# Chapitre 1 : L’analyse de comportements dans le domaine de la surveillance maritime

## 1.1. Introduction

Les progrès technologiques en systèmes de localisation (GPS, RFID, etc.), de télétransmission (VHF, satellite, GSM, etc.), en systèmes embarqués et leur faible coût de production a permis leur déploiement à une large échelle. Enormément de données sur le déplacement d'objets sont acquises par le biais de ces technologies et utilisées dans diverses applications comme le suivi de foules de piétons (Giannotti et al., 2011) (Buard & Christophe 2012), le suivi des animaux (Buard & Brasebin 2011), la gestion du trafic routier, aérien et la surveillance maritime (Etienne 2011). Des bases de données de suivi d'objets mobiles sont mêmes mises en libre utilisation sur internet comme les bases de données AISHUB[23] de suivi de navires.

La plupart du temps ces données sont utilisées pour des besoins temps-réel. L'analyse *a posteriori* des historiques de données peut présenter des perspectives très intéressantes pour l'analyse des comportements et la compréhension des mouvements, des situations et de leur interconnexion.

Dans ce chapitre, nous présentons l'analyse de comportements d'objets mobiles et l'intérêt produit pour plusieurs domaines d'application. Nous décrivons ensuite, les approches d'acquisition et de construction de connaissances utilisées pour la modélisation des comportements. Puis, nous allons nous focaliser sur le domaine maritime et présenter les principales méthodologies de modélisation qui ont été utilisées pour l'analyse et la modélisation des comportements de navires. Ces travaux ont montré l'intérêt accru des méthodes d'analyse de comportement pour la surveillance maritime. Enfin, nous allons présenter une synthèse récapitulative de ces méthodologies (description, avantages et limites).

## 1.2. Définition d'un comportement

Le comportement est défini dans le dictionnaire LAROUSSE[24] comme étant une action, une réaction, un fonctionnement et une évolution spatio-temporelle dans certaines situations d'un objet. Cet objet peut être un véhicule, un navire ou toute autre chose. Dans la suite de ce travail, nous considérons un comportement comme un ensemble de

[23] http://www.aishub.net/
[24] http://www.larousse.fr/dictionnaires/francais/comportement/17728

mouvements dans une situation. Le mouvement est un changement de position et la situation est une combinaison de caractéristiques de l'objet, de son contexte et de son environnement d'évolution spatio-temporel. Le mouvement est influencé par la situation qui peut être liée à des facteurs internes à l'objet (objectif, panne, etc.) ou des facteurs externes (mouvements des voisins, météorologie, etc.) (Le Pors et al., 2009) dans (Etienne 2011). Un navire qui change son cap par exemple pour éviter de rentrer en collision avec un autre navire est un mouvement influencé par l'environnement dans lequel il évolue.

## 1.3. L'analyse de comportements

L'une des problématiques de l'étude des comportements d'objets mobiles est de comprendre les mouvements, les interactions entre les objets, leur environnement et comment il est possible d'interpréter ces comportements à partir des propriétés quantitatives et qualitatives observées des déplacements. L'étude des déplacements a pour objectif de faciliter la compréhension des causes, des mécanismes, des patrons spatio-temporels du mouvement et leur rôle dans l'évolution du système (Nathan et al., 2008). Les régularités et les irrégularités dans les mouvements peuvent être décrites par des motifs (patrons ou modèles). Un motif peut être considéré comme la synthèse des mouvements, une description des comportements et un modèle de prédiction.

Les positions des mouvements réels sont représentées par des coordonnées affichées sur une cartographie. La jonction de positions d'un même objet permet de représenter l'évolution de l'objet par une trajectoire qui caractérise son déplacement. Une trajectoire est donc une suite de positions ordonnées dans le temps.

L'analyse de comportements des objets à partir de ces trajectoires ouvre des perspectives intéressantes dont la compréhension, la description et la prédiction de la mobilité. Cette capacité de recueille et d'analyse de quantités massives de données de déplacement ont transformé plusieurs domaines comme la biologie et les sciences sociales en permettant d'interpréter les comportements (Lazer et al., 2009). Qu'en est-il du domaine de la surveillance maritime ?

La problématique d'analyse automatique des comportements de navires a sollicité un fort intérêt au niveau des organismes et centres de recherche traitant de la question de la sûreté maritime. Le projet Predictive Analysis for Naval Deployment Activities (PANDA) (Darpa 2005) du ministère de la défense américain a pour objectif d'évaluer automatiquement les comportements de tous les grands navires et pas seulement ceux qui font l'objet d'un suivi (*Vessel Of Interest*), afin de déterminer quels sont ceux qui s'écartent de leur comportement normal et attendu[25]. L'objectif est d'indiquer les menaces relatives à la sûreté maritimes par analyse automatique de comportements de navires. Ce projet est considéré comme initiateur et a inspiré de nombreux travaux dont le projet SARGOS (Chaze et al., 2012) pour la protection de plateformes pétrolières, le projet SECMAR pour la surveillance des ports, le projet ScanMaris (Morel et al., 2008; Morel et al., 2010), TaMaris (Morel et al., 2011), SisMaris (Morel 2009) pour l'analyse du trafic maritime et les projets I2C et Perseus pour l'analyse du trafic maritime et l'interopérabilité des systèmes de surveillance européens.

La recherche et développement (R&D) pour la Défense Canadienne, s'est intéressé aussi à cette question en proposant l'analyse visuelle des données maritimes (Gouin et al., 2011; Lavigne & Gouin 2011). D'autres travaux académiques comme les travaux de N. Willems et M. Riverio ont proposé des méthodes d'analyse visuelle des données maritimes (Willems et al., 2009; Willems 2011; Willems et al., 2011; Riveiro et al., 2008; Riveiro & Goran Falkman 2009; Riveiro & Göran Falkman 2011).

Afin de détecter automatiquement les comportements anormaux, il est possible de modéliser ce qui est anormal pour identifier les comportements qui suivent ce modèle ou de modéliser ce qui est normal pour identifier les comportements qui s'écartent de cette normalité. Les deux approches sont utilisées pour la construction de modèles de comportements (Kai-Lin et al., 2013). Prenons l'exemple d'une étude de mouvements d'utilisateurs de parking à partir d'enregistrements vidéos (Figure 1-1). La découverte d'un mouvement inhabituel ou suspect faisant des déplacements aléatoires entre les véhicules du parking peut décrire un comportement d'un voleur qui cherche quelle voiture voler (partie B de la Figure 1-1). La partie A décrit un comportement normal et la partie B, un comportement aléatoire qui s'écarte de la normalité.

[25] https://www.fbo.gov/spg/ODA/DARPA/CMO/BAA05-44/listing.html

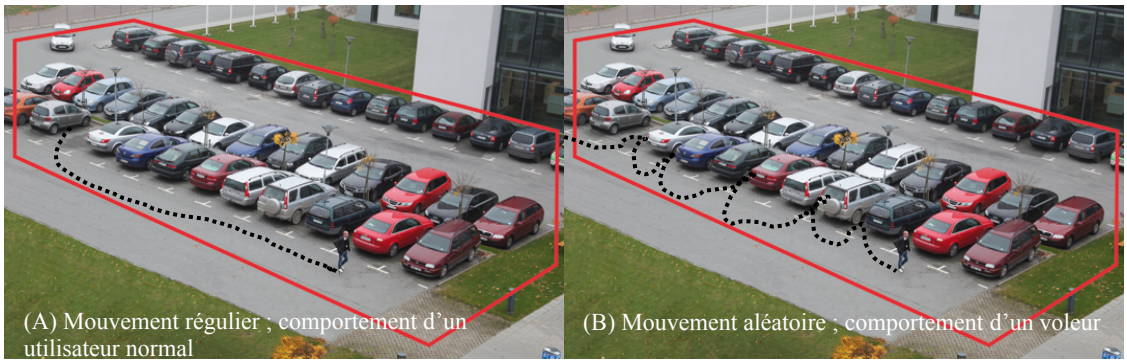


Figure 1-1 : Identification d'un comportement suspect à partir de l'analyse du déplacements d'une personne sur une vidéo de surveillance (Han, 2010) [26].

