Rapport de Projet

Détection de patterns dans des séries de voyages

Membres du groupe :

OULMAHDI Riad
KACED Louheb
DJEHA Wassim
KHEYAR Aya
HALIT Rafik
AZOU Ali

Encadrants :

Abdelkader OUALI
Jacques EVERWYN

Entreprise :

Sinay SAS

Date : 13 mai 2025

Année universitaire : 2024-2025

Table des matières

1 Introduction

1.1 Contexte

Le commerce maritime constitue un pilier fondamental de l'économie mondiale, assurant l'essentiel des échanges internationaux. Dans ce secteur stratégique, l'optimisation des itinéraires et de la gestion portuaire représente un enjeu majeur, tant sur le plan économique qu'environnemental. Face à la complexité croissante des flux maritimes, l'exploitation intelligente des données de navigation, notamment celles issues du système AIS, ouvre de nouvelles perspectives d'analyse et de prédiction.

Ce projet s'inscrit dans cette dynamique en mobilisant des approches d'analyse séquentielle pour mieux comprendre et anticiper les trajectoires des navires. Après une première phase axée sur l'exploration des données et l'identification de motifs récurrents dans les parcours, nous avons orienté nos travaux vers la modélisation probabiliste du comportement des navires, en particulier à travers l'usage de chaînes de Markov. L'objectif global a été de construire des outils capables de prédire les escales futures et d'extraire des informations utiles à la planification maritime.

1.2 Entreprise SINAY

SINAY est une entreprise travaillant sur des solutions pour améliorer l'efficacité des entreprises maritimes. L'entreprise collecte et agrège des données provenant de multiples sources telles que la position des navires, la météo, les courants, la faune et la flore et bien d'autres encore. Ces données sont combinées et analysées avec différents algorithmes d'apprentissage automatique pour obtenir les meilleurs indicateurs de performance pour les entreprises maritimes.

1.3 Objectifs du projet

L'objectif principal de ce projet est d'exploiter les données historiques de navigation maritime afin d'extraire des motifs récurrents dans les trajectoires des navires et de développer des outils prédictifs pour améliorer la gestion des itinéraires.

Pour atteindre cet objectif global, plusieurs axes d'analyse complémentaires ont été explorés :

- Extraction de motifs séquentiels fréquents à partir des trajectoires de navires.
- Analyse des ports les plus fréquentés.
- Modélisation des trajectoires à l'aide de chaînes de Markov, dans le but de prédire les escales futures en fonction de l'historique de navigation.
- Comparaison de plusieurs approches de prédiction.
- Évaluation des performances des modèles à l'aide d'une méthodologie de validation rigoureuse.
- Développement d'un outil de génération de plannings automatisés.

2 Fouille de motifs séquentiels

2.1 Notions de base

2.1.1 Base de données séquentielle

Une base de données séquentielle est une structure de données spécifiquement conçue pour stocker et organiser des séquences d'objets, où la dimension temporelle revêt une importance capitale. Les séquences ainsi définies correspondent à des événements ou des transactions qui se déroulent dans un ordre chronologique bien défini. Contrairement aux bases de données transactionnelles classiques, où les transactions sont indépendantes et non ordonnées, les bases de données séquentielles intègrent explicitement l'ordre temporel au sein de leur structure.

Caractéristiques principales :

- Chaque séquence est associée à un identifiant unique, tel qu'un client ou un utilisateur (dans le contexte présent, les **navires**).
- Les éléments (dans notre cas, les **ports**) sont organisés selon leur ordre d'occurrence temporelle.
- La dimension temporelle des événements est cruciale pour l'analyse des séquences.

Exemple de base séquentielle

Nous considérons un tableau illustrant les séquences d'items pour différents clients, avec les dates correspondantes. Les items sont représentés par des lettres et les dates par des valeurs spécifiques :

Client	Date	Items
C_1	01/01/2004	B, F
C_1	02/02/2004	B
C_1	04/02/2004	C
C_1	18/02/2004	H, I
C_2	11/01/2004	A
C_2	12/01/2004	C
C_2	29/01/2004	D, F, G
C_3	05/01/2004	C, E, G
C_3	12/02/2004	A, B
C_4	06/02/2004	B, C
C_4	07/02/2004	D, G
C_4	08/02/2004	I

`https://www.lirmm.fr/~poncelet/publications/papers/motifs_sequentiels.pdf`

Les motifs contenus dans ces séquences jouent un rôle clé dans l'analyse des comportements des utilisateurs et permettent l'identification de tendances ou de motifs récurrents.

2.1.2 Motif séquentiel

Un motif séquentiel est défini comme une séquence d'items ou d'événements qui apparaissent dans un ordre spécifique au sein des données. La **fréquence** d'un motif séquentiel est déterminée par la proportion de séquences contenant ce motif. Un motif est considéré comme fréquent si sa fréquence dépasse un seuil de support minimal prédéfini par l'utilisateur.

