Chapitre 4 : Exemples d’extraction de connaissances sur les comportements de navires potentiellement à risques

4.1. Introduction

Nous allons présenter dans ce chapitre quelques exemples d'extraction de connaissances pour l'analyse de comportements de navires potentiellement à risques dans l'objectif de tester et valider notre méthodologie. L'interprétation des connaissances générées à l'aide de ShipMine sous forme de mouvements et de situations permet de qualifier un comportement comme étant à risque.

Nous avons choisi d'extraire des exemples de comportements potentiellement à risques par type de comportements présentés sur la Figure 3-1 (Cf. section 3.2.2 du chapitre 3) :

- Des associations entre facteurs d'accidents pour le type « Facteurs de risques »,
- Des zones accidentogènes pour « Zones à risques »,
- Des trajectoires de dérives pour « Trajectoires et sous-trajectoires à risques »,
- Des comportements d'abordage et de pêche parallèle pour « Navigations proches »,
- Des routes à risques de naufrage pour « Routes de navigation à risques ».

Pour réaliser ces extractions, nous avons fait l'acquisition de bases de données sur des accidents maritimes et sur les pistes AIS de déplacements de navires qui sont présentées ci-dessous. La préparation de ces données, à savoir leur nettoyage et la restructuration de l'espace de données à explorer par les méthodes de fouille de données, a été aussi exposée. Le nettoyage de ces données permet d'enlever les bruits, les incohérences et traiter les valeurs manquantes.

Ce chapitre est organisé en trois parties : la première partie concerne la préparation de l'espace de données à explorer ; la deuxième s'intéresse à l'extraction de comportements pouvant décrire des comportements à risques et la troisième traite des limites et améliorations de l'approche de validation et des méthodes de fouille de données utilisées. La première partie de ce chapitre est subdivisée en trois sous-parties, à savoir l'acquisition des bases de données, leur nettoyage et la modélisation des espaces de données à explorer. La deuxième partie quant à elle, est subdivisée en deux sous-parties

pour distinguer l'extraction de situation à risques de l'extraction de mouvements à risques. Enfin, la troisième partie est subdivisée en deux sous-parties pour spécifier les limites et améliorations de la méthodologie et du prototype.

4.2. Préparation de l'espace de données à explorer

4.2.1. Acquisition de bases de données

Une phase d'acquisition des données à explorer est nécessaire dans toute approche de fouille de données. Dans le cadre de l'expérimentation de notre méthodologie, nous avons fait l'acquisition de trois bases de données : une base de données d'accidents maritimes, une base de données météorologiques et une base de données de localisation de navires. Nous présentons ci-dessous ces trois bases de données.

4.2.1.1. Données de Marine Accident Investigation Branch (MAIB)

Nous avons contacté la *Marine Accident Investigation Branch*[79] (MAIB) qui a mis à notre disposition une base de données *Microsoft Office Access* recensant les accidents/incidents de navires qui se sont produits entre 1991 et 2009. Cette base de données est une extraction à partir d'une plus large base de données tenue à jour par le MAIB qui est l'équivalant du bureau sur les événements en mer[80] Français (BEAmer). Ce dernier tient à jour des dossiers papiers décrivant les accidents qui demandent une intégration dans une base de données avant de pouvoir les exploiter.

Les données du MAIB recensent les accidents impliquant les navires britanniques se trouvant n'importe où dans le monde et tous les accidents se trouvant dans les eaux territoriales britanniques. La base de données d'une taille de 16.7 Mo, contient 14 900 cas d'accidents et d'incidents qui concernent 16 230 navires. Dans notre étude nous nous sommes limités aux eaux territoriales britanniques. Le modèle conceptuel de la base de données est représenté sur la Figure 0-1.

[79] http://www.maib.gov.uk/home/index.cfm
[80] http://www.beamer-france.org/index.php

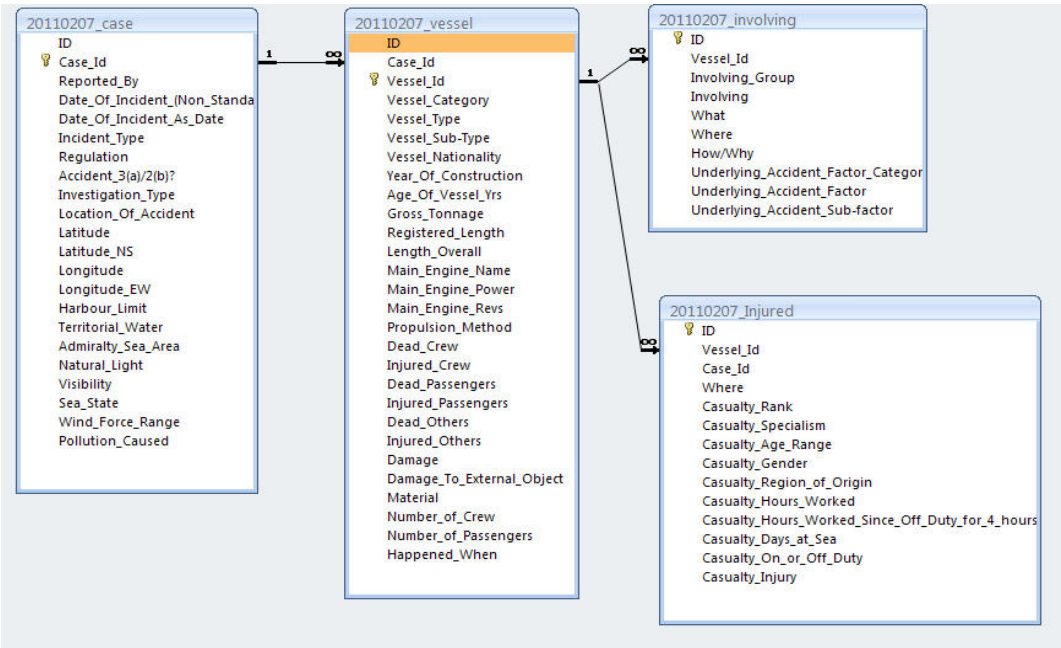

Figure 0-1 : Modèle conceptuel de la base de données MAIB.

4.2.1.2. Données de Modern-Era Retrospective analysis for Research and Applications (MERRA)

Nous avons téléchargé des données météorologiques à partir du site de la National Aeronautics and Space Administration[81]. Cette agence gouvernementale plus connue sous son acronyme NASA, a en charge la plupart des programmes spatiaux civils des États-Unis.

Un historique de données entre 1991 et 2009 a été extrait d'une table appelée IAU 2d atmospheric single-level diagnostics (tavg1_2d_slv_Nx) pour compléter les données météorologiques manquantes de la base de données MAIB. Ces données présentent des mesures de force, direction du vent, pression, humidité, etc. prises toutes les heures du 01/01/1990 au 31/12/2010. La Figure 4-2 illustre l'emprise spatiale des données téléchargées, c'est-à-dire le Royaume-Uni.

Figure 0-2 : Etendue de la zone géographique de téléchargement des données MERRA.

[81] http://disc.sci.gsfc.nasa.gov/daac-bin/FTPSubset.pl?LOOKUPID_List=MAT1NXSLV.

Les données téléchargées concernent les variables suivantes (voir Table 4-1) :

Variable	Description	Unité / Type variable
Lon_Merra	Longitude du point Merra	Degrés décimaux
Lat_Merra	Latitude du point Merra	Degrés décimaux
Merra_time	Timestamp du relevé météo Merra	aaaa-mm-jj hh:mm:ss
SLP	Pression au niveau de la mer	Pa
DISPH	Hauteur de déplacement	m
U2M	Composante Ouest->Est du vent à 2 m au-dessus de la hauteur de déplacement	m.s-1
V2M	Composante Sud->Nord du vent à 2 m au-dessus de la hauteur de déplacement	m.s-1
T2M	Température à 2 m au-dessus de la hauteur de déplacement	Kelvin
QV2M	Humidité spécifique à 2 m au-dessus de la hauteur de déplacement	Kg.kg-1
Winddspeed	Vitesse du vent	m.s-1
Winddir	Direction du vent	Degrés

Table 0-1 : Description des données MERRA téléchargées.

4.2.1.3. Données AIS (Automatic Identification System)

Ce sont des données de cinématiques de navires (Numéro MMSI, vitesse, heure UTC, position, cap, vitesse, etc.) transmises en quasi-temps-réel par les capteurs AIS installés à bord de navires. Ces données transmises sont reçues par d'autres navires navigant à côté et des récepteurs implantés sur les côtes.

Les données AIS anonymes[82], nous sont fournies par DCNS[83] à titre gracieux sous forme de trames d'informations de la National Marine Electronics Association (*NMEA)* envoyées au fur et à mesure de leur réception. Ces données alimente presque en continue notre serveur de données PostgreSQL depuis un an avec un volume de messages d'environ 1.5 Go par jour.

[82] Vu le caractère sensible de l'utilisation de ces données, il a été procédé à leur anonymisation en modifiant les numéros MMSI des navires.
[83] Groupe Français intervenant dans le domaine de l'armement naval et de l'énergie.