Un autre exemple présente ci-dessous une trajectoire inhabituelle d'un navire qui fait un demi-tour et revient s'arrêter devant la côte. Cette trajectoire peut correspondre à une dérive d'un navire comme le montre la Figure 1-2 représentant la dérive puis l'échouement du Costa Concordia sur les côtes italiennes en janvier 2012.


Figure 1-2 : Trajectoire du Costa Concordia au moment de l'échouement.

Le mouvement à lui seul n'est pas toujours suffisant pour analyser les comportements, il est alors nécessaire d'étudier la situation dans laquelle évolue le mouvement. L'analyse de situations peut aider à mieux qualifier un comportement.

[26] Exemple inespéré de la présentation de J. Han à DASFAA 2010, slide 97, les images sont récupérées de http://www.axis.com/fr/academy/10_reasons/camera_intelligence.htm

## 1.4. Approches d'acquisition et de construction de connaissances pour la modélisation de comportements

Dans la construction de connaissances, deux approches peuvent être utilisées (Roddick & Lees 2009) :

- Approche déductive : modélisation du monde réel avec des modèles mathématiques dont la prévision ou la description des modèles est calculée,
- Approche inductive : correspondance de modèles dont la prévision est faite par rapport aux observations passées :
  - A partir de l'expérience (*top-down*) dont l'observation et le raisonnement est fait par l'humain,
  - A partir de l'analyse des bases de données[27] (*bottom-up*)
    - Analyse statistiques,
    - Analyse visuelle de données,
    - Fouille de données.

La déduction permet de déduire des conséquences observables à partir d'hypothèses générales (prémisses) (Martin 2012). Contrairement à la déduction, l'induction produit des prémisses et offre la possibilité de générer de nouvelles connaissances. Cette approche de raisonnement passe d'observations particulières à des hypothèses générales en faisant des restrictions sur un espace d'hypothèses jusqu'à ce qu'une description restrictive de cet espace puisse être formée (Roddick & Lees 2009).

L'induction part de l'idée que « *la répétition d'un événement augmente la probabilité de le voir se reproduire* ». La répétition d'un événement n'implique pas forcement sa reproduction. Par conséquent, cette approche privilégie l'observation, l'analyse et l'expérimentation pour tirer des conclusions générales (Martin 2012). La validation des connaissances issues de l'induction est donc primordiale pour éviter d'avoir des connaissances inutiles ou erronées. C'est pourquoi, nous considérons que cette approche est à privilégier pour la découverte de connaissances.

[27] Une base de données est un ensemble de données organisées dans des structures pour faciliter la gestion des données (ajout, suppression, mise à jour, interrogation, etc.) et leur utilisation par des applications informatiques. (Gardarin 2011).

La Figure 1-3 résume bien les approches d'acquisition et de construction de connaissances à savoir, la déduction et l'induction. La fouille de données appartient à la correspondance de modèles (Pattern Matching) qui correspond à l'approche inductive.

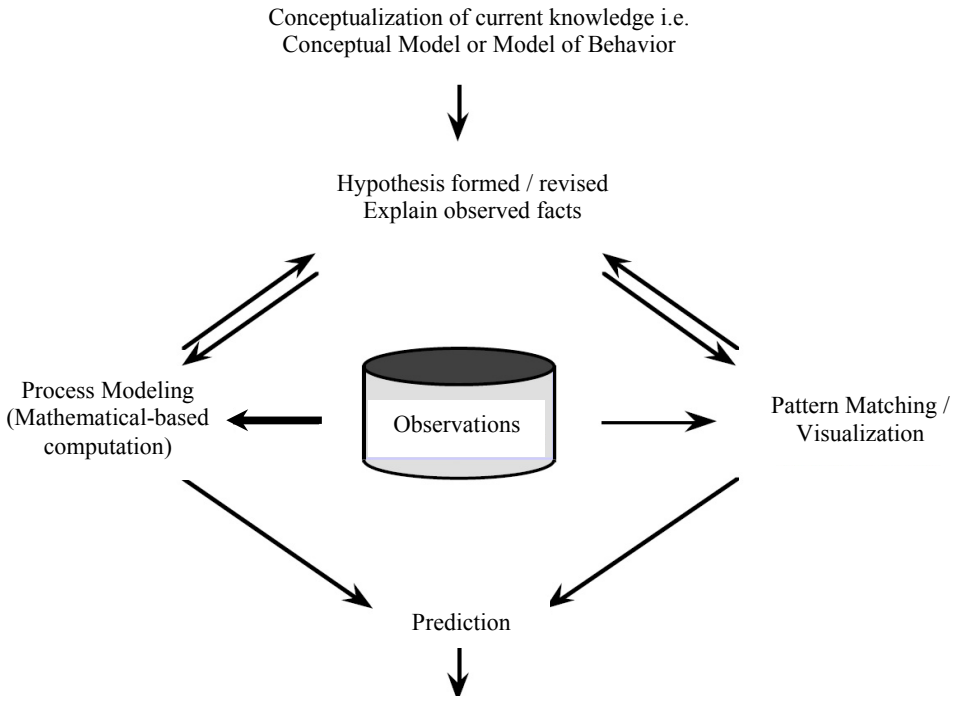


Figure 1-3 : Les approches d'acquisition de connaissances
(Roddick & Lees 2009).

Dans l'approche inductive, deux sous approches sont utilisées : une approche *top-down* qui permet la découverte de connaissances à partir du raisonnement humain sur des observations et une approche *bottom-up* guidée par des explorations automatiques de bases de données d'observations.

Nous les avons appelés ainsi car dans l'approche *top-down,* le mécanisme d'apprentissage commence à partir de l'expertise alors que dans l'approche *bottom-up,* il commence à partir d'observations stockées dans des bases de données.

## 1.4.1. Approche *top-down*

Dans cette approche basée sur l'expertise, la méthode de brainstorming est souvent utilisée ((Roy 2008) ; ((Nilsson et al., 2008) dans (Laere & Nilsson 2009))). Le *brainstorming* consiste à réunir des experts pour l'acquisition de connaissances sur les

comportements habituels, inhabituels ou suspects. Ces méthodes sont compliquées à mettre en œuvre et couteuses ; les scénarios en sortie dépendent beaucoup de l'expérience personnelle de chaque expert. De plus, elles ne permettent pas la découverte de connaissances nouvelles.

### 1.4.2. Approche *bottom-up*

Cette approche d'apprentissage basée sur les données, utilise plusieurs méthodologies pour l'acquisition de connaissances. Parmi ces méthodologies nous pouvons citer les analyses statistiques, la visualisation, l'apprentissage automatique et la fouille de données.

L'intérêt de la fouille de données est qu'elle rassemble plusieurs techniques statistiques, d'apprentissage machine, algorithmiques, etc. pour exploiter les avantages de chacune de ces techniques. De plus, les avancées informatiques en calcul et en stockage permettent l'analyse rapide de grands volumes de données.

## 1.5. Méthodologies d'analyse de comportements de navires

Le domaine d'application qui a été privilégié dans cette thèse est l'étude des comportements de navire. Pour l'analyse de ces comportements, différentes méthodologies ont été proposées. Dans les sous sections suivantes, nous allons présenter l'analyse statistique, l'analyse visuelle et la fouille de données.

### 1.5.1. Analyse statistique

Les statistiques peuvent être définies comme un ensemble de techniques et méthodes permettant le traitement et l'interprétation des données. Selon J. Tukey (Tukey 1980), il existe deux approches d'analyse statistiques, une analyse exploratoire et une autre confirmatoire. L'analyse exploratoire part des données qu'elle permet d'analyser sous différentes facettes pour mettre en évidence des structures cachées par le volume important des données et aider ainsi à construire des modèles (Ladiray 1997). Le volume de données analysé peut atteindre plusieurs dizaines de milliers d'individus et des dizaines de variables. Cette analyse peut précéder une analyse confirmatoire pour mettre en exergue les propriétés qualitatives, quantitatives des données et poser des hypothèses

plausibles. Pour infirmer ou confirmer une hypothèse, c'est l'analyse confirmatoire qui intervient. Cette analyse est faite sur un échantillon de données représentatif de la population globale puis les informations résultantes sont inférées à la population entière avec un certain degré de confiance. Cette inférence peut être vue comme une extrapolation de nouvelles informations à partir de celles connues déjà. L'analyse exploratoire et confirmatoire peuvent et doivent passer d'une à l'autre car elles sont complémentaires.