Explication d'un motif

Examinons le cas du client C_2 et analysons ses interactions avec les événements survenus dans ses séquences. Soit un motif particulier de la forme :

$$B \to C$$

- **Motif choisi :** $B \to C$, support = 50%.
- **Interprétation :** Les clients C_1 et C_4 effectuent une action associée à l'item **B**, suivie par une action liée à l'item **C**.
- **Utilité :** Ce motif met en évidence une relation entre les produits **B** et **C**. On peut en déduire que l'achat du produit **B** est souvent suivi de l'achat du produit **C**.

2.1.3 Fréquence dans les motifs séquentiels

Le support d'un motif séquentiel est défini comme la proportion de séquences dans la base de données qui contiennent ce motif particulier. Formellement, si un motif σ apparaît dans n_σ séquences sur un total de N séquences, le support S_σ de ce motif est donné par la relation suivante :

$$S_\sigma = \frac{n_\sigma}{N}.$$

Par exemple, si un motif $\{B, I\}$ apparaît dans les séquences de C_1 et C_4, soit dans 2 séquences sur un total de 4, son support est de :

$$S_{\{B,I\}} = \frac{2}{4} = 50\%.$$

Ainsi, un motif est dit fréquent si son support dépasse un seuil minimal prédéfini, souvent noté θ. Par conséquent, un motif σ est fréquent si :

$$S_\sigma \geq \theta.$$

2.2 Algorithme PrefixSpan

Dans le cadre de ce projet, nous avons utilisé l'implémentation Java de l'algorithme PrefixSpan fournie par la bibliothèque SPMF (Sequential Pattern Mining Framework). Cette implémentation permet une extraction efficace des motifs séquentiels fréquents à partir d'une base de données de séquences.

L'algorithme PREFIXSPAN optimise l'extraction de motifs séquentiels fréquents en réduisant le nombre de candidats générés. Il exploite les préfixes communs des séquences pour projeter la base de données en sous-ensembles plus petits, facilitant ainsi la recherche.

1. Extraction des items fréquents
- Une première passe sur la base de données identifie les items fréquents. Ces items servent de préfixes initiaux pour diviser l'espace de recherche.

2. Projection des bases
- Chaque préfixe fréquent définit une base projetée, contenant uniquement les séquences partageant ce préfixe. Seuls les suffixes sont considérés dans ces bases projetées. Cela réduit significativement l'espace de recherche.

3. Exploration récursive
- À partir de chaque base projetée, PREFIXSPAN identifie de nouveaux préfixes fréquents et génère des motifs plus longs. Ce processus se répète de manière récursive jusqu'à ce qu'aucune séquence fréquente ne puisse être étendue.

Une explication plus détaillée de cet algorithme est disponible dans le document complet : `https://www.lirmm.fr/~poncelet/publications/papers/motifs_sequentiels.pdf`.

2.3 Données de l'entreprise

L'entreprise nous a fourni un fichier `JSON` contenant des informations détaillées sur les voyages effectués par différents navires. Ce fichier constitue la base principale de notre analyse. Il contient notamment des éléments pertinents tels que :

- **id** : identifiant unique attribué à chaque voyage ;
- **mmsi** : *Maritime Mobile Service Identity*, un numéro d'identification radio utilisé pour suivre les navires. Ce numéro peut changer au cours de la vie du navire ;
- **imo** : *International Maritime Organization number*, identifiant permanent et unique attribué à chaque navire ;
- **departure_date** : date et heure de départ du port d'origine ;
- **arrival_date** : date et heure d'arrivée au port de destination ;
- **departure_port** : nom du port de départ ;
- **arrival_port** : nom du port d'arrivée.

2.4 Traitement initial

Dans un premier temps, nous avons utilisé les données disponibles pour construire des séquences de ports visités, en utilisant comme identifiant principal le champ `imo`, et en cas d'absence, le champ `mmsi`. Pour cela :

1. Les champs `departure_port` et `arrival_port` ont été sélectionnés pour créer des séquences ordonnées par la date d'arrivée.
2. Plusieurs anomalies ont été identifiées et filtrées :
   — Ports manquants (`null`).
   — Trajets où le port de départ est identique au port d'arrivée.
   — Duplication consécutive de ports due à des chevauchements entre voyages successifs.

Exemple de séquence avant nettoyage :

$$\text{Séquence :}\langle \text{AUBTB} \rightarrow \text{AUPOR} \rightarrow \text{AUPOR} \rightarrow \text{VNSGN} \rightarrow \ldots \rangle$$

Séquence obtenue après correction :

$$\text{Séquence :}\langle \text{AUBTB} \rightarrow \text{AUPOR} \rightarrow \text{VNSGN} \rightarrow \ldots \rangle$$

Lors de la deuxième phase du projet, à la suite de la réception de données supplémentaires fournies par l'entreprise , nous avons identifié la nécessité d'adapter notre traitement initial afin d'assurer une qualité optimale et une meilleure représentativité des séquences. Ces ajustements ont permis de traiter des cas spécifiques et d'affiner le découpage des séquences. Les principales modifications apportées sont les suivantes :

— **Remplacement des ports manquants** : Les ports absents dans les données ont été remplacés par une valeur générique `FICTIF`. Afin d'éviter des erreurs dans le tri chronologique, une date par défaut (le 1er janvier de l'année du voyage) a été attribuée aux trajets commençant par un port `FICTIF` (car généralement c'est à cette date là que les ports sont nulls) .