4.2.2. Nettoyage des données

De nos jours, d'importantes bases de données sont constituées à partir de flux continus d'informations ou de mises à jour régulières. Ces données peuvent contenir des bruits, des données manquantes, des incohérences ou des imprécisions. En fouille de données, contrairement à l'analyse statistique, les données utilisées sont souvent créées à d'autres fins, elles sont donc souvent inexploitables directement par la fouille de données.

La qualité des connaissances extraites par fouille de données dépend beaucoup de la qualité et de la quantité des données en entrée. En effet, plus il y a de données (cas observés) meilleure est la précision des connaissances. Une analyse statistique des valeurs d'attributs de ces données est nécessaire étant donné que la qualité des résultats d'analyses exploratoires dépend généralement plus de la préparation des données que de la méthode exploratoire utilisée. La première étape de toute investigation dans les données est le calcul des statistiques univariées[84] pour connaître la distribution des variables et identifier les anomalies. Il est important de manipuler les incohérences et les valeurs manquantes avec intelligence car elles peuvent véhiculer des informations intéressantes. En effet, l'absence d'un signal AIS transmis par un chalutier dont la dernière position connue se situe à proximité d'une zone de pêche illégale peut être un indicateur d'une tentative de fraude. Une trajectoire d'un navire aberrante par rapport aux trajectoires du groupe (navires de même type) ou par rapport à ses trajectoires habituelles peut indiquer un comportement suspect.

Avant toute exploration de données, une étape de préparation de ces données est nécessaire pour permettre leur exploitation. Cette préparation est difficile et demande plusieurs itérations compte-tenu de son lien fort avec la qualité des résultats. En effet, la quantité et la qualité des données ont un impact direct et significatif sur la qualité des connaissances générées. Nous nous proposons d'étudier dans les sous-sections suivantes, la distribution des variables pour identifier les anomalies, les corriger et préparer le contexte d'exploration. Les anomalies peuvent être des données manquantes, des incohérences et des imprécisions.

[84] Analyse statistiques concernant une seule variable

4.2.2.1. Données manquantes

Dans le but d'améliorer la qualité des résultats, plusieurs approches peuvent être envisagées pour nettoyer les données comme le remplacement de données manquantes par pondération de moyennes et de médianes ; la prédiction des valeurs manquantes (Jami et al., 2005) et la suppression des observations concernant les valeurs manquantes. Nous avons choisi dans notre cas de remplacer les valeurs manquantes par des valeurs issues d'autres sources de données. Dans le cas de la variable décrivant la force du vent des données d'enquêtes accidents, nous avons remplacé les valeurs manquantes par des valeurs de mesures issues de la base de données MERRA. Comme on le voit sur la Figure 0-3, les coordonnées des données MERRA sont représentées par la grille en bleu et les positions des accidents en orangé. Une requête relie chaque cas de la base de données MAIB avec le point de données MERRA se trouvant à proximité en tenant compte de l'heure la plus proche avec une confiance de 30 minutes. La requête a permis de générer un fichier CSV contenant les résultats de la jointure interne entre les incidents MAIB et les données MERRA.

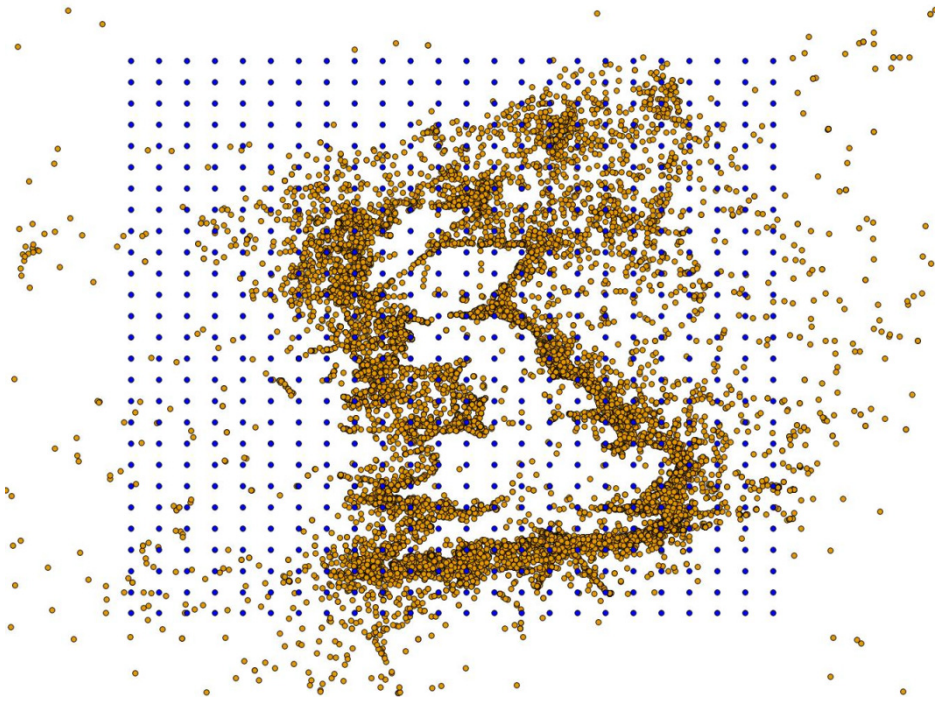

Figure 0-3 : Compléter les données manquantes du MAIB par superposition des données MERRA.

Dans MERRA, la plus petite résolution pour l'intervalle de temps et la région choisis est « 1/2° Latitude x 2/3° Longitude ». Même si cette résolution spatiale semble grande, nous supposons que les mesures météorologiques sont bien interpolées. De plus, sur les 12 000 observations de MAIB bien renseignées, presque 8 000 correspondent à celles de MERRA.

4.2.2.2. Variables continues et distribution hétérogène

Les algorithmes d'extraction de règles d'association ne prennent pas en considération les variables ou les attributs continus dans leur processus d'extraction. Pour ne pas perdre d'informations en entrée de ces algorithmes, une distribution des variables continues a été effectuée pour les séparer en classes d'intervalles. Nous avons choisi les effectifs égaux comme critère de séparation des classes pour éviter de biaiser les résultats de ces algorithmes. En effet, ces algorithmes sont basés sur la découverte d'itemsets fréquents par un calcul d'effectifs, donc une classe ayant plus d'effectifs a plus de chance d'apparaitre dans les règles d'association en sortie.

Nous avons identifié aussi dans la base de données MAIB, des variables discrètes ayant une distribution hétérogène de leur effectif. Cette hétérogénéité va empêcher l'apparition des fameuses pépites d'or ou les cas rares dans les règles d'association en sortie. Les valeurs des variables de la base de données ayant les plus grandes fréquences d'apparition, vont apparaître plus souvent dans les règles au détriment des autres. En prenant l'exemple des navires de pêche dans notre base de données d'enquêtes d'accidents et d'incidents de navires britanniques, nous avions remarqué qu'ils apparaissaient presque systématiquement dans les règles d'association. Ce qui est normal car ils représentent *66%* de l'effectif total de la base de données. Pour faire apparaître les autres catégories de navires, nous avons dû regrouper toutes les catégories de navires en deux grandes classes :

- Classe Transport : Avec un effectif de *34%*, elle regroupe tous les navires de transports de personnes, d'hydrocarbures (Tanker) et de marchandises,

- Classe Pêche : Les navires de pêche représentent un effectif de *66%*.

D'autres algorithmes de fouille de données ne peuvent pas être appliqués sur des variables continues comme le cas de l'algorithme ID3, c'est pour cela que des discrétisations de variables continues sont effectuées.

4.2.2.3. Données aberrantes

Les données aberrantes sont des données atypiques qui sont éloignées de l'ensemble des observations. Elles peuvent être des observations intéressantes pour un domaine et erronées pour un autre. La présence d'une donnée aberrante peut signifier une erreur de saisie, de mesure ou un cas particulier, appelé aussi atypique. L'identification et la décision de garder ou non ces données demande d'avoir une bonne connaissance du domaine d'application étudié.

4.2.2.3.1. Données accidents

La répartition des accidents de la base de données MAIB sur une carte numérique nous a permis par exemple d'identifier et d'écarter les positions aberrantes localisées sur terre, loin des zones de navigation (Figure 0-4). Ces erreurs de positionnement sont peut être dues aux dysfonctionnements de GPS ou à une mauvaise saisie des coordonnées.

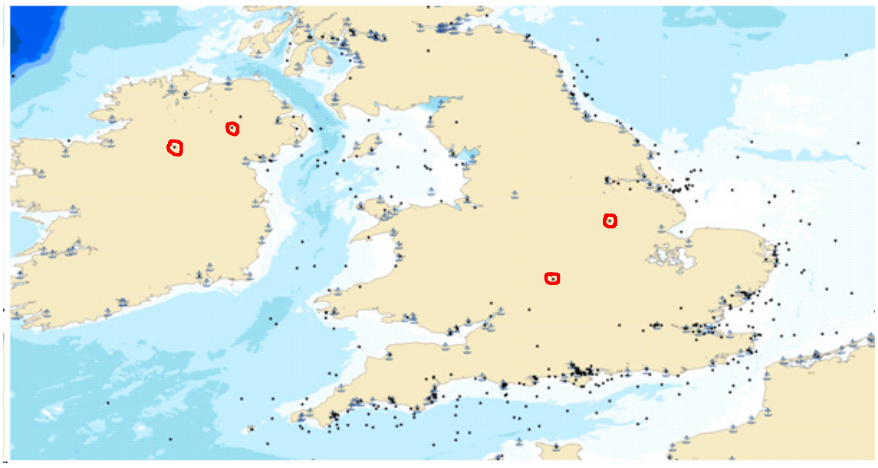

Figure 0-4 : Identification de positions aberrantes par une répartition des positions d'accidents et d'incidents de navires sur une carte numérique. Quelques positions aberrantes sont entourées en rouge.