Dans (Etienne et al., 2010), les auteurs ont proposé une analyse statistique appliquée à un historique de déplacement de navires qui est décrit par des trajectoires spatio-temporelles. Leur proposition se fond sur une ingénieuse extension de la boite à moustaches[28] à l'analyse spatio-temporelle. Une médiane spatiale et temporelle sont calculées dans l'objectif de les utiliser dans la définition du couloir spatio-temporel (Figure 1-4). Le couloir spatial et temporel est défini d'une manière à ce qu'ils incluent 90% des positions les plus proches de la médiane. C'est le 9$^{ème}$ décile[29] qui a été pris comme séparateur entre les 90% des positions et les 10% restantes. Le couloir spatio-temporel résultant de l'analyse a été utilisé comme motif pour détecter les comportements inhabituels à partir de trajectoires. Les comportements inhabituels sont découverts par intersection d'une position avec ce motif pour la qualifier à un instant t de position en retard, à l'heure, en avance, sur la route ou en dehors du couloir spatio-temporel.

[28] Appelée en anglais Box Plot, c'est une représentation graphique en boite permettant de représenter schématiquement la distribution d'une série statistique.

[29] Est une statistique descriptive contenant neuf valeurs séparant un ensemble d'individus en parts égales en effectif.

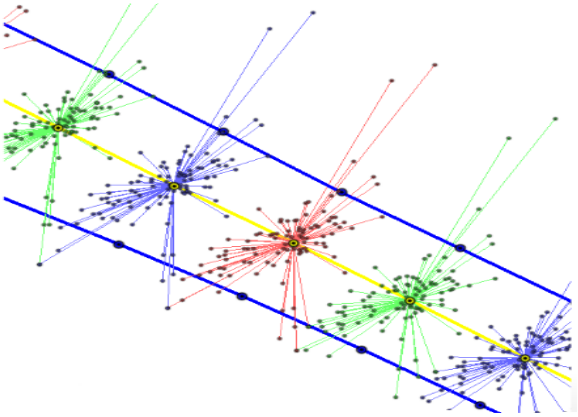

Figure 1-4 : Définition de positions médiane, la trajectoire médiane et du couloir spatio-temporel pour l'identification des comportements inhabituels de navires (Etienne et al., 2010).

## 1.5.2. Analyse visuelle

L'analyse de comportements de navires par exploration visuelle des données utilise des techniques de géo-visualisation pour présenter les informations d'une manière à pouvoir en extraire du sens. Willems (Willems et al., 2011) par exemple propose un algorithme de visualisation des trajectoires de navires basé sur la densité : plus un navire navigue rapidement, moins il laisse de traces. Ces traces représentent des informations sur les routes maritimes, des zones d'ancrage, les lenteurs et déplacements rapides et bien d'autres informations (Figure 1-5).

Dans ce genre d'approche c'est à l'utilisateur d'interagir avec les représentations visuelles pour identifier les anomalies et construire des connaissances. Les techniques de visualisation vont aider à améliorer les capacités de perception, de compréhension et de raisonnement de l'être humain (Riveiro & Goran Falkman 2009).

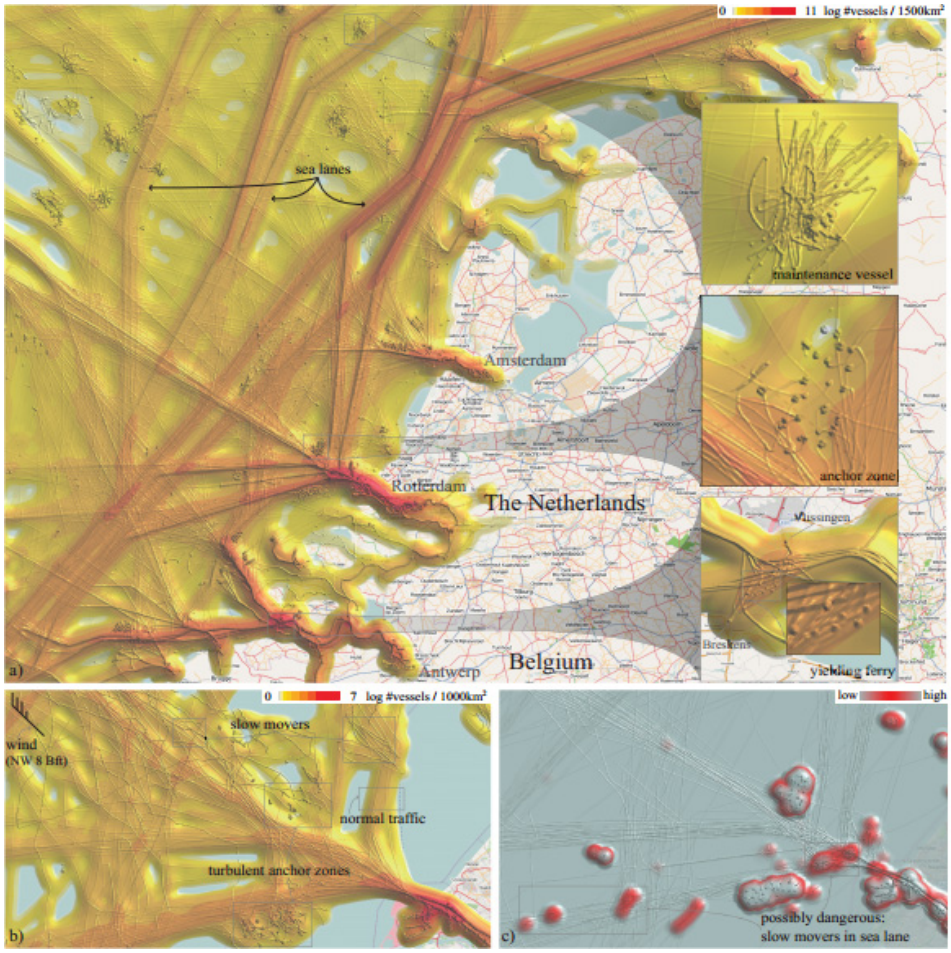


Figure 1-5 : Analyse des densités de déplacement de navires (Willems et al., 2009).

Dans (Riveiro & Goran Falkman 2009), les auteurs ont utilisé un graphique de fréquence de vitesses pour extraire par visualisation des comportements inhabituels de navires par rapport à leur vitesse (profil de navires lents et trop rapides). Dans un autre travail (Etienne et al., 2011), les auteurs utilisent un indice de similarité pour détecter par géo-visualisation les trajectoires outliers qui ne suivent pas la même progression temporelle et spatiale que le groupe de trajectoires.

L'avantage de l'analyse visuelle et géo-visuelle est lié à l'intégration de la dimension humaine dans l'exploration de données. Dans cette analyse, les utilisateurs interagissent avec différentes formes de visualisation dans l'objectif d'extraire des connaissances. Cette participation des utilisateurs dans la construction des connaissances est intéressante car ils connaissent le domaine d'application.

Les inconvénients de cette méthode d'analyse de comportements sont la complexité de certaines visualisations et le temps requis pour l'affichage et l'exploration des données. Prenons l'exemple d'une forme de géo-visualisation appelée Trajectory Wall (G. Andrienko et al., 2014) qui montre la complexité de certaines représentations pour des utilisateurs non habitués à ce genre de représentation (Figure 1-6). Les objets rapides sont représentés en vert et les objets lents en rouge. La représentation circulaire quant à elle, permet de voir la moyenne des intervalles de vitesses par tranche horaire.