— **Suppression des doublons d'escales** : Dans le cas où un navire présente plusieurs escales successives identiques (même port et même date), ces doublons ont été supprimés pour éviter les biais dans l'analyse des trajets.

— **Tri temporel systématique** : Tous les trajets ont été rigoureusement triés par date d'arrivée pour chaque navire, afin de refléter fidèlement l'ordre réel des escales et de garantir la cohérence des séquences.

— **Découpage automatique des séquences** : Pour mieux distinguer les phases de navigation des périodes d'arrêt prolongé, une segmentation automatique est effectuée. Le principe est le suivant :
   — La durée de chaque trajet (en jours) est calculée à partir des dates de départ et d'arrivée.
   — Une fonction identifie l'écart **le plus grand** entre deux durées successives dans l'ensemble des trajets.
   — Cet écart est utilisé comme **seuil optimal** : toute accumulation de durées dépassant ce seuil déclenche le début d'une nouvelle séquence.

Exemple illustratif :

Considérons un navire dont les trajets sont les suivants (chaque ligne correspond à un voyage) :

AUPOR → HKE03	(30 jours)
HKE03 → CNBJS	(31 jours)
CNBJS → KRBUS	(32 jours)
KRBUS → USLAX	(90 jours)
USLAX → USPOR	(91 jours)

Les durées successives sont : `30, 31, 32, 90, 91`

Les écarts entre durées : `1, 1, 58, 1`
Le plus grand écart est **58 jours**, entre `32` et `90`

Ce seuil est utilisé pour segmenter les voyages en deux séquences :
— `AUPOR → HKE03 → CNBJS → KRBUS`
— `USLAX → USPOR`

Cette méthode permet une segmentation dynamique qui reflète les interruptions significatives dans l'activité du navire, contrairement à une coupure fixée arbitrairement.

2.5 Adaptation au format de données SPMF

Lors de l'analyse des séquences sous leur forme actuelle (chaînes de caractères représentant les ports), nous avons observé une perte de performance dans les traitements. Cela s'explique par le coût élevé de manipulation des chaînes de caractères, en particulier lorsque le volume de données est important.

Solution proposée

Pour optimiser le traitement des séquences, nous avons adopté une approche consistant à convertir les chaînes de caractères en nombres entiers. Voici les étapes mises en œuvre :

1. **Création d'un dictionnaire de conversion** : Chaque port est associé à un entier unique. Cela permet de représenter les ports de manière plus compacte et rapide à traiter.
2. **Séparation des éléments dans les séquences** :
   — Les différents itemsets (ports) sont séparés par la valeur **-1**.
   — La fin d'une séquence est indiquée par la valeur **-2**.

   Voici un exemple de séquence transformée avec cette méthode :

$$1\ \mathbf{-1}\ 3\ \mathbf{-1}\ 5\ \mathbf{-1}\ 6\ \mathbf{-1}\ 1\ \mathbf{-1}\ 2\ \mathbf{-1}\ \ldots\ \mathbf{-2}$$

Avantages de cette méthode

— **Efficacité** : La manipulation de nombres entiers est bien plus rapide que celle de chaînes de caractères, réduisant ainsi le temps de traitement.
— **Simplicité** : Le format compact des séquences facilite leur analyse et leur exploitation dans des algorithmes.
— **Scalabilité** : Cette méthode est mieux adaptée à un grand volume de données.

2.6 Filtrage des motifs non pertinents

Après l'éxecution de l'algorithme PrefixSpan sur les séquences, un processus de post-traitement est appliqué pour garantir que seuls les motifs pertinents sont retenus. Ce processus a pour but d'exclure les motifs qui ne reflètent pas notre objectif d'identifier des lignes maritimes régulières. Les étapes suivantes sont effectuées :

1. **Exclusion des motifs de taille 1** : Les motifs contenant un seul port ne sont pas pertinents, car ils ne représentent pas des itinéraires maritimes significatifs.
2. **Suppression des motifs avec des ports consécutifs identiques** : Si un motif contient deux ports identiques consécutifs, il existe forcément un motif équivalent sans cette duplication, ayant le même support. Par conséquent, ces motifs sont éliminés pour éviter les redondances.
3. **Exclusion des motifs avec un unique port répété plusieurs fois** : Les motifs composés d'un seul port répété (par exemple : $\langle AUPOR \rightarrow AUPOR \rightarrow AUPOR \rangle$) sont exclus, car ils ne représentent pas des lignes maritimes, mais plutôt des mouvements locaux ou des anomalies.

3 Expérimentations et analyse

Dans cette section, nous présentons le protocole expérimental ainsi que les résultats obtenus à partir de notre base de données séquentielle.

3.1 Protocole expérimental

Les expérimentations ont été réalisées sur une base de données séquentielle contenant 1552 navires. Nous avons fait varier le support minimal de 1% à 20%. Les métriques observées incluent :

— le temps d'exécution (en milli-secondes),
— le nombre de motifs trouvés,
— la taille moyenne des motifs,
— l'écart-type des tailles des motifs.