4.2.2.3.2. Données AIS

Un autre type de valeurs aberrantes a été identifié dans la variable *Age-of-vessel* qui comporte des valeurs négatives. L'analyse des valeurs extrêmes qui représentent les valeurs maximales et minimales de la distribution de la variable, nous a permis de détecter ces valeurs erronées pour les éliminer de la base.

Nous avons suivi la même approche que la répartition des accidents MAIB sur une carte numérique pour détecter les trajectoires aberrantes. La reconstitution des données AIS sous forme de trajectoires sur une cartographie, nous a permis de découvrir des trajectoires aberrantes qui passent sur les terres.

La Figure 4-5 montre l'exemple d'une trajectoire aberrante passant entre le port de Malaga et Valence. Cela est dû la plupart du temps à une perte de signal AIS entre deux points éloignés.

Figure 0-5 : Une trajectoire aberrante due à une perte de signal AIS.

Pour nettoyer ces données, nous avons choisi une approche simple. On affiche sur une cartographie toutes les trajectoires d'un jeu de données puis les trajectoires passant sur terre sont sélectionnées pour récupérer leur identifiant MMSI. Les trajectoires ayant des partitions aberrantes sont alors supprimées à partir du fichier de données source. Cette procédure est simple et rapide pour un nombre raisonnable de trajectoires. Pour les trajectoires passant sur les terres à proximité de la mer, une interpolation est effectuée pour compléter ces trajectoires. Après suppression des trajectoires aberrantes, interpolation des trajectoires qui passent près de la mer dans certaines partitions et découpages des trajectoires aux ports, nous avons conservé 65 trajectoires de tankers.

Il y a d'autres données aberrantes dues à la perte de signaux. Ces trajectoires ne passent pas par les terres mais présentent des partitions très longues. L'algorithme de détection de trajectoires aberrantes va ainsi détecter ces partitions alors qu'elles ne nous intéressent pas. Ces aberrations sont dues à une perte de signale et non aux déplacements réels des navires. Nous proposons pour éliminer ces aberrations, une limitation de la

longueur des partitions. Des conditions sont fixées dans le calcul des partitions aberrantes pour filtrer les parties jugées trop longues ou trop courtes. Pour éviter des analyses statistiques couteuses, nous avons fixé des seuils arbitrairement : si la partition est plus petite que 0.00009° - environ 10 mètres - ou plus grande que 0.09° - 10 kilomètres environ - elle n'est pas prise en compte dans le calcul.

4.2.3. Modélisation des espaces de données

Nous présentons ci-après la modélisation des espaces de données concernant les données accidents et les données AIS.

4.2.3.1. Données Accidents

Nous avons sélectionné dans la base de données MAIB, les données qui décrivent les accidents (type d'accidents, position, temps, eaux territoriales, etc.), les caractéristiques des navires (identifiant IMO, type du navire, âge du navire, longueur, etc.) et la description de l'environnement (visibilité, état de la mer, force du vent, etc.). La sélection des variables sur lesquelles va porter notre analyse va réduire le nombre de variables à considérer, le nombre de règles générées et ainsi faciliter l'interprétation des résultats. Cette sélection de données va constituer le contexte d'exploration sur lequel va porter l'extraction de règles d'association dans le but de trouver les relations d'implications entre les différents facteurs de situations.

Quelques variables prises toutes seules permettent déjà d'évaluer le risque comme le type du navire. En effet, les autorités maritimes s'appuient beaucoup sur cette dimension pour mesurer le risque. L'entrée d'un navire militaire et d'un cargo dans la rade de Brest par exemple, n'est pas perçue au même niveau de risque par les autorités maritimes. On imagine alors l'apport d'une mise en relation entre plusieurs variables à l'évaluation des risques.

Dans l'analyse des accidents de navires, nous avons gardé les types d'accidents les plus fréquents et pouvant avoir une relation avec les conditions météorologiques, à savoir, l'échouement et le naufrage.

En ce qui concerne les variables, nous avons choisi de garder les conditions de navigation (vent, courant, etc.), les caractéristiques de navires (type, âge, etc.) et les risques maritimes (type du risque, catégorie, etc.) pour la détection de facteurs de risques et la localisation (latitude et longitude) de l'accident ainsi que le type d'accident pour la détection des zones de risques.

Nous avons aussi enlevé les données fluviales de la base de données MAIB car nous travaillons sur une problématique maritime.

Les données résultant de la sélection effectuée sur la base de données MAIB contiennent les attributs suivants (Table 4-2) :

Attribut	Description	Unité / Type variable
Case_id	Identifiant de l'incident	Entier
Incident_type	Type de l'incident	Texte
Vessel_id	Identifiant du navire impliqué	Entier
Vessel_Category	Catégorie de navire impliqué	Texte
Age_Slice_Of_Vessel	Intervalle d'âge du navire impliqué	Text
Incident_time	Timestamp de l'incident	aaaa-mm-jj hh:mm:ss
Location	Localisation de l'accident (Coastal waters, High seas, Non-tidal waters, Port/harbour area)	Texte
Territorial_Water	Indique l'eau territoriale de l'accident	Texte
Lat_vf	Latitude de l'incident	Degrés décimaux
Lon_vf	Longitude de l'incident	Degrés décimaux
Sea_state	Etat de la mer	Texte
Wind_force	Force du vent	Echelle Beaufort
Visibility	Visibilité au moment de l'accident (Poor, Good, Mod.)	Texte

Table 0-2 : Description des attributs de la base de données MAIB sélectionnés et préparés.

4.2.3.2. **Données AIS**

Concernant les données de déplacement, nous avons dû créer plusieurs fichiers de données selon l'algorithme d'exploration car chacun possède un format de données d'entrée particulier. Les données ont été aussi limitées par type de navires car les

comportements et plus particulièrement les mouvements sont différents d'un type de navire à un autre. Les tankers par exemple ont tendance à faire des trajets presque rectilignes entre deux ports (chemin le plus court) alors que les navires de pêche font des déplacements souvent spécifiques autour d'un port.

La préparation des données est certes une condition nécessaire mais elle n'est pas pour autant suffisante pour avoir des connaissances de qualités. En effet, d'autres conditions doivent être satisfaites comme par exemple le choix d'une méthode d'exploration pertinente.

Nous présentons ci-après, un tableau présentant quelques attributs de la base de données AIS (Table 4-3) :

Attributs	Description
MMSI	Identifiant unique du navire
IMO	Identifiant de l'équipement AIS
Nom	Nom du navire
Type	Type du navire
Latitude	Latitude du navire au moment de l'envoi du signal AIS
Longitude	Longitude du navire au moment de l'envoi du signal AIS
Vitesse	Vitesse du navire au moment de l'envoi du signal AIS
Cap	Direction du navire au moment de l'envoi du signal AIS
Horodatage	Timestamp de l'envoi du signal AIS
Cargaison	Type de cargaison transportée par le navire

Table 0-3 : Description de quelques attributs des données AIS.

4.3. Extraction de comportements à risques

L'extraction de comportements à risques passe préalablement par l'extraction et l'interprétation des situations et mouvements de navires. Dans les sous-sections suivantes, nous allons commencer par l'extraction de quelques exemples de situations à risques et mouvements à risques à partir de données réelles d'enquêtes d'accidents maritimes et de déplacements de navires.

4.3.1. Extraction de situations à risques

Parmi les méthodes d'extraction de situations à risques que nous avons choisies dans le chapitre précédent (Cf. section 3.2.3.1) pour extraire des connaissances sur les facteurs et les zones à risques, certaines ne sont pas intégrées à ShipMine. C'est le cas des méthodes d'extraction de règles d'association et la construction d'arbres de décision. Ces méthodes sont utilisées à partir de programmes tiers. La méthode de découverte de zones à risques quant à elle est intégrée à ShipMine.

L'exploration des données d'enquêtes d'accidents maritimes de MAIB, peut permettre la génération de connaissances sur des situations à risques mettant en relation des conditions de navigation (vent, courant, etc.), des caractéristiques de navires (type, âge, etc.) et les risques maritimes (type du risque, catégorie, etc.). La découverte de ces relations peut aider à la prédiction et au ciblage des accidents maritimes. Cette exploration peut aussi permettre la découverte de connaissances sur les zones à forte densité d'accidents pour l'identification de zones à risques.

La discussion des résultats découverts automatiquement par ces trois méthodes d'extraction de situations est présentée ci-après.

4.3.1.1. Extraction de règles d'association

La base de données de transactions utilisée est la base de données des accidents et incidents maritimes du MAIB (Cf. section 4.2.3). Les *items* sont les différents facteurs comme le type du navire, la force du vent, la localisation relative et les types d'accidents. La découverte des associations dans cette base consiste à chercher les ensembles *d'items* fréquemment liés dans les accidents (Idiri & Napoli 2012a).

