessels, Cover($\tau_{current}$), Int θ, Float q
Result: Set_{on}(cdtBH)
initialization;
foreach *each pane* **do**
  foreach *new vessel position* **do**
    if $v \in C_{i,j}$ *for the first time* **then**
      $Nvessel_{i,j}(\tau_\beta) \Leftarrow Nvessel_{i,j}(\tau_\beta) + 1$
      $Ncross_{i,j}$ $(\tau_\beta) \Leftarrow Ncross_{i,j}(\tau_\beta) + 1$
  end
end
Cover($\tau_{current}$) $\Leftarrow$ $\sum_{i=0}^{\omega/\beta} Cover[t_{current-i*\beta}, t_{current-(i+1)*\beta}]$
`// we sum every cover histogram considering vessel and cross numbers;`
Sort Cover(G,$\tau_{current}$) by $Nvessel_{i,j}$ $\tau_{current}$
`// we sort the Cover Histogram by vessel number;`
$\theta \Leftarrow Nvessel_{i,j}/Nvessel_{i,j} = Cover(\tau_{current})[q * Length(G)]$ `// θ is the upper Nvessel value of the first quantile q;`

foreach β *time period* **do**
  foreach *grid cell* $C_{i,j}$ **do**
    if $Nvessel_{i,j} < \theta$ **then**
      Set_{on}(cdtBH) $\Leftarrow Set_{on}$(cdtBH) $\cup C_{i,j}$
  end
end

4.3 Black Hole extraction from both offline and online data

So far and at every β time period the cells $C_{i,j}$ candidate black holes are extracted. In fact, the size of the window range ω might generate in a few cases false candidate cells for the *Black Hole* set (i.e., cells that value the Supposed covered cells set($\tau_{current}$)). Due to the size of the cells, no positions are recorded on these cells given a period of time. In order to tackle this problem, let us compare the number of trajectories that have intersected a given cell $C_{i,j}$ by considering the trajectories involved as polylines. Let us first compare the number of vessels $Nvessel_{i,j}$ that have recorded their information in $C_{i,j}$ to the number $Ncross_{i,j}$ that have intersected the cell. Let us then categorize such cells using the following rules described further in our algorithm(cf. Algorithm 3, Figure 4) :

- *Covered Cell (i.e. $C_{i,j} \in \bar{BH}(\tau_{current})$).* When the number of recorded vessels in the cell is close to the number of trajectories that have crossed this cell with $Ncross_{i,j} > 0$ then let us consider that such cell is covered (i.e. when $Nvessel_{i,j}(\tau_{current})$ / $Ncross_{i,j}$ $(\tau_{current}) \simeq 1$ and $Ncross_{i,j} > 0$). Accordingly, the difference between the two numbers is the consequence of some vessels switching off their AIS emitter. For the following parts and experiments, let us consider that cells $C_{i,j}$ lie in this category if $Nvessel_{i,j}(\tau_{current})$ / $Ncross_{i,j}$ $(\tau_{current}) >= \rho$.

- *Black Hole Cell (i.e. $C_{i,j} \in BH(\tau_{current})$).* When the number of recorded vessels in the cell is much lower than the number of trajectories that have crossed this cell with $Nvessel_{i,j} > 0$ then let us consider that the cell is not covered (i.e. when $Nvessel_{i,j}$ $(\tau_{current})$ / $Ncross_{i,j}$ $(\tau_{current}) \simeq 0$ with $Ncross_{i,j}$ $(\tau_{current}) > 0$). In this case, the difference between the two numbers can be explained by the non coverage of the cells. Indeed, when vessels are into the cell, positions emitted by the associated boats are not received by coastal antennas. For the following part and experiments, let us consider that a cell lies in this category if $Nvessel_{i,j}$ $(\tau_{current})$ / $Ncross_{i,j}$ $(\tau_{current}) < \rho$.
- *Empty cell (i.e. $C_{i,j} \in E(\tau_{current})$).* When the number of recorded vessels in the cell and the number of trajectories that have crossed the cell $C_{i,j}$ are equal to 0 offline data is required to determine if the region is covered or not (i.e. when $Nvessel_{i,j}$ $(\tau_{current}) = Ncross_{i,j}$ $(\tau_{current}) = 0$).
- The last case $Nvessel_{i,j} > Ncross_{i,j}$ cannot happen by definition as $Ncross_{i,j}$ is defined, indeed if a vessel lies in a cell $C_{i,j}$ then $Ncross_{i,j}$ is incremented, so $\forall$ $C_{i,j}$, $\tau \in T$, $Ncross_{i,j}(\tau) >= Nvessel_{i,j}(\tau)$

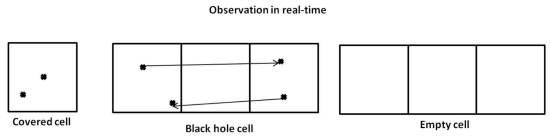

Figure 4: Online part results

Regarding the specific cells that have neither be crossed, nor have positions inside considering the online part (i.e. empty cells), they still need to be identified as covered or not (cf. Figure 5).