3.2 Résultats expérimentaux

Support minimum	**Nombre de navires**	**Temps pris (ms)**	**Nombre de motifs trouvés**	**Taille moyenne des motifs**	**Écart type**
1%	16	131600.55	6462944	10.19	2.02
4%	62	569.71	1539	4.47	2.10
7%	109	437.76	225	3.29	1.39
10%	155	347.67	51	2.37	0.62
13%	202	337.82	8	2.00	0.00
16%	248	334.06	1	2.00	0.00
19%	295	332.50	0	nan	nan
20%	310	339.00	0	nan	nan

TABLE 1 – Résultats d'expérimentation

3.3 Interprétation des résultats

Les résultats expérimentaux montrent que plus le seuil de support minimum augmente, moins il y a de motifs fréquents trouvés, certains seuils élevés (comme 0.17) ne produisant aucun motif. Le temps d'exécution suit la même logique : il est élevé avec un support faible en raison du grand nombre de motifs à explorer, et devient très court lorsque le support augmente. De plus, la taille moyenne des motifs ainsi que leur écart-type diminuent avec un support plus élevé, traduisant une élimination progressive des séquences longues et une homogénéité croissante des motifs conservés.

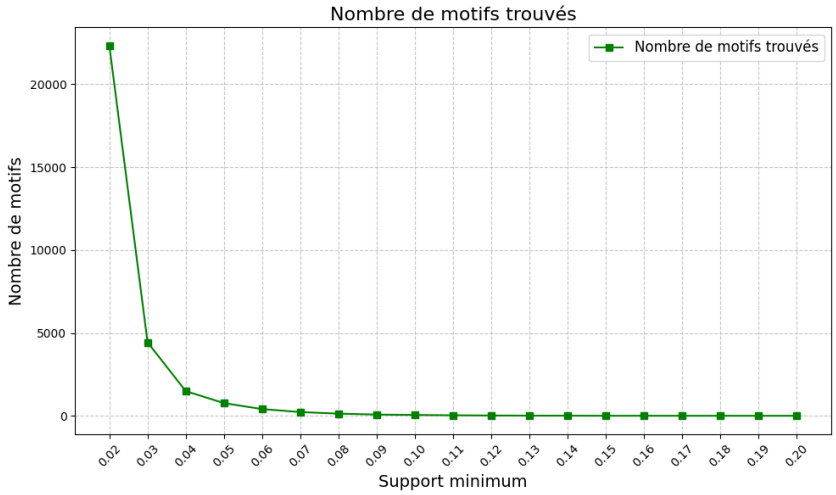

FIGURE 1 – Exemple d'un graphe visualise le nombre de motifs trouvés

3.4 Analyse des résultats de PrefixSpan

Voici un échantillon de motifs fréquents extraits à l'aide de l'algorithme PrefixSpan :

— {PATBG} → {PACTB} → {PAROD} → {PACTB} → {NLMSV} → {PAROD} → {PACTB} → {BEANR} → {PACTB} → {PAROD} → {PATBG} → {PAROD} → {PACTB} #SUP : 16

— {CNE76} → {KRBNP} → {CNE76} → {KRBNP} → {CNE76} → {KRBNP} #SUP : 43

— **{GBLGP} → {BEANR} → {NLMSV} #SUP : 63**

— {USSAV} → {PATBG} #SUP : 62

— {COBUN} → {PATBG} → {PAROD} → {PATBG} → {COBUN} #SUP : 33

Chaque motif correspond à une séquence consécutive de ports fréquents dans la base de données, où SUP indique le support, c'est-à-dire le nombre de séquences dans lesquelles ce motif apparaît.

Analyse :

Si nous prenons, par exemple, le motif **GBLGP → BEANR → NLMSV** avec SUP : **63**, cela signifie que 63 navires ont suivi ce trajet, c'est-à-dire qu'ils sont passés par ces ports dans cet ordre. Cependant, entre deux ports, comme **GBLGP → BEANR**, il est possible qu'ils aient effectué d'autres trajets intermédiaires, tout comme il est possible qu'ils aient effectué ce trajet directement.

4 Génération des plannings

Dans le cadre de ce travail, nous avons développé un script pour produire un planning spécifique pour chaque navire donné, à partir de son numéro IMO.

Pour la conception de ce script, nous nous sommes basés sur l'exemple de planning fourni par l'entreprise, en suivant la même structuration des données. Le script lit le fichier source `merged_voyages.json`, contenant les informations sur plusieurs voyages. Il invite l'utilisateur à saisir l'IMO du navire désiré, puis filtre et extrait uniquement les voyages correspondant à cet IMO.