L'exploration des données MAIB par l'algorithme *Apriori* après les avoir préparées (Cf. section 4.2.3), a permis de découvrir beaucoup de règles. Ces règles ont été obtenues en faisant varier les valeurs des seuils du support "*minsupp*" (indicateur de fiabilité de la règle) et de la confiance "*minconf*" (indicateur de précision de la règle). L'exploration des 4 247 observations a permis de générer :

- 631 règles avec des seuils de support et de confiance supp=0.04 et conf= 0.6,

- 371 règles avec supp=0.1 et conf= 0.6,

- 803 règles sur une sélection de 715 observations concernant les navires de transport avec supp=0.05 et conf= 0.5.

La sélection des règles intéressantes est effectuée en deux étapes. La première étape est une sélection selon les mesures de support, de confiance et de lift. La deuxième étape est effectuée sur un critère de pertinence par rapport à l'analyse des situations à risque. Dans la première étape, nous avons sélectionné plusieurs règles intéressantes en termes de métrique (support, confiance et lift) mais il s'est avéré dans la deuxième étape, après une analyse de pertinence que la majorité de ces règles n'apportaient pas de nouvelles connaissances.

Nous présentons ci-dessous, une règle d'associations par classe de résultats obtenus :

- **Règle 1** (Règle de prédiction) : Location = Coastal waters, Vessel_Category = Fish catching/processing, Age_Slice_Of_Vessel = 11 to 18 years → Incident_Type=Machinery Failure (supp=0.086 ; conf= 0.725 ; lift=1.47),

  La première règle informe que les accidents de navires de pêche âgés de 11 à 18 ans et navigant dans les eaux côtières britanniques sont causés dans 72% des cas par une panne mécanique. Cette règle est assez fréquente, elle représente presque 9% de la base de données MAIB.

- **Règle 2** (Règle de ciblage) : Vessel-Category=Fish catching → Vessel-Type= Trawler (supp= 0.14 ; conf= 0.43 ; Lift= 3),

  La deuxième règle informe que si un accident concerne un navire de pêche alors dans 43% des cas c'est un chalutier. La fiabilité de cette règle est vérifiée par 14% des cas de la base de données. Selon un sous-officier de la marine marchande, les chalutiers sont les plus exposés au risque de naufrage car ils tirent un chalut qui peut s'accrocher et entrainer vers le fond le chalutier. Donc cette règle confirme une information connue auparavant par les navigateurs.

- **Règle 3** (Règle Banale) : Vessel-Category=Passenger → Pollution-Caused=No (supp=0.15 ; conf= 0.73 ; lift= 1.2)

  La dernière règle, présente une règle triviale (inutile) qui signifie que les accidents de navires transportant des voyageurs ne causent pas de pollution dans 73% des cas, ce qui semble logique car ils transportent des passagers et non des substances polluantes. Il est important de noter que la notion de pollution ou non est ici relative aux autorités maritimes britanniques qui décident ou non si une pollution est avérée.

Concernant les règles impliquant des conditions météorologiques, océanographiques et de contexte, la majorité des règles trouvées, ayant le support (Supp >= « minsupp ») et la confiance (Conf >= « minconf ») mettent en relation des conditions normales : mer calme, force de vent faible, dans les ports, etc. voici ci-après, un exemple de quelques règles de ce type (Table 0-4).

Partie 1	Partie 2	Supp	Conf
Location=Port/harbour area, Wind_Force_Range=0-3 (calm)	Vessel_Category=Transport	5.7%	70%
Incident_Type=Grounding, Sea_State=Calm <2 ft	Vessel_Category=Fish catching/processing	4.6%	66%
Sea_State=Sheltered Waters, Vessel_Category=Transport	Location=Port/harbour area	8%	57%

Table 0-4 : Exemple de règles d'associations impliquant des conditions météorologiques, océanographiques et de contexte normales dans les accidents.

L'un des inconvénients de cette exploration est le fait que les *items* sont tous au même niveau, c'est pour cela que toutes les implications possibles ayant le seuil de support et de confiance sont générées. On se retrouve avec énormément de règles qu'il faut analyser pour identifier celles qui décrivent de nouvelles connaissances auxquelles les experts n'avaient pas pensés auparavant, des banalités (par exemple, il fait froid en hiver et chaud en été) et des idées reçues.

Par la suite, nous allons utiliser les arbres de décision pour définir une variable à expliquer à partir d'autres variables dites explicatives. Cela va focaliser la recherche sur

les règles permettant d'expliquer les types d'accidents par rapport aux facteurs (variables) d'environnement, de contexte et de caractéristiques des navires.

4.3.1.2. Construction d'un arbre de décision

Nous présentons ci-dessous le résultat obtenu de l'exploration des données MAIB par les arbres de décision. Comme on le voit sur la Table 4-5 la représentation graphique de l'ensemble des règles obtenues permet déjà une exploitation plus aisée et rapide des règles d'implication.

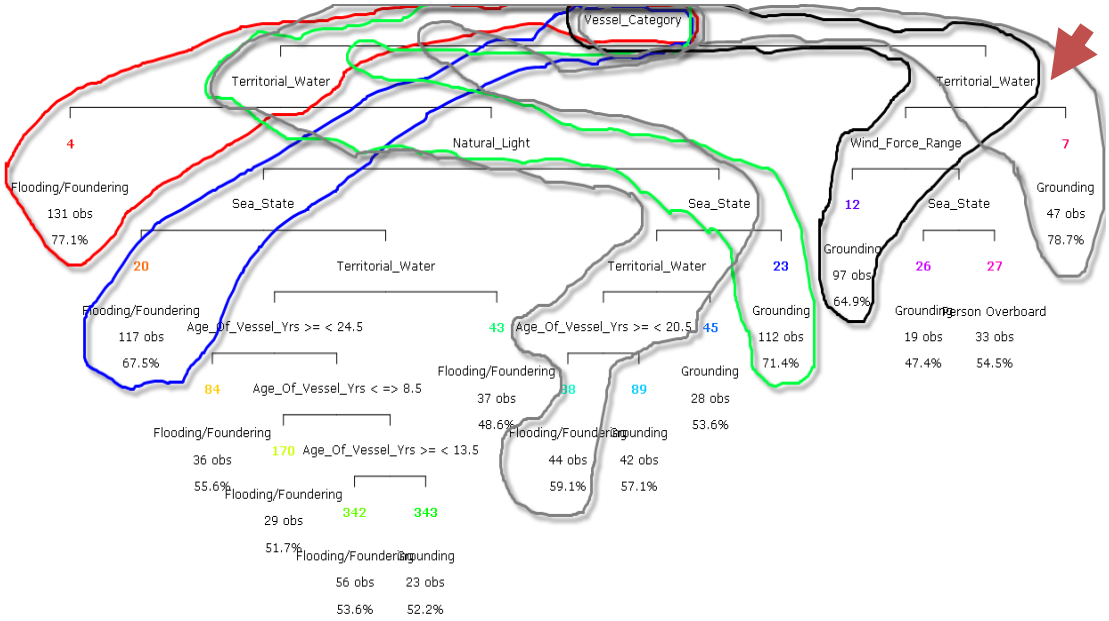

Figure 0-6 : Arbre de décision expliquant les types d'accidents par rapport à des facteurs météorologiques, océanographiques et des caractéristiques de navires.

Prenons par exemple, la règle 7 qui est entourée en gris sur l'arbre de décision (Figure 4-6). Cette règle identifie une relation entre les échouements de navires de transport britanniques et des localisations relatives : les accidents de navires de transport se trouvant sur les territoires : Northen Irish, Scottish ou Welsh sont des échouements dans 78,7% des cas.

L'utilisation des arbres de décision n'a pas été plus fructueuse que celle des règles d'association. La majorité des règles trouvées montre que les accidents ont eu lieu dans

des conditions météorologiques, océanographiques et de contextes les plus normales possible : mer calme, force de vent faible, bonne visibilité.

Pour confirmer que ces résultats ne sont pas dus à une sous-estimation des conditions météorologiques au cours de l'enquête accident et incident, nous les avons comparées aux données MERRA. Après vérification, il s'est avéré qu'il y a eu plutôt une surestimation des données MAIB, comme il est possible de le percevoir sur le tableau croisé des 4 334 observations différentes entre les données MAIB et MERRA (Table 4-5).

Table de wind_force par wind_force_Merra

		wind_force_Merra(wind_force_Merra)			
		0-3	4-6	7-9	Total
wind_force					
0-3	Pctage en ligne	0.00	99.73	0.27	
10-12	Pctage en ligne	8.61	60.93	30.46	
4-6	Pctage en ligne	99.60	0.00	0.40	
7-9	Pctage en ligne	7.07	92.93	0.00	
Other	Pctage en ligne	56.70	41.24	2.06	
Total	Fréquence	1915	2361	58	4334

Généré par le Système SAS ('Local', XP_PRO) le 07 février 2013 à 3:51:25 PM

Table 0-5 : Comparaison des forces du vent mal renseignées pendant l'enquête accident avec ceux de la base de données MERRA.