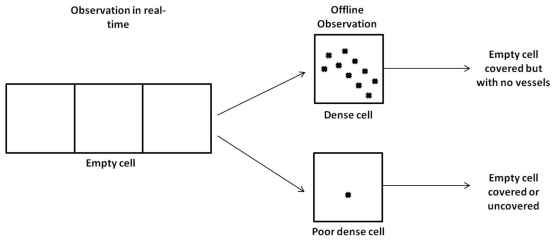

Figure 5: Regions prune process

In order to address this case, let us consider the offline part as follows. Let us consider the Density Value $d_{i,j}(\tau_{Past})$ associated to the cells $C_{i,j}$ to identify *Black Hole* the cells included in the *Empty cells* set denoted $E(\tau_{current})$. As mentioned above, and due to some granularity aspects, we might have some specific time periods where no vessels cross a given region even if this region is usually covered (i.e., have a sufficient Density Value). Given a threshold D :

- If $C_{i,j} \notin Setoff(cdtBH)$ and $d_{i,j}(\tau) > D$, one can reasonably guess that $C_{i,j} \in E(\tau_{current}) \cap \bar{BH}(\tau_{current})$ (i.e. $C_{i,j}$ is in the Supposed covered cells set).
- Otherwise, the cell $C_{i,j}$ is empty and can be either covered or uncovered and need further investigations to determine its coverage.

If $d_{i,j}(\tau_{Past})$ is under this threshold value, we consider that we cannot infer any information concerning that specific cell. Further works should be made to take care of antennas positions or the nature of adjacent cells to determine if those uncertain cells tend to be covered or not.

Algorithm 3: Black Hole extraction and merge with the offline part

```
   Data: [v] Vessels, Cover(t_current−β, t_current),
         Set_on(cdtBH), Float ρ
   Result: Set_on(cdtBH)
 1 initialization;
 2 foreach β time period do
 3     foreach grid cell C_{i,j} ∈ Set_on(cdtBH) do
 4         if Nvessel_{i,j}(τ_current) =
           Ncross_{i,j}(τ_current) = 0 then
 5             E(τ_current) ⇐ E(τ_current) ∪ C_{i,j}
 6         else
 7             if Nvessel_{i,j} (τ_current) / Ncross_{i,j}
               (τ_current) < ρ then
 8                 BH(τ_current) ⇐ BH(τ_current) ∪ C_{i,j}
 9
10             else
11                 Nvessel_{i,j} (τ_current) / Ncross_{i,j}
                   (τ_current) >= ρ
12                 C_{i,j} ∈ B̄H (τ_current)
13
14     end
15     foreach grid cell C_{i,j} ∈ E(τ_current) do
16         if d_{i,j} (τ_Past) > D then
17             C_{i,j} ∈ Supposed covered cells
18
19         else
20             C_{i,j} ∈ is undefined (supposed covered
               or uncovered cell)
21
22     end
23 end
```

While new streaming data flows in the system, synopses in both online and offline parts need to be updated. The maintenance of such synopsis for the online part has been studied in the previous part where one synopsis is kept for every pane and merged at every β time period in the whole sliding window range ω to compute the current *Black Hole* cells. The issue appears when the window range ω is reached for the online part, then old streaming data need to be load-shedded (here the last pane) to have enough free space in memory for online processing. For the following parts of this paper, an overall aggregate synopsis is considered for the offline part and further discussed in the experiment analysis.

5. EXPERIMENTAL RESULTS

This section describes the results extracted from a historical dataset corresponding to the maritime region of Brest and that contains more than 18 millions positions over a period of six months.