Pour chaque voyage trouvé, les informations essentielles sont extraites et organisées de manière structurée sous forme de dictionnaire. Ces informations incluent :
— L'ID et l'IMO du navire ;
— Les ports d'origine et de destination ;
— Les dates de départ et d'arrivée ;
— La durée de transit : nombre de jours écoulés entre le départ et l'arrivée (calculée par soustraction des dates) ;
— D'autres champs structurants tels que `SCAC CODE` ou `SCHEDULE TYPE`, qui sont actuellement laissés vides (`None`). Ces champs pourront être complétés ultérieurement si les données deviennent disponibles.


```
{
    "P2P_ID": 228283,
    "VOYAGE_ID": null,
    "CARRIER_ALIAS": "Unknown Carrier",
    "SCAC_CODE": null,
    "CARRIER_SERVICE_DES": null,
    "VESSEL_NAME": "Vessel 9625906",
    "VESSEL_IMO": "9625906",
    "VOYAGE": null,
    "ORIGIN": "AUBTB",
    "ORIGIN_PORT_CODE": "AUBTB",
    "ORIGIN_TYPE_CODE": null,
    "ORIGIN_EVENTDATE": "07/11/2022 21:10",
    "DESTINATION": "AUPOR",
    "DESTINATION_PORT_CODE": "AUPOR",
    "DESTINATION_TYPE_CODE": null,
    "DESTINATION_EVENTDATE": "07/16/2022 13:27",
    "ROUTING": "AUBTB-AUPOR",
    "MODIFIED_DATE": null,
    "AMENDMENT_CODE": null,
    "NO_OF_TRANSSHIPMENTS": null,
    "TRANSIT_TIME": 4,
    "SCHEDULE_TYPE": null
},
```


FIGURE 2 – Exemple d'un extrait de planning généré pour un navire

5 Prédiction des ports

5.1 Chaînes de Markov

Pour prédire les escales futures d'un navire, nous avons utilisé un modèle probabiliste basé sur les chaînes de Markov. Ce type de modèle repose sur une hypothèse simple : la prochaine position dépend uniquement des positions précédentes (sur un nombre défini d'étapes). Ce type de modélisation est particulièrement adapté pour représenter les comportements séquentiels, comme les trajets maritimes.

Pourquoi une chaîne de Markov ?
— Elle permet de capturer les régularités de transition entre ports.
— Elle est adaptée à la prévision à court terme dans des séquences ordonnées.
— Elle est facile à interpréter, grâce à une matrice de transition.

5.2 Modèle Markovien d'ordre variable

Nous avons construit un modèle Markovien de manière flexible, permettant de choisir l'ordre n de la chaîne (avec $n \geq 1$). Cela signifie que la prédiction du prochain port peut être conditionnée à un, deux, ou plusieurs ports précédents.

La formule suivante correspond au **modèle de Markov d'ordre 2** : la probabilité de transition vers un port k, après avoir visité successivement les ports $i \to j$, est donnée par :

$$P(k \mid i, j) = \frac{\text{Occ}(i \to j \to k)}{\sum_{l \in \mathcal{P}} \text{Occ}(i \to j \to l)}$$

où $\mathcal{P}$ est l'ensemble de tous les ports possibles, et où l représente un port quelconque (y compris k). Cette formule exprime la fréquence relative des transitions vers k parmi toutes les destinations observées après la séquence $i \to j$.

Cette formule peut être généralisée à un **modèle de Markov d'ordre** n, en considérant les n derniers ports visités ($s_1 \to s_2 \to \cdots \to s_n$). La probabilité de transition vers un port k s'écrit alors :

$$P(k \mid s_1, s_2, \ldots, s_n) = \frac{\text{Occ}(s_1 \to s_2 \to \cdots \to s_n \to k)}{\sum_{l \in \mathcal{P}} \text{Occ}(s_1 \to s_2 \to \cdots \to s_n \to l)}$$

où, de la même manière, l désigne un port quelconque de $\mathcal{P}$, y compris k. Cette généralisation permet de prendre en compte des séquences plus longues dans l'historique des ports visités, ce qui peut améliorer la précision du modèle dans certains contextes.

Le choix de l'ordre est laissé à l'utilisateur, et nous avons testé plusieurs ordres ($n = 1, 2, 3$) pour comparer leur impact sur la précision et la complexité du modèle. Un ordre plus élevé permet d'être plus spécifique mais demande plus de données pour estimer correctement les transitions.

5.3 Implémentation et simulation

Nous avons implémenté deux approches complémentaires de modélisation par chaînes de Markov, en nous basant sur les séquences de ports visités par les navires.

1. Modèle de Markov basé sur les séquences de ports

Dans cette approche, nous avons considéré les séquences brutes de ports visités par un navire ou plusieurs, sous forme de listes ordonnées comme l'exemple suivant :

`[FICTIF, KRPUS, KRSBU, HKE03, HKE03, VNHPH, VNNHD, ..., CNE83]`

La méthode de calcul des probabilités de transition se déroule en trois étapes :

1. Parcours de la séquence de ports avec une fenêtre glissante de taille égale à l'ordre n.
2. À chaque position, incrémentation du compteur d'occurrence de la transition observée (par exemple : $i \to j \to k$), grâce a une HashMap contenant toutes les transitions observées et leurs nombre d'occurences respectif.
3. Calcul de la probabilité de transition en divisant chaque occurrence par le total des transitions sortantes depuis le même état (historique de taille n).

Cela nous a permis de construire un modèle de Markov classique d'ordre n, où les probabilités de transition sont fondées directement sur la fréquence d'apparition des séquences dans les données complètes de trajets.