Le problème vient probablement de l'hypothèse de départ qui suppose une relation entre les mauvaises conditions météorologiques et les accidents maritimes. Il est possible que cette hypothèse ne soit pas si souvent vérifiée, c'est pour cela que les résultats ne sont pas ceux attendus.

Selon les résultats, il est possible que les accidents maritimes aient une relation avec la localisation. L'analyse des localisations d'accidents et incidents maritime va être abordée dans la section suivante.

4.3.1.3. Extraction de zones à risques

Des zones denses ont été découvertes par regroupement des localisations des accidents et incidents maritimes. Ces zones peuvent être utilisées pour suivre l'évolution des risques, suivre de plus près les navires qui fréquentent ces zones (Vessel Of Interest) et avoir une meilleur planification des moyens maritimes de surveillance et d'intervention.

L'utilisation de la fonctionnalité « Zones à risque » intégrée à ShipMine sur les données de localisation des accidents et incidents du MAIB, nous a permis d'identifier des zones accidentogènes. Pour une distance de voisinage égale à 1km et un seuil minimum d'accidents égal à 50, nous avons obtenu trois zones à risques localisées respectivement à côté des villes suivantes : Portsmounth, Milford Haven et Bournemouth. Comme nous le voyons sur la Figure 4-7, la zone autour de Portsmounth contient 178 accidents, Milford Haven 102 accidents et Bournemouth 61 accidents survenus entre 1991 et 2009.

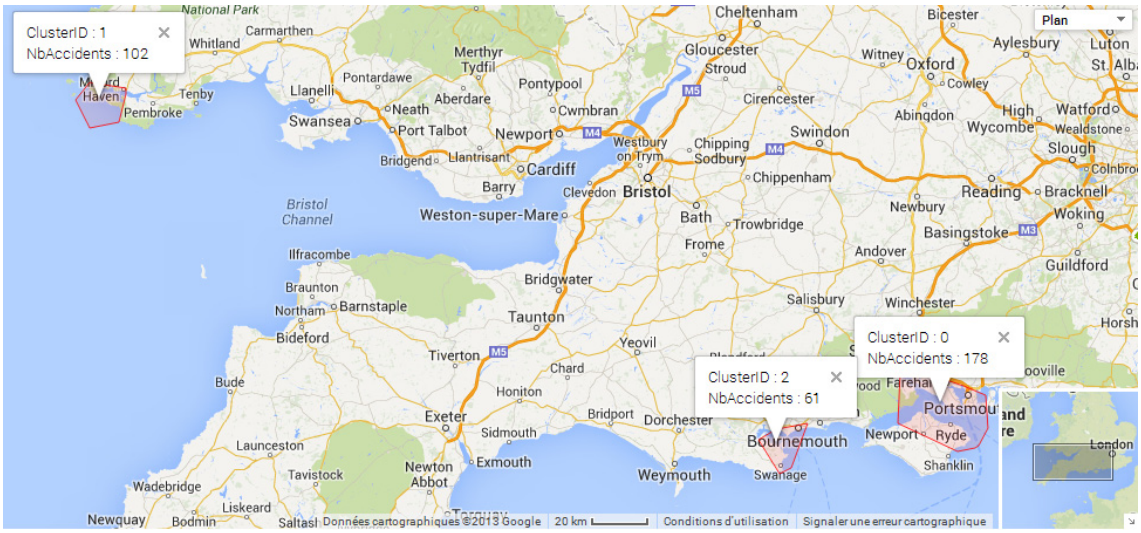

Figure 0-7 : Découverte de zones accidentogènes dans le sud de l'Angleterre pour une distance de voisinage de 10 km et un minimum de 50 accidents.

Plus nous augmentons la distance de voisinage, plus les zones accidentogènes deviennent larges et leur nombre augmente ou diminue selon la distribution des accidents. Comme on le voit sur la Figure 4-8-(a), pour une distance égale à 20 km nous avons obtenu 9 zones. La zone de Portsmounth vu dans le premier exemple est passée à 235 accidents et le nombre de zones a augmenté car avec une distance de 20 km, il y a plus de concentrations d'accidents identifiées.

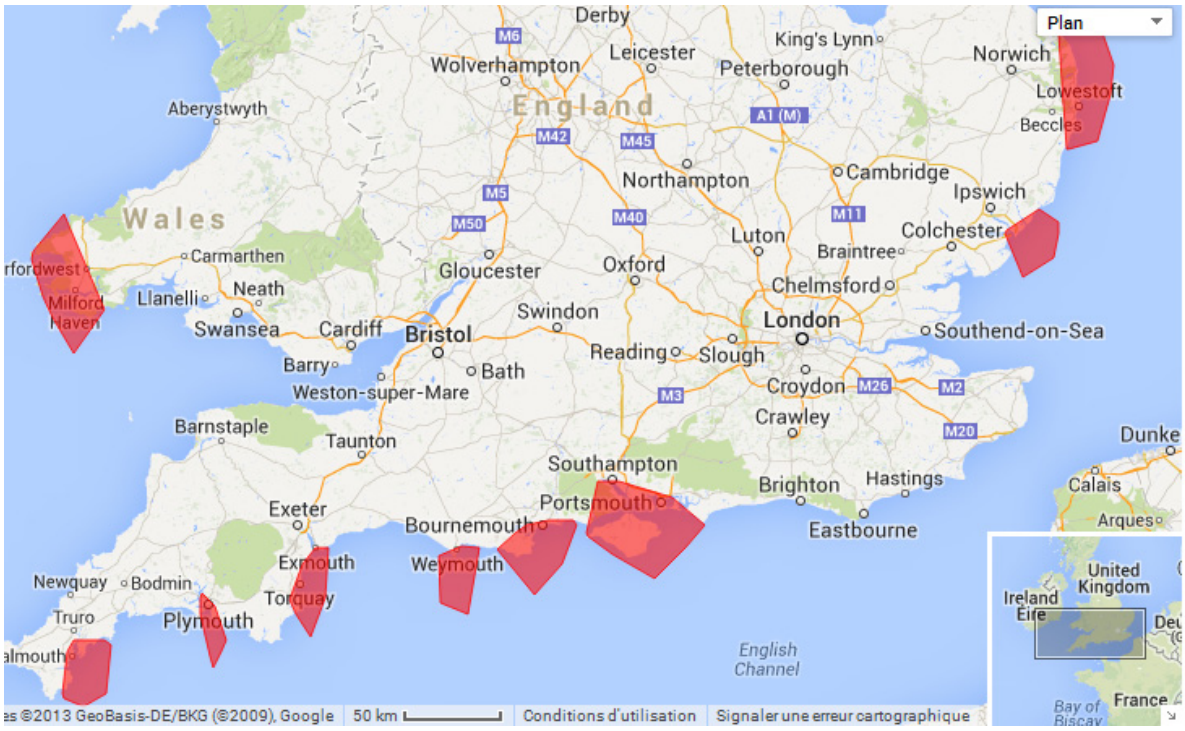

Figure 0-8 : Evolution de la largeur des zones accidentogènes par rapport à la distance de voisinage D=20 km.

Il peut s'avérer judicieux d'afficher les zones à risques sur des ENC pour identifier d'éventuelles relations entre des objets maritimes et ces concentrations d'accidents, comme par exemple des rochers affleurants.

4.3.2. Extraction de mouvements à risques

Dans cette section, nous allons procéder au test d'extraction de mouvements potentiellement à risques et plus particulièrement à la découverte de motifs de trajectoires aberrantes, de navigations proches et de motifs de routes de navigation à partir des historiques de déplacements de navires (position, vitesse, cap, etc.). La procédure de test utilisée est exposée ci-après.

4.3.2.1. Trajectoires aberrantes

L'exploration des 65 trajectoires de 16 tankers se déplaçant en Méditerranée en utilisant la fonctionnalité de ShipMine permettant de détecter les trajectoires anormales a permis d'extraire des mouvements anormaux de navires.

Pour une distance de voisinage égale à 100 mètres et une proportion de trajectoires non ressemblantes dans le voisinage égale à 98%, nous avons obtenu presque 5 000 partitions anormales. La Figure 4-9 présente un focus sur le détroit de Gibraltar[85] qui présente quelques comportements anormaux. Nous distinguons des changements de cap brusques, des attentes loin des ports et des changements de destinations à quelques encablures du port. Ces comportements sont anormaux et peuvent décrire un risque. Nous allons discuter par la suite quelques exemples de motifs de mouvements que nous avons obtenus.