5.1 Dataset Description

The dataset used for the evaluation of our processing approach and the Black Hole detection algorithm have been derived from a preliminary study of maritime traffic in the western part of France. This dataset covers a period of 6 months between October 2014 and March 2015. It has been received and parsed in real-time, then stored in a flat file acting as an input for the Black Hole detection. Over this six-month period 24,467,196 AIS messages have been received (i.e., a mean value of more than one message and a half per second), out of which 1,316,689, that is, a ratio of 6.41% that did not comply with the ITU standardized outline [1]. Such erroneous messages have been discarded. The AIS broadcast messages of 27 different types. The dataset only retains messages conveying a position report. Such messages represent in average 81.3% of the received messages (with an average error rate on position reports of 7.3%). The dataset finally contains 18,115,534 positions. The spatial extent of this positions report is illustrated by Figure 6. While most of the data are located in a radio range (15,991,493 located within a circle of 50km), the AIS coverage evolves continuously, and under certain weather conditions, the receiver can obtain positions over the line of sight (2,124,041). The maximum distance of the receiver considered for this study is bounded to 1000 km (other points are considered as outliers).

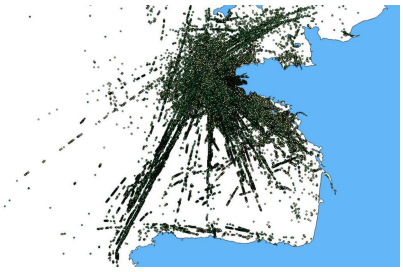

Figure 6: Locations of all messages received during the study

5.2 Methodology Description

The experiments have been made via Flink [2], a *Big Data* platform that provides an architecture permitting to run both batch and stream processing. However, this system doesn't give any support for spatial and spatio-temporal operators. We plan to implement our own operators to design the architecture proposed in [14]. The search for *Black Holes* is an illustration of the kind of problems we want to address with the suggested hybrid architecture. The offline part should permit to prune the space of answers, while online processing should allow to give an answer taking into account the results extracted from offline synopses. The cell size has been fixed empirically to 0.2*0.2 and the quantile chosen for θ determination is 5%. In order to define the offline set, we consider the initial five months of data while the last month is considered as the online one. For the online part, the data is read as if it was arriving in real-time with faster speed and with ordered timestamps (i.e., taking into account unordered timestamp will be part of our future work).

5.3 Offline experiments

As mentioned previously, for each cell of the grid, the number of distinct vessels that have recorded their position on those cells is computed. In Figure 7 each point (cell of the grid) represents the number of positions $Npos_{i,j}(\tau_{Past})$ recorded in the cell $C_{i,j}$, red color for the most recorded cells and white for those with no records. It appears that the maritime area close to Brest is relatively crowded and one can observe the routes commonly used by vessels that go along France and especially the Ouessant fairway. The candidate black holes are cells denoted in white as well as the orange ones with the lowest counts value.

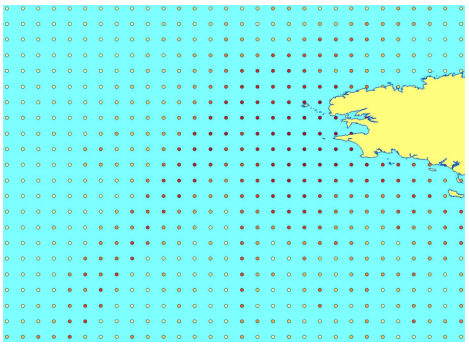

Figure 7: Overall view of the AIS messages considered

5.4 Synopses update and maintenance

So far, an overall aggregate synopsis has been considered for the offline part, but intermediate aggregate synopses can be also taken into account for days, weeks, months etc., for estimating for instance density values. In a stream-oriented system, recent data should be more important than older ones, considering that recent data are more relevant to understand some specific events. Let us consider for our case study candidate *Black Holes* for two successive days, each point represented in the map corresponds to a cell $C_{i,j}$ that belongs to the candidate *Black Hole* cells set (Figure 8).

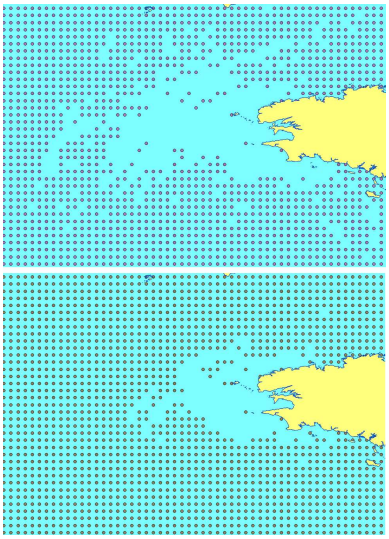

Figure 8: Experimental results (March 28 and 29, 2015)

The fluctuations of the covered cells seems too important to consider only recent data and they may not be so representative of the actual coverage. This being a reason for further experimental studies as a whole aggregate over historical data is considerer in this paper. This example also shows how fluctuations over covered regions can be important for a short time period.