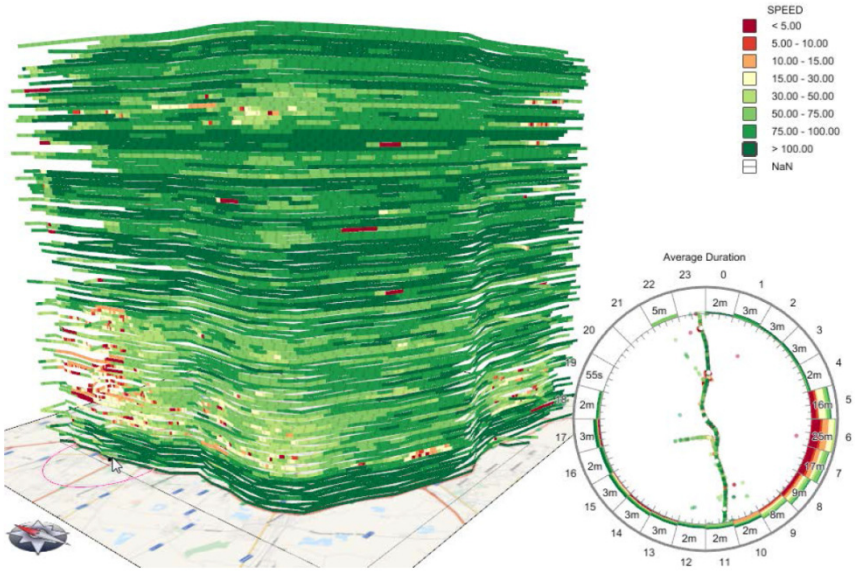


Figure 1-6 : Géo-visualisation de comportements d'objets mobiles par distribution d'intervalles de vitesses dans le temps
(G. Andrienko et al., 2014).

Généralement le problème ne réside pas dans la complexité des visualisations mais plutôt dans l'inadaptation de ces visualisations aux profils des utilisateurs et à l'analyse de comportement effectuée. C'est pour cela que Vatin (Vatin & Napoli 2013a)(Vatin & Napoli 2013b), propose dans son travail de recherche, une approche permettant d'adapter automatiquement la visualisation selon le profil de l'utilisateur et l'analyse de comportements qu'il souhaite effectuer.

## 1.5.3. Analyse par fouille de données

Par analogie à la recherche des pépites d'or dans un gisement, la fouille de données vise à extraire des informations cachées par analyse globale et à découvrir des modèles,

appelés motifs, difficiles à percevoir directement du fait du volume important des données, du nombre de variables à considérer et enfin du fait qu'il y ait des hypothèses imprévisibles (Gardarin 2011).

La fouille de données peut être vue comme un générateur automatique d'hypothèses pour examen et validation dans le but de transformer les données en connaissances. Plusieurs définitions de la fouille de données existent. Parmi ces définitions, nous pouvons citer celle de Frawley (Frawley et al., 1992) qui présentent la fouille de données comme une extraction non triviale de connaissances implicites et potentiellement utiles à partir des données. Parsaye (Parsaye 1995), le définit comme un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données. Une autre définition plus anecdotique est celle de Chorafas : la fouille de données consiste à torturer les données jusqu'à ce qu'elles avouent.

Il y a dans la littérature des travaux initiateurs dans l'utilisation des méthodes de fouille de données au problème de surveillance maritime (LeBlanc & Rucks 1996) (Torun & Düzgün 2006) (Marven et al., 2007) (Etienne et al., 2010). Nous allons décrire quelques-uns de ces travaux.

#### 1.5.3.1. Analyse de situations par clustering d'événements

Le Blanc et Rucks sont parmi les pionniers à avoir utilisé la fouille de données au domaine maritime (LeBlanc & Rucks 1996). Ils ont appliqué une méthode de clustering des plus proches voisins KNN (Anderberg 1973) sur un échantillon de 936 cas d'accidents qui se sont produits sur le fleuve du Mississippi. Cette méthode de clustering avait permis d'identifier quatre clusters nommés zones dangereuses, mauvaises conditions de navigation, accidents pouvant être évités, accidents qui n'auraient pas dû arriver. Les zones dangereuses regroupent les accidents qui se sont produits dans des parties dangereuses du fleuve.

Après avoir identifié ces clusters une analyse discriminante est réalisée pour déterminer les valeurs d'attributs qui séparent le mieux les cas d'accidents dans les différents clusters. Cette analyse discriminante a comme objectif de prédire le groupe de nouveaux cas d'accidents et de montrer l'intérêt de l'utilisation des systèmes de suivi de

navires. Les accidents classés comme pouvant être évités sont caractérisés par la non utilisation de ces systèmes.

L'un des inconvénients de cette méthode est le fait qu'elle construit les clusters en connaissant au préalable les zones dangereuses, comme les zones avec des rochers affleurant, les passages étroits et les zones à forte densité de trafic. La vocation de ce travail est d'utiliser les résultats obtenus pour classer les nouveaux accidents selon les quatre classes et à montrer ainsi l'intérêt d'utiliser un système de suivi de navires.

Torun et de son équipe (Torun & Düzgün 2006) ont aussi utilisé la fouille de données dans le domaine de la sécurité maritime. L'existence de points étroits dans le détroit[30] d'Istanbul, le fait qu'il soit rocheux, qu'il contienne des virages et des courants, etc., augmente les risques du transport maritime dans cette zone. Cela a motivé l'équipe de Torun à proposer un modèle linéaire de vulnérabilité[31] pour calculer plus précisément les risques en se basant sur les données de danger. Le danger peut être défini comme étant la menace qui pèse sur la sécurité, la sûreté, les personnes, l'activité et les biens. La particularité du danger est qu'il ne dépend pas de l'existence d'objets vulnérables mais existe indépendamment.

Torun et son équipe utilisent les techniques de fouille de données spatiales pour évaluer la vulnérabilité des personnes et des zones par rapport au transport de pétrole et de gaz dans le détroit d'Istanbul. Les techniques utilisées sont le clustering spatial avec les algorithmes K-means et ISODATA, l'autocorrélation, les hots spots et l'analyse de densité.

Le résultat de ce travail est présenté sur la Figure 1-7 où la partie (a) identifie les clusters de distribution d'accidents obtenus par l'indice de Moran[32]. Les distributions ayant un nombre d'accidents plus élevé que la majorité ou celles à proximité ont une couleur noir foncé. La partie (c) est le résultat de la superposition des zones de risques d'accidents (a) avec la vulnérabilité des personnes (b) décrivant les endroits pouvant avoir des conséquences graves en cas d'accident à proximité.

[30] Passage maritime naturel
[31] Conséquences de réalisation d'un risque sur les objets exposés.
[32] Mesure statistique d'autocorrélation proposée par Patrick Alfred Pierce Moran.

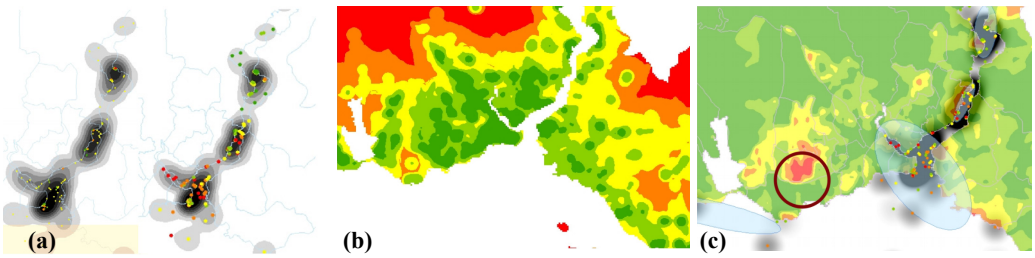


Figure 1-7 : (a) Zones à risques d'accidents, (b) Vulnérabilité des personnes, (c) Chevauchement de (a) et (b) (Torun & Düzgün 2006).

Dans un article de C. Marven (Marven et al., 2007), les auteurs ont montré l'intérêt des méthodes exploratoires à l'analyse des risques maritimes. Ils ont appliqué des méthodes d'analyse spatiale pour aider les gardes côtes canadiens à planifier leur recherche, prendre des décisions et améliorer le sauvetage en mer. Pour cela ils se sont basés sur le clustering spatial pour identifier et visualiser les concentrations d'incidents et d'accidents maritimes.