2. Modèle de Markov basé sur les motifs séquentiels réguliers

Dans cette deuxième approche, nous avons extrait des motifs séquentiels fréquents (ou réguliers) à partir des mêmes données, à l'aide de l'algorithme PrefixSpan. Ces motifs sont des sous-séquences de ports récurrentes, associées à leur support (fréquence d'apparition).

Par exemple :

— `{HKE03} →{VNHPH}` #SUP : 11
— `{HKE03} →{VNHPH} →{HKE03}` #SUP : 10
— `{HKE03} →{VNHPH} →{CNE83}` #SUP : 4

À partir de ces motifs extraits, nous avons construit une chaîne de Markov où :

— Chaque état est défini par un ou plusieurs ports apparaissant dans une séquence fréquente.
— Les transitions sont pondérées selon le support (nombre d'occurrences) des motifs contenant la transition.

Les probabilités de transition sont obtenues en normalisant les supports : pour une même condition de départ (par exemple `HKE03 →VNHPH`), on divise le support de chaque transition candidate par la somme des supports des transitions possibles depuis ce même préfixe.

Cette méthode exploite la **régularité structurelle** des motifs fréquents pour rendre le modèle plus robuste et interprétable, en se concentrant sur les séquences typiques plutôt que l'ensemble exhaustif des trajets.

Ainsi, les deux modèles s'appuient sur les mêmes données sources mais avec des logiques de construction différentes : le premier est fondé sur les trajectoires complètes, tandis que le second s'appuie sur les motifs récurrents identifiés dans ces trajectoires.

5.4 Exemple de sortie textuelle simulée

```
Si le navire était à HKE03 → VNHPH, il ira probablement vers :
  - HKE03 avec une probabilité de 0.92
  - CNE83 avec une probabilité de 0.08
```

5.5 Visualisation

Nous avons également généré une matrice de transition sous forme de carte thermique (heatmap), facilitant l'interprétation des probabilités. Chaque ligne représente un état (ex. deux derniers ports), et chaque colonne un port de destination possible.

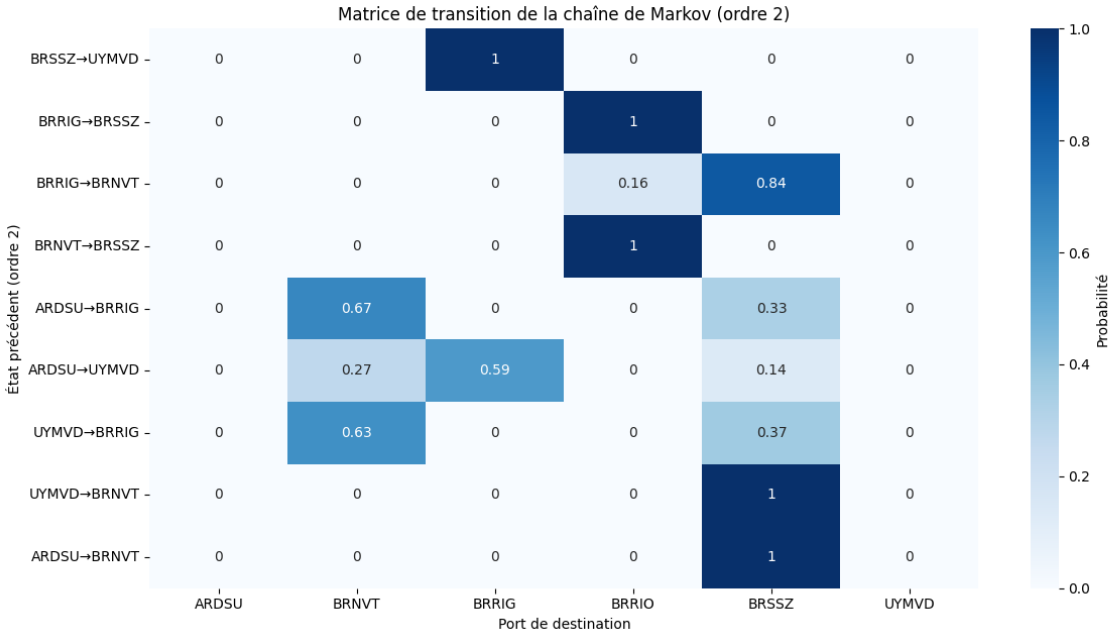

FIGURE 3 – Matrice de transition d'une chaîne de Markov (ordre 2)

3. Comparaison des performances

Nous avons évalué empiriquement les deux approches de chaînes de Markov (basée sur les séquences complètes et basée sur les motifs fréquents) en utilisant un modèle d'ordre 5 sur l'ensemble des trajectoires de la flotte. Les résultats obtenus montrent une nette supériorité du modèle fondé sur les séquences :

Modèle	**Moyenne de précision**	**Écart type**
Séquences (ordre 5)	**63.64%**	**0.84%**
Motifs fréquents (ordre 5)	**22.07%**	**9.82%**

TABLE 2 – Comparaison des performances

Ces résultats indiquent que le modèle de Markov basé sur les **séquences complètes** est nettement plus performant, à la fois en termes de précision moyenne et de stabilité (faible écart type). À l'inverse, le modèle basé sur les **motifs séquentiels fréquents** présente une précision plus faible et une variabilité plus élevée d'une exécution à l'autre.