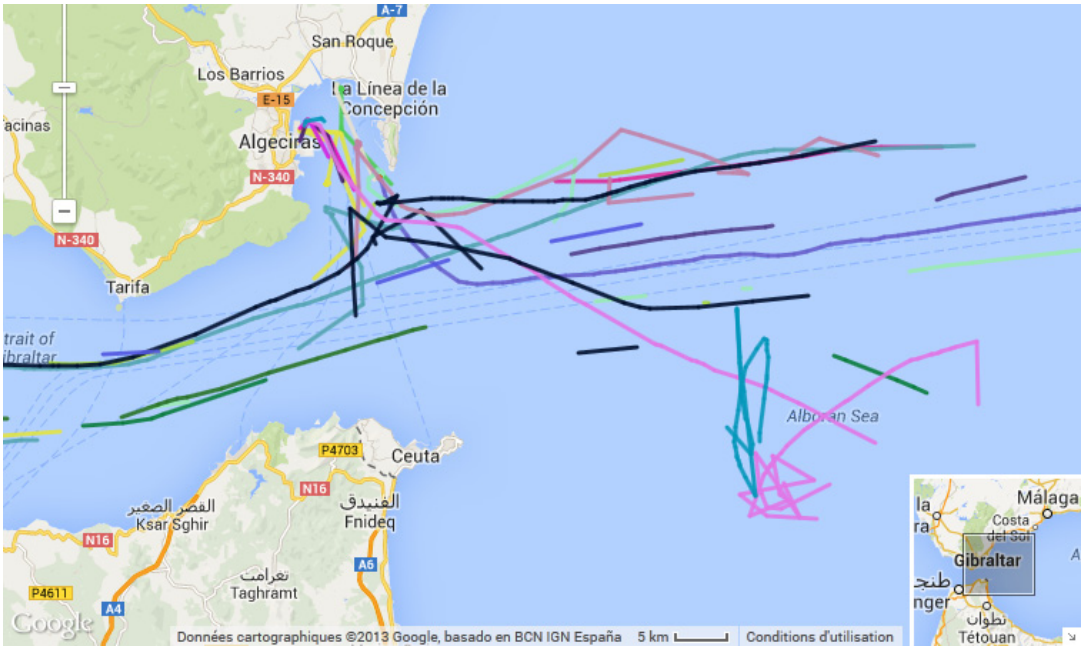

Figure 0-9 : Détection de comportements anormaux de tankers à Gibraltar.

Prenons le cas d'un tanker qui a l'air d'attendre à l'écart du passage de Gibraltar. Ce navire a eu deux comportements consécutifs similaires comme on peut le voir sur la Figure 4-10. Que signifie ce comportement ?

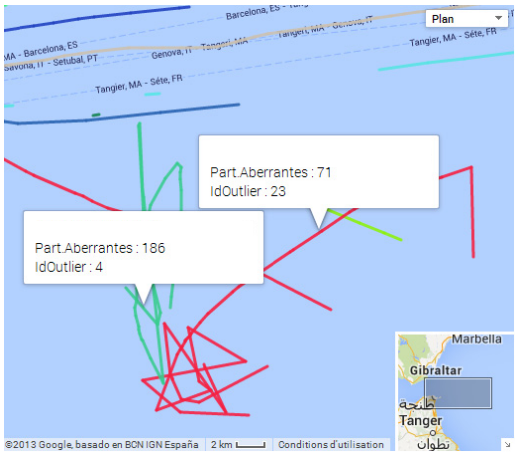

Figure 0-10 : Comportement anormal d'un navire

Ce comportement peut être juste une attente pour rentrer au port à défaut de place

[85] De l'arabe Jabel Tariq, c'est un passage maritime situé au sud de l'Espagne et qui relie entre la mer Méditerranée et l'océan Atlantique.

ou d'autorisation. Il peut être aussi en négociation pour vendre la marchandise transportée au plus offrant comme il peut être un comportement à risque (avarie, trafic illégal, etc.).

Le fait que ce tanker reste loin des ports habituellement fréquentés par ce type de navires, lève des doutes sur ses objectifs réels. En effet, ce comportement peut représenter un risque d'échange frauduleux entre navires ou à partir de la côte en utilisant des embarcations rapides. Le tanker est à 25 km des côtes. De plus, un comportement peut présenter des risques sur le trafic maritime surtout la nuit où il n'y a pas de visibilité.

Le même comportement près des ports (Figure 4-11), peut être considéré comme une attente de déchargement d'une partie ou de toute la marchandise vers un navire plus petit. Ce genre de procédure se fait dans les ports qui ne sont pas adaptés pour recevoir de grands navires. Justement, les Figures suivante (Figure 4-12 et Figure 4-13) montrent des mises en couple[86] réelles entre deux tankers dans un port de Gibraltar.

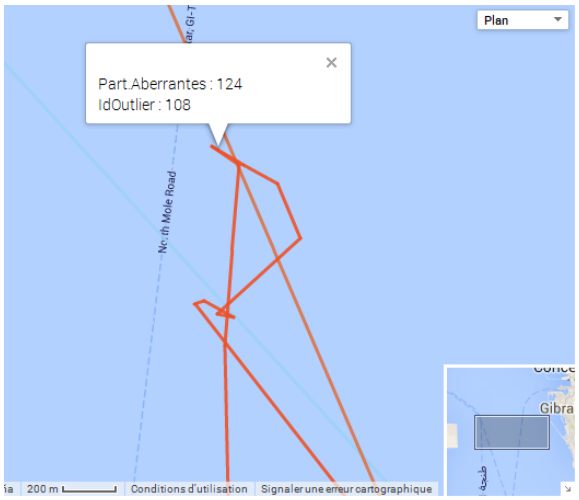

Figure 0-11 : Comportement d'attente et de mise en couple d'un tanker pour déchargement de marchandise.

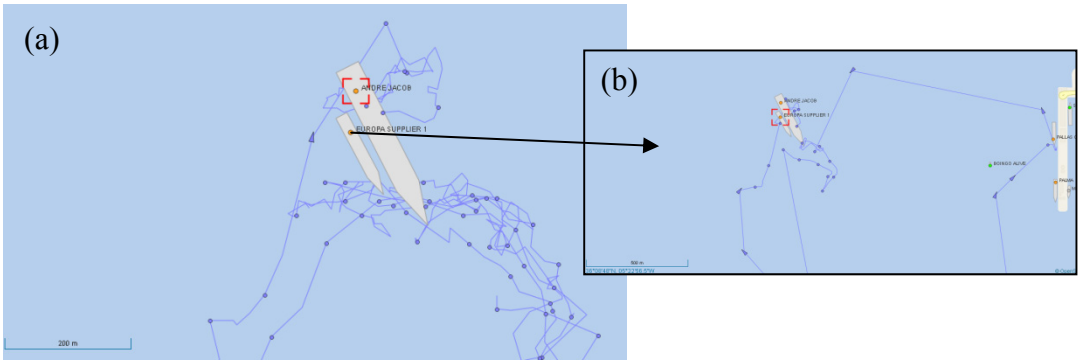

Figure 0-12 : Comportement de déchargement de marchandise -Le grand tanker se met en couple (a) avec un navire qui vient du port (b). (http://www.vesselfinder.com/fr)

[86] Expression utilisée par les navigateurs pour designer deux navires se mettant l'un à côté de l'autre.

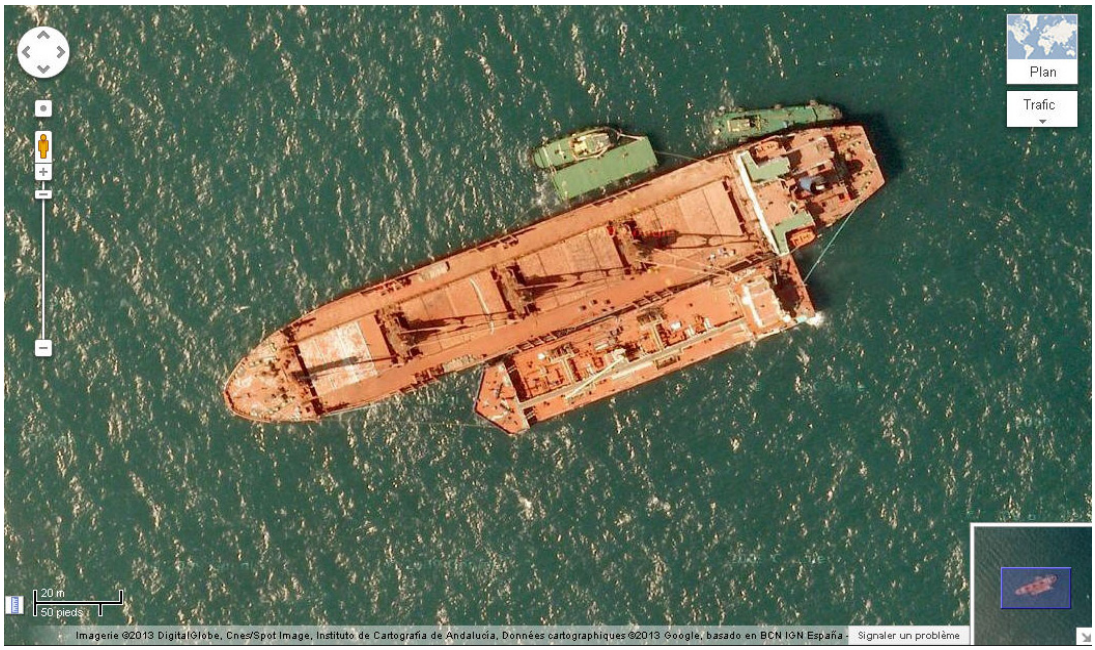

Figure 0-13 : Mise en couple de deux tankers dans un port de Gibraltar
(source Google Maps)

Nous présentons un autre comportement anormal sur la Figure 4-14. Ce comportement peut indiquer une manœuvre d'un tanker pour récupérer un objet tombé à la mer ou pour prendre de la distance avec un autre navire. Ce genre de manœuvre est anormale, risquée et peut engendrer un accident du trafic.