Les auteurs de l'article montrent l'opportunité d'appliquer au domaine maritime, des méthodes d'analyse spatiale utilisées en épidémiologie et criminologie. Parmi ces méthodes, on trouve Spatial and Temporal Analysis of Crime (STAC[33]) et Nearest Neighbour Hierarchical Cluster Analysis (NHH).

Concernant les deux dernières méthodes, ce ne sont pas des limites qui sont dressées mais une perspective. Les résultats des méthodes auraient pu être utilisés pour faire une étude de causalité des accidents, afin de comprendre les causes des concentrations anormales. La matérialisation de ces concentrations d'accidents sous forme de zones dans une base de données pourrait alerter en temps-réel des navires fréquentant ces zones.

### 1.5.3.2. Analyse du comportement par clustering de trajectoires

Dans (Etienne et al., 2010), les auteurs ont proposé une méthode d'extraction de trajectoires homogènes appelée Groupe Homogène de Trajectoires (GHT) qui donne de meilleures performances de calcul que la méthode T-Clustering, proposée par Andrienko

[33] http://www.icjia.state.il.us/public/index.cfm?metaSection=data&metaPage=stacfacts

(G. Andrienko et al., 2009) et intégrée dans M-Atlas (voir section 2.3.2). Concernant les résultats obtenus par les deux méthodes, ils sont presque identiques (Figure 1-8). Cette extraction de trajectoires ayant des mouvements similaires est basée sur deux notions différentes, à savoir les graphes d'intérêt et la similarité entre trajectoires (Etienne et al., 2008)(Etienne et al., 2009).

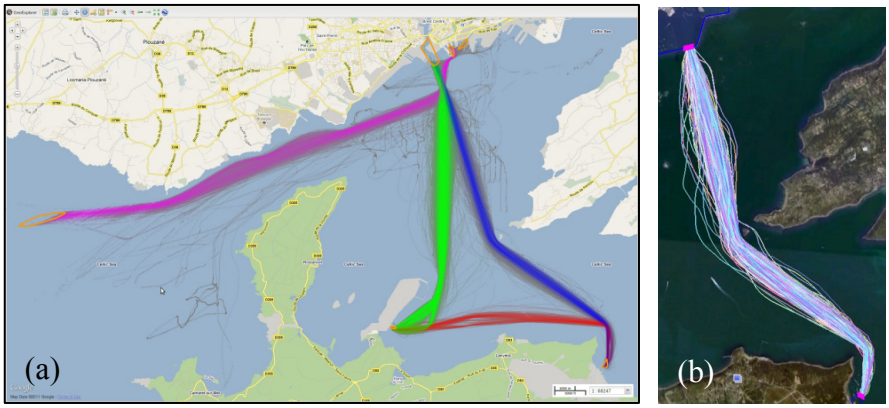


Figure 1-8 : Extraction de groupes de trajectoires-(a) par T-Clustering et (b) par GHT (Etienne et al., 2010).

La méthode de fouille de données proposée par Etienne (Etienne et al., 2010) permet d'extraire des routes types et des couloirs spatio-temporels pour des itinéraires donnés. La route type est construite en se basant sur le calcul de médianes à chaque ensemble de positions d'un GHT.

Le point fort de cette méthode est sa rapidité de calcul. Elle utilise des critères de sélection qui permettent d'optimiser les temps de réponses. Son point faible est dans le nombre important d'étapes de préparation des données au préalable de l'analyse spatio-temporelle. En effet, pour traiter et nettoyer les trajectoires, il faut suivre les étapes suivantes :

- Définir pour chaque trajectoire une position de départ et d'arrivée,
- Couper les trajectoires après un temps estimé à un trajet entre la position de départ et d'arrivée, pour ne pas avoir des allers-retours,
- Filtrer le groupe de trajectoires homogènes par l'algorithme Douglas-Peucker (2.2.3.1.4) pour gommer les imprécisions et les aberrations,

- Projeter les positions de départ sur une ligne de biais pour recaler la dimension spatiale (Figure 1-9),

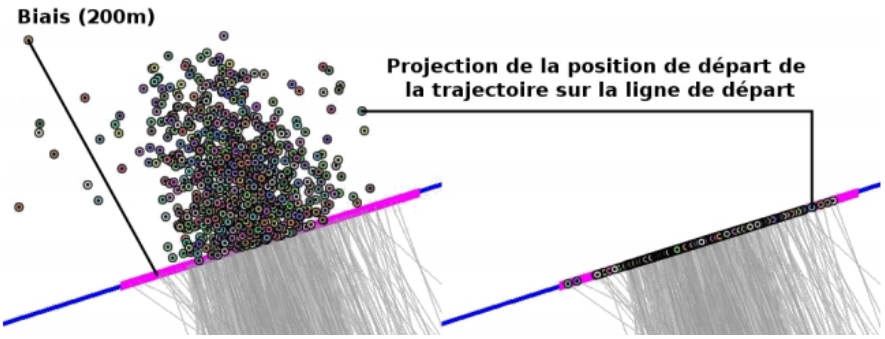


Figure 1-9 : Recalage spatial dans la méthode GHT (Etienne et al., 2010).

- Ré-échantillonner les positions des trajectoires homogènes pour avoir le même pas d'échantillonnage spatial. Par exemple, mettre une position chaque 100 mètres,

- Normaliser la dimension temporelle pour travailler sur des temps relatifs (temps écoulés depuis la position de départ).

Ces travaux montrent bien l'intérêt de l'utilisation de la fouille de données au domaine de la surveillance maritime : analyse de comportements inhabituels (navire en retard, en avance, en dehors de la route habituelle, etc.), découverte de routes types, de zones vulnérables, de concentrations d'accidents et la classification automatique de ces accidents.

## 1.6. Méthodologies de modélisation de comportements de navires

Différentes méthodologies de modélisation ont été utilisées pour l'analyse de comportements de navires. Dans les sous sections suivantes, nous allons décrire les principales méthodologies à savoir la modélisation par règles d'inférences, la modélisation ontologique et par classifieur Bayésien.

### 1.6.1. Modélisation par règles d'inférence

Le raisonnement automatique est un sous-champ de l'Intelligence Artificielle[34]. Il permet de simuler le raisonnement humain sur une machine pour déduire de nouvelles connaissances à partir de flux d'événements en entrée (faits avérés) et des connaissances mémorisées au préalable. Nous distinguons les faits issus d'événements en entrée du système de raisonnement que nous appelons *faits avérés* et les faits qui sont déduits par le système que nous appelons *faits inférés*.

Les connaissances dans un système de raisonnement sont souvent codées sous forme de règles (généralisation d'exemples) ou de cas (exemples). Une règle est de la forme « si *Antécédent* alors *Conséquent* » telle que, « *Antécédent* » et « *Conséquent* » sont des expressions de conjonction de disjonctions des occurrences d'objets. Une connaissance permet de mettre en relation des informations connues (« *Antécédent* ») et des informations qu'on cherche à déduire (« *Conséquent* ») ou des actions que l'on veut exécuter comme : si une condition est vérifiée alors déclencher une alerte (Jones 2007). Un cas est une description d'un problème avec sa solution associée. Ce paradigme enregistre des cas sources résolus dans une base de cas pour résoudre de nouveaux problèmes appelés cas cibles. Suivant la formulation de la connaissance, plusieurs types de raisonnement peuvent être appliqués. Les plus utilisés sont le raisonnement déductif et le raisonnement analogique. Dans le déductif, les valeurs en sortie sont déduites à partir des valeurs en entrée et ; dans le raisonnement analogique, le nouveau problème est ramené à un problème dont la solution est connue. La solution connue est donc adaptée au nouveau problème.

Un système de raisonnement à base de règles (RAPR) est choisi la plupart du temps par rapport au raisonnement à base de cas (RAPC) pour plusieurs raisons : tout d'abord, il est facile à comprendre car l'humain raisonne souvent sous forme de règles (si conditions alors actions). Puis, il permet la modularité sous forme de règles de connaissances. Un problème complexe peut être décomposé en règles simples. De plus, le raisonnement se fait par déduction et non par analogie. Le raisonnement par analogie peut aboutir parfois à des conclusions erronées (Lieber 2001).