Cela peut s'expliquer par le fait que les motifs séquentiels, bien qu'informatifs, ne couvrent qu'une partie des trajectoires observées et ignorent les transitions moins fréquentes, ce qui peut nuire à la généralisation du modèle pour certaines séquences de ports.

4. Limites d'un modèle individuel par navire

Nous avons également testé l'efficacité des chaînes de Markov sur les trajectoires d'un **seul navire**, en comparant les performances pour différents ordres du modèle. Cependant, ces expérimentations se sont révélées peu concluantes, en raison de la faible quantité de données disponibles. En effet, un seul navire fournit en moyenne une suite de seulement 30 à 40 ports visités, ce qui est insuffisant pour estimer de manière fiable les probabilités de transition, en particulier pour les ordres élevés.

Les résultats obtenus illustrent cette instabilité :

Ordre du modèle	**Précision (%)**
1	58.82
2	78.57
3	63.64
4	25.00
5	66.67

On observe une **grande variabilité** des performances selon l'ordre choisi, ce qui reflète le manque de données d'entraînement. À titre de comparaison, la flotte complète regroupe environ **1550 fois plus de données** qu'un seul navire, ce qui permet de construire des modèles statistiquement plus robustes et stables. Ces observations confirment l'intérêt de regrouper les trajets de plusieurs navires pour obtenir des prédictions fiables.

5.6 Analyse de la stabilité des performances selon la quantité de données

Comparaison entre un seul navire et la flotte complète

Nous avons évalué les performances du modèle de chaînes de Markov sur deux échelles différentes :

— Sur les séquences complètes de **toute la flotte**, qui regroupent les trajectoires de plusieurs milliers de navires.

— Sur les séquences de **voyages d'un seul navire**, contenant en moyenne 30 à 40 ports visités.

Résultats obtenus

Sur l'ensemble de la flotte :
(environ 17 000 transitions observées), les résultats montrent une amélioration progressive de la précision avec l'augmentation de l'ordre de la chaîne :

Ordre	Précision
1	21.55%
2	34.77%
3	48.11%
4	58.02%
5	63.65%
6	66.40%
7	67.82%
8	68.83%
9	69.13%
10	71.92%

Sur un seul navire :
Bien que certaines précisions puissent paraître élevées (jusqu'à 78.57% à l'ordre 2), les performances sont très instables d'un ordre à l'autre :

Ordre	Précision
1	70.59%
2	78.57%
3	63.64%
4	25.00%
5	75.00%

Interprétation

Malgré quelques taux de précision élevés sur un seul navire, ces résultats sont à relativiser. En effet, le nombre de transitions observées est très faible (par exemple 11 transitions à l'ordre 3), ce qui rend le modèle peu fiable et très sensible à la moindre erreur.

Par exemple :

— **Un score de 63.64% sur 11 transitions** (7 bonnes prédictions) semble élevé ;
— Mais il est bien moins pertinent que **58.02% sur 6 687 transitions**, obtenu avec le modèle basé sur toute la flotte.

Plus la base de données est importante, plus les probabilités estimées sont robustes et les prédictions fiables. Ainsi, le modèle global fondé sur l'ensemble de la flotte est statistiquement plus stable et pertinent pour des applications pratiques, notamment en termes de généralisation.

6 Validation croisée

Afin d'évaluer rigoureusement les performances de nos modèles de chaînes de Markov, nous avons mis en place une procédure de validation croisée. Celle-ci permet de mesurer la capacité du modèle à généraliser sur des données nouvelles, tout en limitant le risque de surapprentissage. La démarche adoptée est structurée comme suit :

Nous avons utilisé une approche classique de validation croisée en **répartissant les données en deux ensembles** :

— **80% des trajectoires** sont utilisées pour l'apprentissage ;
— **20% restantes** sont réservées à l'évaluation des performances du modèle.

Extraction des motifs (cas du modèle basé sur les motifs)

Sur l'ensemble d'entraînement, nous avons appliqué l'algorithme **PrefixSpan** afin d'extraire les motifs séquentiels fréquents. Ces motifs sont ensuite utilisés pour construire des matrices de transition, dans lesquelles chaque transition représente la probabilité de passer d'un motif ou d'un port à un autre.

Évaluation des prédictions

Sur l'ensemble de test, la procédure de test repose sur les étapes suivantes :

1. **Masquage du port suivant** : dans chaque séquence de test, le port suivant à prédire est temporairement masqué.
2. **Prédiction** : le modèle tente de prédire le port masqué à partir du contexte (les ports précédents).
3. **Évaluation du score** : un score de 1 est attribué si la prédiction est correcte, 0 sinon.
4. **Moyenne des scores** : la moyenne de ces scores sur l'ensemble des séquences de test donne la **précision du modèle** et sa **variance**.

Répétition et robustesse de l'évaluation

Pour garantir la robustesse de l'évaluation et éviter toute dépendance à un découpage particulier des données, la procédure complète est répétée **5 fois** avec des répartitions aléatoires différentes des données sans chevauchement.