5.5 Experiment results

It has been long observed and not only in the region of study that meteorological changes can happen relatively quickly. The temporal granularity considered is the one of the day as the upper limit for the size of the temporal sliding window for the online part. Three time intervals are studied and discussed: one day, twelve hours and six hours. For one day, and from the online part approximatively 82% empty cells are identified, 3% formally identified as Black Holes cells and 15% of formally identified as covered cells. For twelve hours, still for the online part approximatively 88% empty cells are identified, 1% formally identified as Black Holes cells and 11% of formally identified covered cells. For six hours, the online part gives approximatively 95% empty cells, 1% formally identified as Black Holes cells and 4% of formally identified as covered cells.
Figure 9 illustrates and makes a difference between empty cells for six hours (represented by both triangles and points) and one day (triangles only).

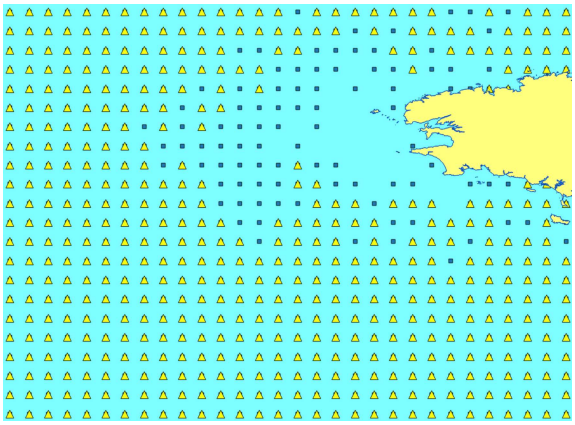

Figure 9: Empty cells for six hours and one day

Figure 10 illustrates our cell classification over a six hour sliding window for Black Hole identification process. Empty cells that are undefined (i.e. empty and that belongs either to the supposed covered cells or supposed Black Hole cells) appears in the map as triangles, while supposed covered cells (identified by cross with one month historical data) are represented by squares. Formally identified Black Holes (identified by pentagon) are found because some vessels have crossed them without emitting any signal, and formally covered cells correspond to region with no points near from Brest coasts.

Whatever the level of granularity, this entails the need for additional information extracted from historical data in order to make a real difference between the cells that have been uncover but which are in fact potentially covered (i.e., the ones where no vessels have crossed them) and the ones that are *Black Holes*.

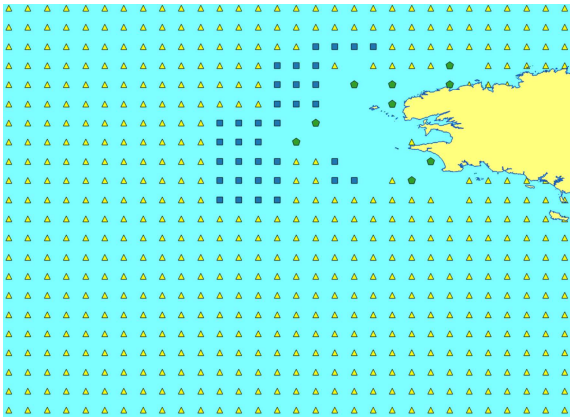

Figure 10: cell classification over six hours and considering one month of historical data

6. PERSPECTIVES AND FUTURE WORK

To the best of our knowledge the concept of black hole, as defined in our study, has not been yet addressed. In [6] the authors use the term *Black Hole* to search for, from an oriented network, the vertices that have more outgoing than arrival flows (and conversely), but that concept is different from the one developed in our paper. One related subject related to ours is the identification of dense queries that have been addressed in both online [5], where the authors use a quadtree to record new arriving positions, and offline points of view [4] where the authors search for dense cells associated to moving objects. With our approach, a subset of the *Black Hole* cells are identified, but still remaining issues concern the derivation of specific behaviors associated to empty cells and integration of the granularity dimension for both online and offline parts. Some other experiments should be developed to observe the nature of regions determined as *Black Holes* and show the relevance of our algorithm to detect those specific regions. Finally, index techniques should be further investigated to update *Black Hole* regions and queries that make use of our *Black Hole* detection process.

Indeed, while the identification of the current black holes is performed, such black holes can be used to detect trajectory anomalies as well as suspicious behaviors. AIS messages are emitted from boats accordingly to some rules defined in [1]. Unfortunately, transponders responsible for sending those messages can be switch off by the sailers in charge on board. While not knowing exactly where the black holes are, one couldn't make any difference between vessels that enter an region with no coverage, and those that intentionally switch off their AIS transponders. Such behaviors are usually related to malversations as flow of illicit products, illegal immigration, overfishing or pollution.
Future works will be oriented towards the identification of on purpose switch-off AIS or detection of fake messages. Indeed, when no positions are received from a vessel outside of uncovered cells, one can infer that such vessels have switched off their emitters. Regarding fake messages, this case is another one to ideally take into account, but any emergence of a signal from a usually uncovered cell has to be further examined.