[34] Domaine de recherche visant à rendre les machines intelligentes (raisonnement, perception, reconnaissance, etc.).

| | RAPC | RAPR |
|---|---|---|
| **Connaissance** | Cas | Génération de cas |
| **Modularité** | problème | règle |
| **Résolution des problèmes** | Adaptation de cas | Application de règles (rapide) |
| **Raisonnement** | Non déductif | déductif |
| **Acquisition** | Facile (épisode de résolution d'un problème) | Difficile (comment faire pour résoudre un problème) |

Table 1-1 : Comparaison entre RAPC et RAPR (Idiri et Napoli, 2012).

L'approche de raisonnement automatique par règles d'inférence présente un intérêt accru pour la problématique de surveillance maritime. Elle est utilisée pour l'identification des comportements à risques par l'implémentation du raisonnement humain. Cette approche est intéressante car elle est facile à mettre en œuvre si nous la comparons avec les modèles mathématiques par exemple. Un problème complexe peut être décomposé en un ensemble de règles. De plus, cette modularité sous forme de règles facilite la maintenabilité du système. En effet, les connaissances sont mises à jour facilement et rapidement dans un système de raisonnement alors qu'il est difficile par exemple de faire évoluer un modèle mathématique.

Pour bien comprendre les systèmes de raisonnement à base de règles, une présentation de son fonctionnement est exposée ci-après. Dans un système de raisonnement à base de règles, trois composantes essentielles sont définies : une base de connaissance ; les faits et un moteur d'inférence. Par analogie au raisonnement humain, la base de connaissance est l'ensemble de connaissances d'un être humain. Les faits sont la perception de son environnement (vue, goût, toucher, etc.). Le moteur d'inférence est le raisonnement humain. Comme nous le voyons sur la Figure 1-10, le moteur d'inférence vérifie en continu dans les événements en entrée (base de faits) s'il y a des règles applicables dans la base de règles. Avant d'exécuter ces règles, le moteur doit résoudre les conflits qui peuvent apparaître entre les règles applicables. Après l'exécution des règles sélectionnées par le moteur, de nouvelles règles et/ou de nouveaux faits vont venir enrichir respectivement, la base de règles et la base de faits (faits inférés, par exemple retourner une alerte).

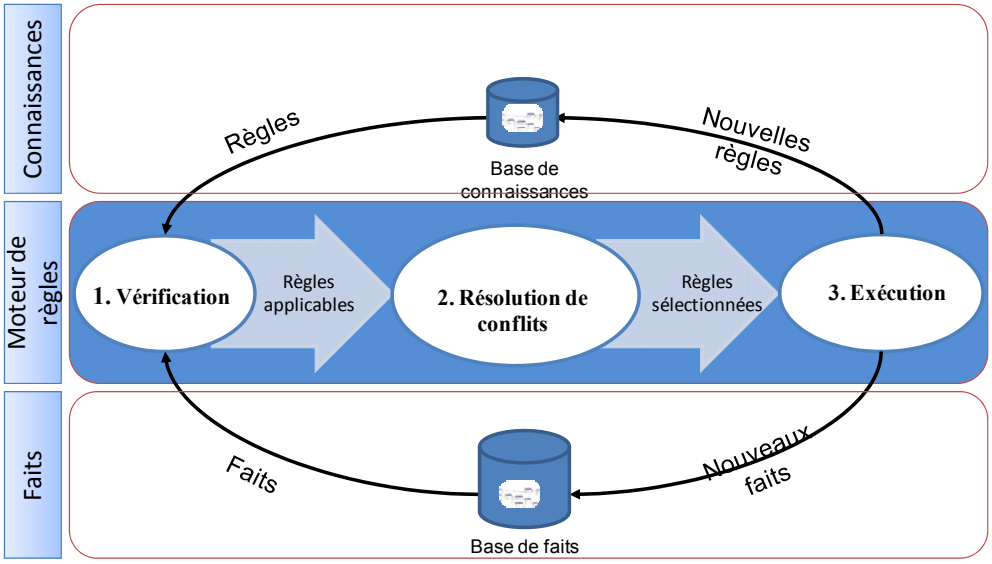


Figure 1-10 : Processus de raisonnement à base de règles
(Idiri et Napoli, 2012).

L'équipe de Jean Roy (Roy 2010) a utilisé le raisonnement à base de règles pour l'identification automatique de comportements anormaux liées à la sécurité maritime. Ils ont décrit un système complet de détection automatique de comportements anormaux, de la constitution de la base de connaissances à l'évaluation de menaces. Les connaissances expertes ont été définies sous forme de règles au cours d'un Workshop organisé au Canada avec des experts du domaine maritime. Ces règles décrivent des situations anormales de navires, connues auparavant par les experts qu'elles portent sur des informations statiques (signal AIS, numéro IMO, etc.) ou dynamique (vitesse, position, équipage, etc.). Quelques exemples de ces règles sont représentés sur la Figure 1-11. Une taxonomie des anomalies a été proposée sous forme d'ontologie (Roy, 2008) au cours de ce workshop. Studer (Studer et al., 1998) avait défini une ontologie comme étant « une spécification formelle et explicite d'une conceptualisation partagée». Pour plus de détail sur les ontologies, voir la section 1.6.2.

Après avoir conçu et développé le système de raisonnement à base de règles, il est possible d'identifier automatiquement les anomalies à partir des flux continus d'informations sur les navires. Concernant l'acquisition des connaissances expertes, J. Roy et son équipe se sont basés sur le brainstorming.

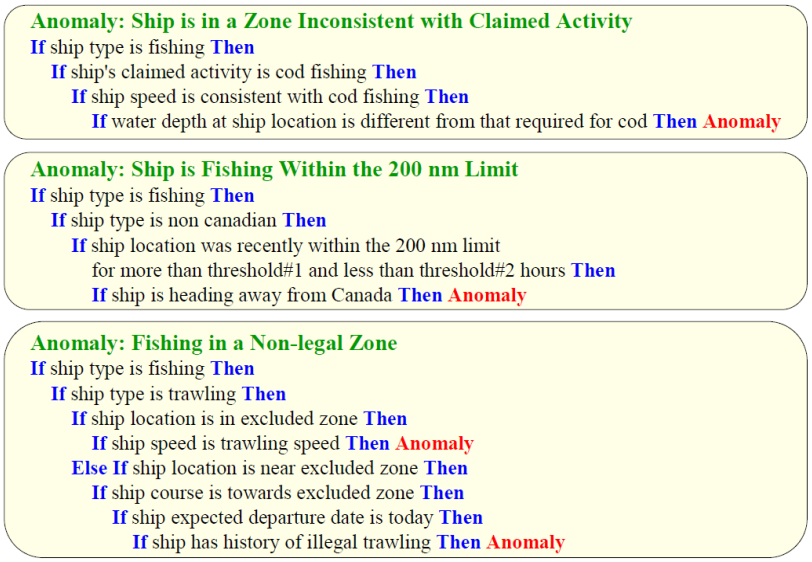


Figure 1-11 : Exemple de règles définies par brainstorming au Workshop organisé au Canada (J. Roy 2010).

Dans le projet SISMARIS, piloté par DCNS[35] et auquel a participé notre centre de recherche, un système de raisonnement à base de règles est aussi proposé pour analyser le trafic maritime sur des zones étendues. Les règles de connaissances intégrées au système sont issues des échanges avec les experts du domaine maritime comme les CROSS et la gendarmerie maritime.

Le point fort du raisonnement à base de règles est sa facilité de compréhension, sa modularité sous forme de règles (si condition alors action), sa maintenabilité et sa rapidité de traitement. Son point faible est lié à la difficulté d'acquisition et de capitalisation des connaissances (Estevez et al., 2006). L'acquisition des connaissances est considérée comme le goulot d'étranglement de cette méthode d'analyse de comportements.