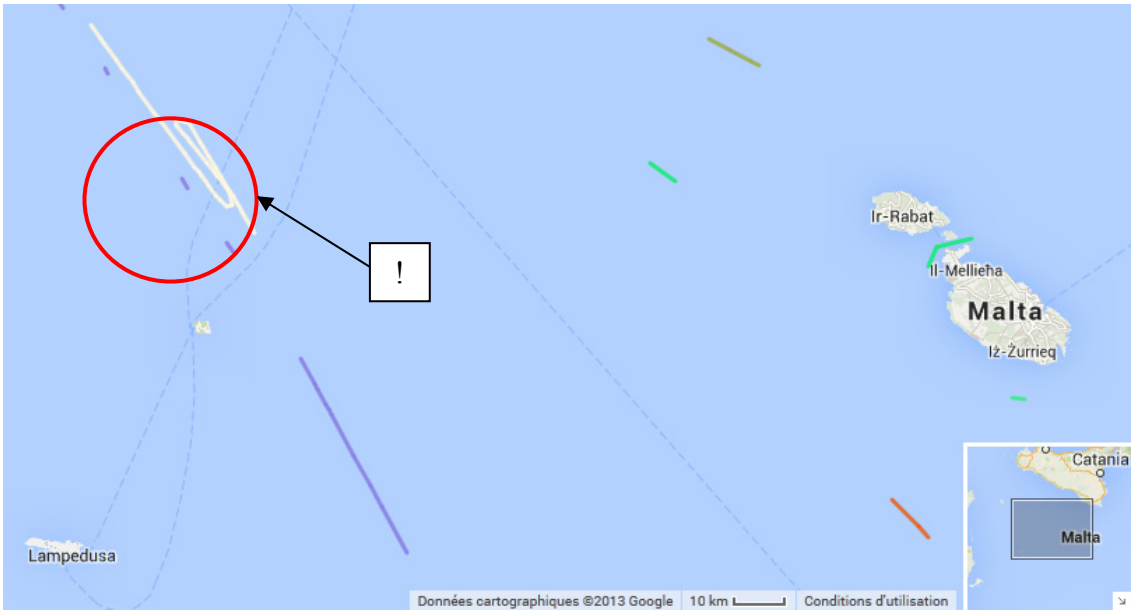

Figure 0-14 : Comportement d'un navire qui change plusieurs fois de destinations.

La Figure 4-15 montre un autre comportement anormal d'un tanker qui sort du port puis qui y revient après plusieurs kilomètres de navigation. Ce comportement peut représenter l'état d'un navire qui a nécessité un retour d'urgence au port pour cause d'avarie par exemple. Nous présentons ci-dessous (Figure 4-16) une sous-trajectoire aberrante qui présente un cas ressemblant à la dérive d'un tanker. Le navire a une trajectoire inhabituelle qui tend vers les côtes puis elle s'arrête à 6 km de la côte.

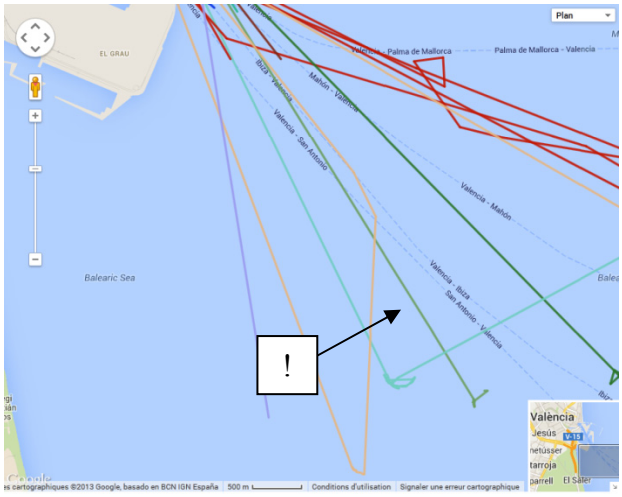

Figure 0-15 : Comportement d'un tanker qui revient au port.

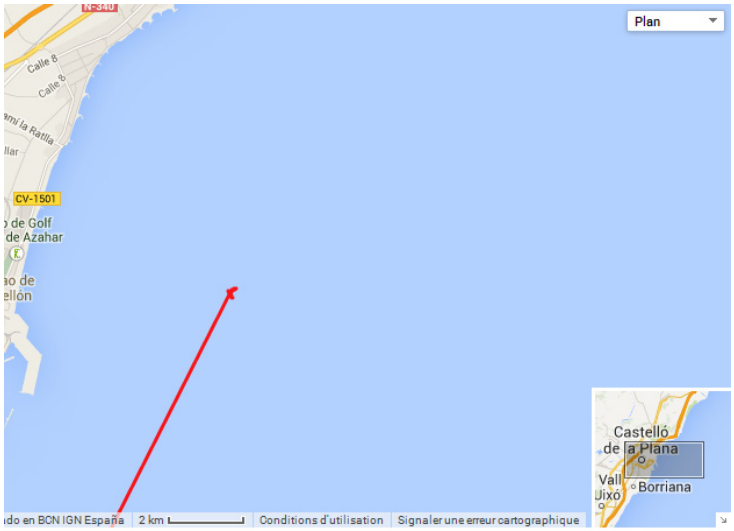

Figure 0-16 : Comportement d'un tanker qui ressemble à une dérive.

Nous avons vu au travers de quelques exemples présentés dans cette section, que la méthode de découverte de trajectoires et sous-trajectoires aberrantes utilisée permet bien d'extraire des motifs de mouvements anormaux qui peuvent décrire des comportements à risques.

4.3.2.2. Navigation proche

L'exploration d'un historique de déplacement de 9 navires de pêches navigants dans les eaux territoriales des îles Féroé (Figure 4-17) a permis d'identifier des comportements de navigations parallèles proches. Ces comportements peuvent indiquer des pêches parallèles qui sont interdites.

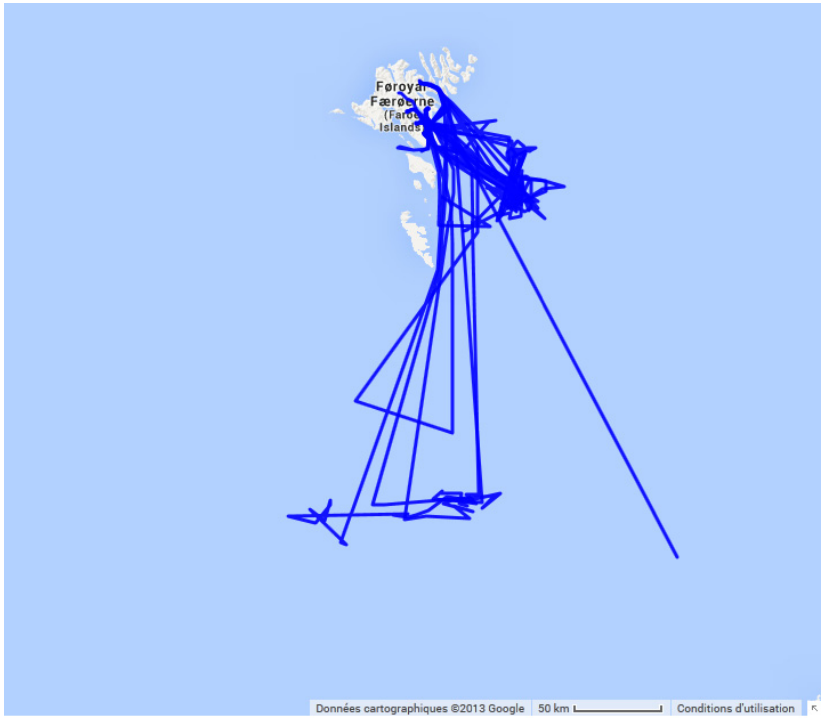

Figure 0-17 : Historique de trajectoires de navires de pêche navigant dans les eaux territoriales des îles Féroé.

L'exploration de ces déplacements de navires de pêche en utilisant la fonctionnalité « Navigation proche » de ShipMine avec un nombre de trajectoires minimal égal à 2, une distance de voisinage maximale égale à 1 kilomètre et une durée minimale de navigation proche égale à 10 minutes, nous a permis de découvrir 8 navigations parallèles. La simplification des trajectoires par l'algorithme Douglas-Peucker temporel (Cf. section 2.2.3.1.4 du chapitre 2) a été effectuée avec une précision (tolérance) de 40 mètres. Nous avons choisi cette valeur par visualisation de différents résultats de trajectoires simplifiés obtenus pour plusieurs valeurs de tolérance. L'idée est de simplifier les trajectoires mais sans les rendre trop lisses.

Nous présentons sur la figure suivante (Figure 4-18), deux comportements de navigation proches pouvant décrire une pêche parallèle. Dans la navigation parallèle numéro 7 par exemple, les deux navires sont restés en parallèle plus de 30 minutes à la date du 15 juin 2013.