7. CONCLUSION

The research presented in this paper introduces the modelling concept of *Black Holes* that denotes maritime regions from where signal emitted by vessels cannot be received by coastal antennas during a given time period. This specific case requires the development of a hybrid computing approach that takes into account the continuous flow of information involved, as well as uncertainties inherent to the problem. The principles behind this hybrid approach have been first described in [14], this allowing us to identify candidate black holes regions by merging offline and online synopse. Under these principles, this paper develops a computational setup that (1) identifies candidate black holes from offline data, (2) determines potential black holes regions from incoming streaming data and (3) merges information from (1) and (2) in order to find probably not covered regions (4). However, additional mechanisms are still required to maintain the offline sysnopses while streaming data are transferred to the historical data. The experiments developed show the relevance of such a hybrid approach for moving objects queries and has been applied to the Brest Dataset described in Section 5. The different steps involved during the process have been identified and should be still considered for further works in other hybrid queries.

8. REFERENCES

[1] *Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band.* Recommendation ITU-R M.1371-5 (02/2014), 2014.

[2] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, Felix Naumann, Mathias Peters, Astrid Rheinländer, Matthias J. Sax, Sebastian Schelter, Mareike Höger, Kostas Tzoumas et Daniel Warneke : The stratosphere platform for big data analytics. *VLDB J.*, 23(6):939–964, 2014.

[3] Sirish Chandrasekaran et Michael Franklin : Remembrance of streams past: Overload-sensitive management of archived streams. *In Proceedings of the Thirtieth International Conference on Very Large Data Bases*, VLDB '04, pages 348–359, 2004.

[4] Marios Hadjieleftheriou, George Kollios, Dimitrios Gunopulos et Vassilis J. Tsotras : On-line discovery of dense areas in spatio-temporal databases. *In Advances in Spatial and Temporal Databases, 8th International Symposium, SSTD 2003, Santorini Island, Greece, July 24-27, 2003, Proceedings*, pages 306–324, 2003.

[5] Xing Hao, Xiaofeng Meng et Jianliang Xu : Continuous density queries for moving objects. *In Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, Mobide 2008, June 13, 2008, Vancouver, British Columbia, Canada, Proceedings*, pages 1–7, 2008.

[6] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang et Lei Zou : Detecting urban black holes based on human mobility data. *In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 35. ACM, 2015.

[7] Christian S. Jensen, Dan Lin, Beng Chin Ooi et Rui Zhang : Effective density queries on continuouslymoving objects. *In Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 71, 2006.

[8] Corrado Loglisci et Donato Malerba : Mining dense regions from vehicular mobility in streaming setting. *In Foundations of Intelligent Systems - 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings*, pages 40–49, 2014.

[9] Nathan Marz : *Big data : principles and best practices of scalable realtime data systems.* O'Reilly Media, [S.l.], 2013.

[10] Jinfeng Ni et Chinya V. Ravishankar : Pointwise-dense region queries in spatio-temporal databases. *In Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 1066–1075, 2007.

[11] Giuliana Pallotta, Michele Vespe et Karna Bryan : Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, 15(6):2218–2245, 2013.

[12] C. Ray, R. Gallen, C. Iphar, A. Napoli et A. Bouju : Deais project: Detection of ais spoofing and resulting risks. *In OCEANS 2015 - Genova*, pages 1–6, May 2015.

[13] Branko Ristic, Barbara F. La Scala, Mark R. Morelande et Neil J. Gordon : Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. *In 11th International Conference on Information Fusion, FUSION 2008, Cologne, Germany, June 30 - July 3, 2008*, pages 1–7, 2008.

[14] Loic Salmon, Cyril Ray et Christophe Claramunt : A hybrid approach combining real-time and archived data for mobility analysis. *In 6th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS)*, page 6, 2015.

9. ACKNOWLEDGMENTS

This research has been started during a scientific research mission of the first author at University of Piraeus (Piraeus, Greece). Helpful discussions with Kostas Patroumpas, Yanis Theodoridis and Nikos Pelekis during this specific mission are kindly acknowledged. This mission has been supported by the "Laboratoire d'Excellence" LabexMER (ANR-10-LABX-19) and co-funded by a grant from the French government under the program "Investissements d'Avenir".