## 1.6.2. Modélisation ontologique

L'analyse de comportements d'objets mobiles a une forte composante spatiale ce qui justifie l'intérêt des ontologies spatiales. Les ontologies spatiales présentent une potentialité intéressante pour l'analyse de comportements en intégrant le raisonnement à base de cas. Pour le raisonnement à base de cas, il a été défini dans la section 1.6.1. Une

[35] Groupe Français travaillant dans l'armement naval et l'énergie, http://fr.dcnsgroup.com/

ontologie est un ensemble de formalismes et de structures pour la formalisation et l'exploitation de connaissances d'un domaine. Les concepts et les termes partagés de ce domaine vont être organisés sous forme d'une hiérarchie de concepts qui est une structure de données en graphe décrivant des concepts partagés et leurs différentes relations. Les relations peuvent être d'inclusion comme le fait « chalutier » inclus dans le concept « navires de pêche » ou sémantiques comme la relation de voisinage ou de proximité.

Le but de cette modélisation ontologique est de formaliser le sens des termes et concepts d'un domaine pour pouvoir les exploiter par les personnes et les ordinateurs conjointement.

L'apport des ontologies pour l'analyse de comportements anormaux de navires a été évalué par Vandecasteele (Vandecasteele 2012) en développant un prototype appelé *OntoMap*. Ce prototype a permis de capitaliser des connaissances expertes et d'identifier automatiquement quelques comportements anormaux sur des données de déplacements de navires. Les informations maritimes utilisées dans ce prototype sont regroupées en trois ontologies : géométrique (position, trajectoires, etc.), cartographique et spécifique au domaine d'application (Vandecasteele 2012). L'ontologie métier, spécifique au domaine d'application a été constituée à l'aide de connaissances expertes issues de la littérature et d'interviews avec les experts (Vandecasteele 2012)[36].

Le système OntoMap (Figure 1-12) permet de définir des scénarios (cas résolus) en fonction d'un ensemble de propriétés et une distance sémantique à partir de laquelle le système décide de faire correspondre un ensemble de faits à un ou plusieurs scénarios. Prenons l'exemple de deux ou plusieurs navires qui naviguent à proximité dans une zone de pêche. Ces faits retournent des alertes associées aux navires concernés. Si la distance sémantique entre les faits et un scénario prédéfini est petite alors il y a de fortes chances que le scénario corresponde.

[36] Page 39

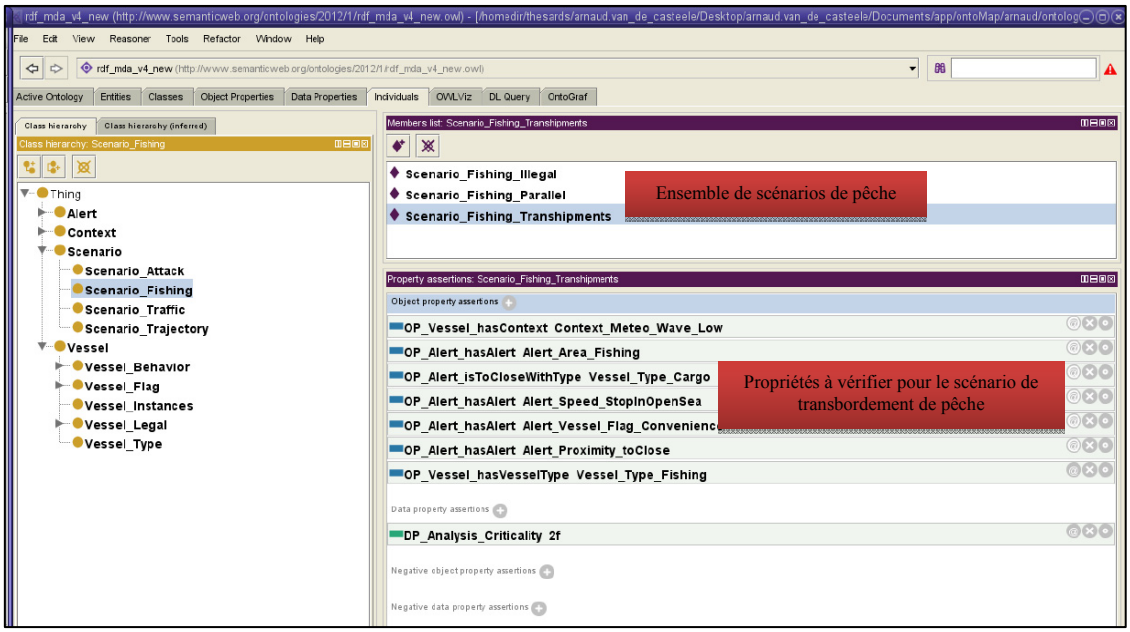


Figure 1-12 : Capture d'écran du système OntoMap (Vandecasteele 2012).

## 1.6.3. Modélisation par Classifieur Bayésien

La classification Bayésienne naïve est une méthode d'apprentissage supervisée qui met en œuvre des classifieurs pour la reconnaissance de formes, la prédiction et le tri. Cette méthode est basée sur le théorème de Bayes et permet de mettre en évidence des combinaisons linéaires entre les observations. L'appellation « naïve » est attribuée à cause de l'hypothèse de départ concernant l'indépendance des caractéristiques des objets classés. Prenons l'exemple d'une embarcation considérée comme étant de classe zodiac si la longueur et la vitesse mesurées appartiennent à des classes de valeurs prédéfinies. Ces deux descripteurs, à savoir la longueur et la vitesse sont supposés être indépendants. Malgré cette hypothèse simpliste de départ, ce type de classifieur est très utilisé et se révèle efficace et robuste.

En se basant sur cette méthode de classification, le projet Système d'Alerte et de Réponse Graduée Offshore (SARGOS), financé par l'agence ANR[37], propose comme son nom l'indique, une réponse à l'encontre d'actes de malveillance comme la piraterie, le terrorisme auxquels les infrastructures offshore sont vulnérables. Il supporte toute la chaîne de traitement, de l'identification de la menace à la proposition d'une réponse

[37] Agence Nationale de Recherche, structure gouvernementale Française qui finance les recherches publiques

graduée. A partir d'informations récupérées par des radars à ondes continues (FMCW[38]) installés sur des plateformes offshores, les petites embarcations vont être détectées et classifiées en se basant sur les classifieurs Bayésiens. Dans le cadre de l'expérimentation de cette méthodologie, l'apprentissage s'est fait sur des données d'observation collectées à partir d'un radar FMCW installé sur le site de la Direction Générale de l'Armement (DGA) de Saint-Mandrier (Giraud et al., 2013). L'apprentissage permet de construire un dictionnaire composé de matrices de vecteurs forme regroupant les caractéristiques de chaque classe de navire (TéSA 2011). Chaque vecteur forme décrit la forme, la géométrie et la topologie d'une observation cible. C'est l'utilisation en situation opérationnelle de ce dictionnaire qui va permettre la classification automatique de nouvelles observations. Dès qu'un objet pénètre dans le périmètre du radar, il est identifié et classé (Figure 1-13).

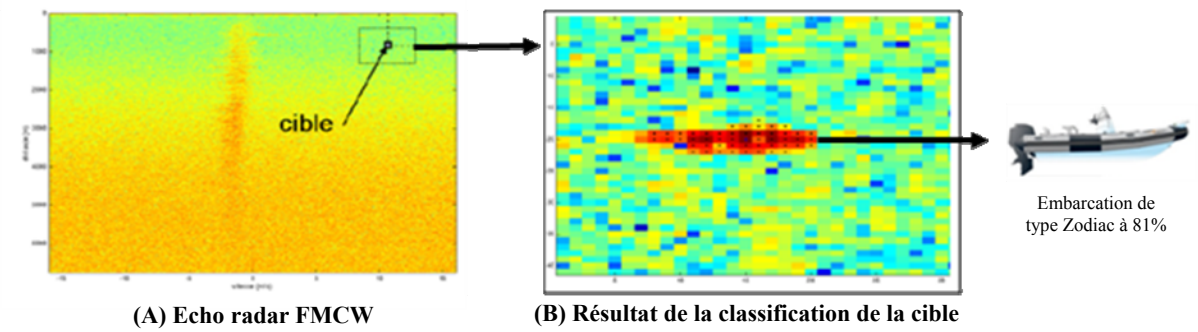


Figure 1-13 : Ciblage et classification d'un comportement d'intrusion d'une embarcation à partir d'images radar Range-Doppler d'un FMCW (TéSA 2011).