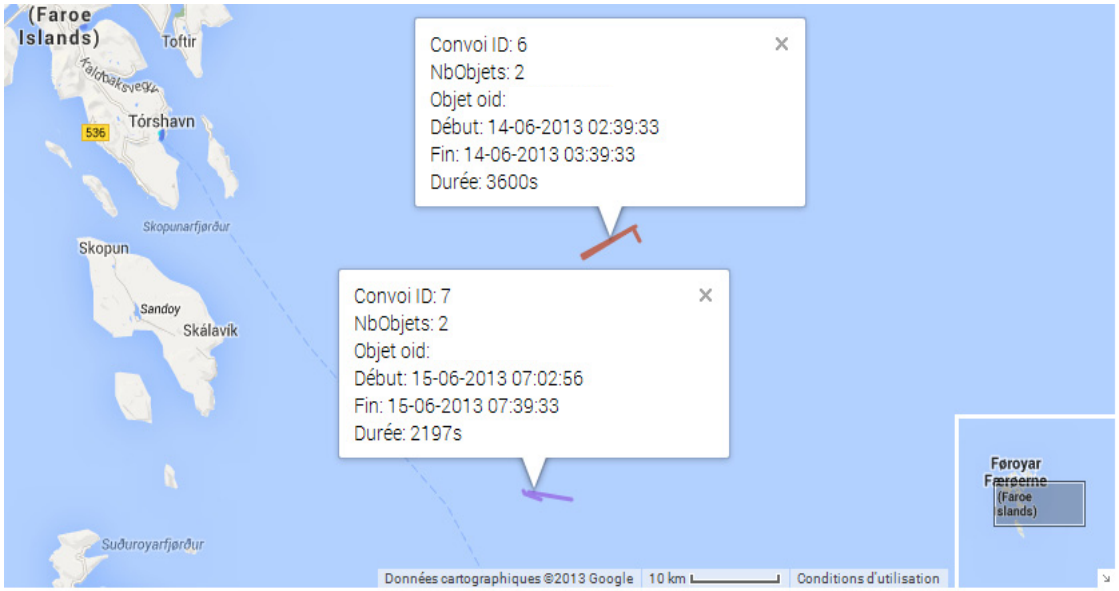

Figure 0-18 : Découverte de pêches parallèles de navires navigant dans les eaux territoriales des îles Féroé.

4.3.2.3. Routes de navigation maritime

Dans cette section, nous allons extraire des trajectoires de navigation habituelles ou type à partir de l'exploration de 14 trajectoires de 2 navires de pêche ayant navigué à proximité du port de Sète. Les traces de ces navigations sont présentées sur la Figure 4-19.

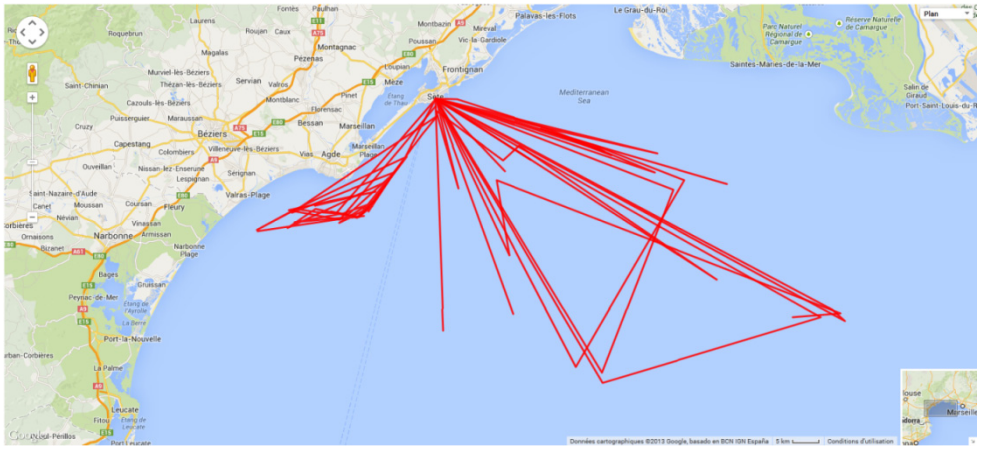

Figure 0-19 : Historique de traces de navigation de 2 navires de pêche dans le port de Sète.

L'exécution de la fonctionnalité « Routes de navigation » de ShipMine sur les 14 trajectoires de navires avec une distance de voisinage de 15 km et un minimum de partitions voisines MinLns égal à 2, nous a permis de découvrir une agrégation des déplacements de ces navires. La route de navigation découverte agrège le comportement de 12 trajectoires comme on peut le percevoir sur la Figure 4-20. Cette route de navigation décrit le comportement général des navires de pêche analysés. Ces navires ont l'habitude de faire des allers-retours entre le port de Sète et la zone sud-est ou sud-ouest avec une navigation parallèle au cap d'Agde. Etant donné que les navires de pêche partent du port de Sète, les deux trajectoires type ont été connectées par densité.

Figure 0-20 : Découverte de routes de navigation de navires de pêche naviguant à proximité du port de Sète.

Les motifs spatiaux décrivant les trajectoires type de navigation peuvent être utilisés pour découvrir les trajectoires aberrantes qui s'écartent du comportement normal. Ces trajectoires anormales peuvent décrire un comportement à risque (cf. section 4.3.2.1) comme ils peuvent décrire des routes à risque de naufrage si on se focalise sur les chalutiers ou des routes à risque de pollution si on se focalise sur les navires de transport de matières dangereuses.

4.4. Limites et pistes d'amélioration

Nous allons exposer dans la section suivante les limites liées à notre proposition méthodologique, à l'atelier mis en œuvre ainsi que quelques pistes d'amélioration possibles.

4.4.1. Méthodologie

La méthodologie utilisée est basée sur la fouille de données qui ne fait pas participer les utilisateurs dans l'exploration de données. C'est une sorte de boite noire pour les utilisateurs. Nous présentons dans les perspectives, une proposition d'amélioration de notre méthode.

La validation de notre méthodologie d'aide à l'extraction de connaissances sur les comportements à risques pourrait être faite sur le terrain avec des experts maritimes comme les capitaines de navire[87]. Cela permettrait de savoir si les connaissances découvertes automatiquement correspondent bien aux idées reçues et aux intuitions des navigateurs. La coïncidence de connaissances permettrait de réconforter les connaissances des navigateurs et valider notre méthodologie. Les connaissances qui ne coïncideraient pas, pourraient être analysées pour en comprendre les raisons : est-ce que c'est dû au fait que les connaissances générées mettent en évidence des situations et mouvements à risques nouveaux, auxquels les navigateurs n'ont pas été confrontés auparavant ou sont-elles biaisées ? Dans ce dernier cas, il serait intéressant d'identifier les causes qui ont amené à générer de telles connaissances.

Les règles d'association n'ont pas permis la découverte de relations négatives entre les itemsets du genre « si A alors non B » (trouver A exclut B et inversement). Les

[87] Contacter par exemple l'Association Française des Capitaines de Navires (http://www.afcan.org/presentation1.html).

relations d'implications des règles d'association sont monodirectionnelles et ne permettent pas la découverte de ce genre de règles. L'extraction de relations de corrélations (Silverstein et al., 1998) peut permettre de découvrir des relations négatives, positives et nulles entre les itemsets. Il serait intéressant d'utiliser cette méthode pour la suite de nos travaux.

Les méthodes de groupement par densité utilisées ne prennent pas en compte les obstacles. Les zones à risques extraites, couvrent des territoires non maritimes. A cette valeur de distance, toutes les localisations d'accidents sont accessibles par densité. Une amélioration possible est d'utiliser des méthodes de clustering spatial prenant en compte cette contrainte d'existence d'obstacles (Tung et al., 2001).

4.4.2. ShipMine

L'atelier ShipMine intègre aujourd'hui quatre algorithmes qui sont DBSCAN, TRAOD, TRACLUS et Convoy. Ces algorithmes supportent des fonctionnalités d'extraction de zones à risques, d'identification de trajectoires aberrantes, de routes de navigation et de navigations proches qui ont permis d'obtenir des résultats prometteurs sur nos données maritimes (cf. section 4.2). Dans une perspective d'amélioration des fonctionnalités de ShipMine, d'autres algorithmes ont été sélectionnés pour leur éventuelle intégration dont deux testés sur les données MAIB, à savoir Apriori et ID3 (cf. sections 4.3.1.1 et 4.3.1.2).

Dans ce travail, nous ne nous sommes pas attardés sur l'étude de la scalabilité, l'optimisation des algorithmes et des programmes de ShipMine pour une utilisation dans le monde opérationnel. Les performances de ShipMine sont suffisantes pour nos besoins expérimentaux mais doivent certainement être améliorées pour un contexte opérationnel.

4.5. Conclusion

Dans ce chapitre, nous avons procédé à la validation de notre méthodologie à l'aide de ShipMine. L'atelier a été utilisé pour extraire des exemples de motifs de mouvements et de situations de navires décrivant des comportements potentiellement à risques. Les tests ont été effectués sur des jeux de données réelles de déplacements et d'enquêtes d'accidents maritimes et ont permis de découvrir des connaissances sur des zones accidentogènes, des trajectoires de dérives, d'abordage, de pêche parallèle et des trajectoires type.

A ce stade de nos recherches, il est possible de répondre à notre problématique qui était de savoir si les motifs et les règles issus de la fouille de données peuvent décrire des comportements à risques. La fouille de données est donc appropriée pour la construction de connaissances sur les comportements à risques de navires.