Pour chaque nouvelle observation détectée, un vecteur forme est mesuré puis comparé aux vecteurs formes se trouvant dans le dictionnaire pour prédire son type. Une distance sémantique est alors calculée entre ce vecteur forme et les vecteurs forme du dictionnaire pour trouver la classe d'appartenance la plus probable. Les descripteurs de l'objet choisis dans le vecteur forme doivent être discriminant entre les objets appartenant à des classes différentes et confondant entre les objets de même classe. Il peut cependant s'avérer difficile de classifier automatiquement des objets ayant des formes qui peuvent se confondre.

[38] Frequency Modulated Continuous Wave

### 1.6.4. Autres méthodologies

On trouve aussi dans la littérature, d'autres travaux de recherche sur l'analyse de situations à risques de navires par des approches probabilistes (Amrozowicz et al., 1997) et fondées sur la simulation numérique (Fournier 2005) (Nishizaki et al., 2011). Ces approches ne considèrent pas les historiques de données qui sont riches en enseignements, ce qui rend la découverte de nouvelles connaissances difficile.

## 1.7. Limites des méthodologies de modélisation actuelles

La plupart des méthodologies de modélisation des risques maritimes proposées actuellement sont basées sur la formalisation de connaissances expertes, qu'elles soient issues de revue de littérature, du brainstorming ou d'interviews avec les experts. Les connaissances issues de ces méthodes dépendent de l'expérience des experts et sont donc subjectives du fait qu'elles peuvent changer d'un expert à l'autre. De plus, ces méthodes ne permettent pas la découverte de connaissances nouvelles décrivant des comportements auxquels les experts n'ont jamais été confrontés.

Selon N. Sumpter (Sumpter & J. Bulpitt 2000), la modélisation de comportements d'objets par apprentissage sur des observations peut produire de meilleurs modèles de comportement, plus réaliste que la modélisation à base d'expertise. L'apprentissage automatique ou la fouille de données permet de bien capturer les caractéristiques réelles à partir des observations et la découverte de nouvelles connaissances. Elle utilise des outils issus de plusieurs domaines comme l'intelligence artificielle, la théorie de l'apprentissage, la théorie de l'information et les statistiques qu'elles soient inférentielles ou descriptives. Elle intègre donc des outils élaborés permettant la découverte de connaissances complexes, ne pouvant être découvertes en tâtonnant par exemple sur des résultats de statistiques descriptives (Tufféry 2010).

## 1.8. Synthèse sur les méthodes d'analyse et de modélisation de comportements de navires

Dans les sections précédentes, des méthodologies d'analyse et de modélisation de comportements de navires ont été décrites. Dans l'objectif de synthétiser ces méthodologies, deux tableaux récapitulatifs sont présentés ci-dessous.

| | **Description** | **Avantages** | **Limites** |
|---|---|---|---|
| **Analyse statistique** | Intègre deux approches (exploratoire et confirmatoire) pour la description synthétique des variables, la recherche de dépendances, la mesure de déviations, la découverte de relations de causalité, etc. | - Etudier les distributions des variables univariées et multivariées,<br>- Infirmer ou confirmer une hypothèse,<br>- Inférence des conclusions à la population entière avec une marge d'erreur calculée,<br>- Utiliser des représentations graphiques intuitives (Box Plot, histogramme, etc.). | - Nécessite de poser des hypothèses de départ,<br>- Ne permet d'analyser qu'une dizaine de milliers d'individus et quelques dizaines de variables. |
| **Analyse visuelle** | Utilise plusieurs formes de visualisation de données pour permettre l'analyse et l'extraction visuelle de connaissances. | - Améliorer les capacités de perception, compréhension et raisonnement des utilisateurs,<br>- Intégrer la dimension humaine dans l'exploration de données pour faire participer les utilisateurs dans la construction des connaissances. | - Raisonnement humain sur les différentes formes visuelles requiert un temps prohibitif pour la construction de connaissances,<br>- Complexité de certaines visualisations et leur inadaptation par rapport aux données analysées, le type d'analyse et le profil de l'utilisateur rend difficile l'interprétation des résultats. |
| **Fouille de données** | Combine plusieurs techniques statistiques, d'intelligence artificielle, d'algorithmiques, d'informatiques, etc. pour l'extraction automatique de connaissances à partir de bases de données. | - Générer automatiquement des hypothèses (meilleure productivité),<br>- Découvrir de nouvelles connaissances,<br>- Très grandes capacités d'analyse en termes de nombre d'individus et de variables,<br>- Découvrir des modèles plus réalistes que les modèles à base d'expertise, | - Ne fait pas participer les utilisateurs dans l'exploration des données,<br>- Les résultats demandent de la compétence en fouille de données pour les interpréter et les valider. |

Table 1-2 : Synthèse des méthodes d'analyse utilisées dans l'étude de comportements de navires

| | Description | Avantages | Limites |
|---|---|---|---|
| **Règles d'inférence** | Formalisation de connaissances sous forme de règles de conjonction de disjonction d'occurrences : « ***si conditions alors actions »*** | - Facilité de compréhension par les utilisateurs,<br>- Modularité des règles,<br>- Réalisation de raisonnements déductifs,<br>- Facilité de mise en œuvre du système modélisé et sa maintenabilité dans le temps. | - Nombre important de règles de connaissances à gérer,<br>- Difficulté d'acquisition de la base de connaissances (comment faire pour résoudre un problème ?). |
| **Modélisation Ontologique** | Formalisation de connaissances sous forme de structures en hiérarchies de concepts et de formalismes pour les partager et les exploiter par les humains et les machines conjointement. | - Partage de connaissances entre les utilisateurs ; entre les utilisateurs et les systèmes et entre les systèmes,<br>- Réalisation de raisonnement automatique sur les ontologies,<br>- Facile à appréhender par les utilisateurs à cause de sa représentation en réseau de graphes. | - Dépendance forte avec le problème à résoudre,<br>- Insuffisance des formalismes exploitant la dimension spatiale. |
| **Classifieurs Bayésiens** | Modèles mettant en exergue des combinaisons linéaires entre les variables pour la reconnaissance de formes, la prédiction et le tri par exemple. | - Efficacité des classifieurs Bayésiens,<br>- Graduation des résultats (modèle probabiliste). | - Condition de départ sur l'indépendance des variables caractérisant les objets classés,<br>- Difficulté de classer automatiquement des objets qui se confondent. |

Table 1-3 : Synthèse des méthodes de modélisation utilisées pour la formalisation de comportements de navires

## 1.9. Conclusion

L'analyse de mouvements d'objets mobiles, leur compréhension, la compréhension des interactions entre les objets et entre les objets et l'environnement présentent des perspectives intéressantes pour des domaines d'application très variés comme la surveillance du trafic maritime.

Identifier des comportements inhabituels, à risques, appréhender les objectifs et les contraintes d'évolution à partir des mouvements est une avancée importante pour la surveillance automatique.

La plupart des travaux de modélisation pour l'identification automatisée des comportements inhabituels, anormaux ou suspects de navires proposés dans la littérature sont certes intéressants mais présentent des limites. Dans ces travaux, la méthode de brainstorming est souvent utilisée. Les scénarios produits dépendent alors beaucoup de l'expérience personnelle de chaque expert. De plus, elle ne permet pas la découverte de connaissances nouvelles. Dans la littérature, la modélisation des risques maritimes par fouille de données automatique est peu explorée (Darpa 2005) alors qu'elle peut combler les limites actuelles.

La fouille de données permet d'extraire des caractéristiques de comportements cachées et réelles à partir d'observations passées ce qui peut aboutir à de meilleurs modèles de comportements.

Dans le chapitre suivant, nous présentons un état de l'art des domaines de fouille de données et nous identifions les méthodes permettant d'extraire des connaissances sur les comportements à risques de navires.