7 Résultats et analyse

Fold	Score correct	Total prédictions	Précision
1	2571	4064	63.26%
2	2592	4162	62.28%
3	3037	4746	63.99%
4	2793	4371	63.90%
5	3002	4634	64.78%
Moyenne	—	—	**63.64%**
Écart-type	—	—	**0.84%**

Interprétation :

Les résultats sont globalement **cohérents et stables** entre les différents plis, avec une faible variance (écart-type de seulement 0.84%). Cela montre que le modèle est fiable et ne dépend pas de manière significative de la répartition des données d'entraînement et de test. La précision moyenne obtenue sur les cinq plis est de **63.64%**, ce qui est en accord avec les performances observées lors de l'évaluation complète sur toute la flotte.

8 Répartition des tâches

Bien que chaque partie du projet est prise par un ou plusieurs membres, le groupe a fonctionné de manière collaborative. Toutes les étapes ont été réalisées collectivement avec entraide et échanges constants. La répartition suivante reflète une organisation formelle pour structurer le travail.

- **Prétraitement des données et gestion des séquences**
  Responsables : Wassim , Louheb
  - Nettoyage des ports fictifs, doublons et valeurs manquantes
  - Transformation des données brutes en séquences de ports
  - Séparation des données par navire ou par flotte
  - Sauvegarde et formatage des séquences pour usage ultérieur
- **Extraction des motifs séquentiels fréquents**
  Responsables : Ali
  - Implémentation de l'algorithme PrefixSpan
  - Réglage du seuil de support minimum
  - Analyse des motifs extraits (fréquence, longueur)
  - Préparation des motifs pour la modélisation de Markov
- **Modélisation avec chaînes de Markov (modèle sur séquences et motifs)**
  Responsables : Riad , Aya
  - Construction des matrices de transition pour différents ordres
  - Implémentation de la prédiction via les chaînes de Markov
  - Construction des matrices de transition à partir des motifs
  - Prise en compte des supports dans le calcul des probabilités
  - Génération de prédictions
- **Évaluation, validation croisée et visualisation**
  Responsables : Ali, Louheb, Rafik
  - Mise en place de la validation croisée (5 folds, 80/20)
  - Calcul des scores de précision pour chaque fold
  - Calcul des moyennes et écarts-types
  - Création de graphiques et tableaux de synthèse
- **Rédaction, présentation et support visuel**
  Responsables : tout le groupe
  - Rédaction du rapport en LaTeX
  - Création du diaporama de présentation
  - Répartition des parties orales entre les membres
  - Intégration des illustrations, schémas et résultats
- **Travail collaboratif (tous)**
  - Relecture croisée et corrections
  - Tests croisés des implémentations
  - Coordination et gestion de projet
  - Entraînement pour la soutenance orale

9 Conclusion

Ce projet a permis d'explorer des techniques avancées d'analyse de données appliquées au domaine maritime, en particulier à travers l'exploitation des données AIS. Dans une première phase, nous avons mis en œuvre des algorithmes de détection de patterns pour identifier des lignes régulières dans les trajets de navires. Cela a nécessité des traitements préalables de nettoyage, de transformation et de représentation des trajectoires sous forme séquentielle.

Nous avons ensuite intégré des méthodes d'analyse de similarité pour regrouper les itinéraires comparables, avant de développer un modèle basé sur les chaînes de Markov afin de prédire la destination des navires en fonction de leur position actuelle. L'ensemble du pipeline a été évalué grâce à des approches de validation croisée, garantissant la robustesse des résultats.

Les outils mobilisés (Python, SPMF, bibliothèques de traitement de données) ainsi que la rigueur méthodologique adoptée tout au long du projet nous ont permis de produire une solution fiable et extensible pour la détection de comportements maritimes répétitifs et la prédiction d'itinéraires.

Ce travail constitue une première étape vers un système d'analyse intelligent des flux maritimes. Il pourra être poursuivi par l'intégration de données contextuelles (météo, ports de transit, etc.) et l'utilisation de modèles plus complexes (réseaux de neurones, apprentissage profond) pour une meilleure prise en compte de la dynamique des trajets.

Perspectives

Ce travail préliminaire constitue une première étape vers une compréhension approfondie des trajectoires maritimes à partir des données AIS. Les développements réalisés pourront être prolongés et enrichis lors des stages à venir au sein de l'entreprise Sinay.

— **Stage 1 – Analyse des Lignes Maritimes Régulières (LMR)** : Ce stage aura pour objectif de proposer une mesure de similarité entre les trajectoires AIS réelles et les LMR théoriques, de vérifier la pertinence des plannings fournis par les compagnies, et de réaliser un regroupement de navires selon leurs motifs de navigation. Il s'appuiera notamment sur l'extraction de motifs séquentiels et l'étude des trajectoires à plusieurs échelles.

— **Stage 2 – Prédiction de Ports de Destination** : Dans un second temps, l'analyse se portera sur la modélisation des séquences de ports visités à l'aide de chaînes de Markov, afin de prédire le port de destination le plus probable d'un navire. Ce stage pourra également intégrer des méthodes plus avancées comme la programmation par contrainte ou des modèles boîte noire, selon le temps disponible.

Ces deux stages permettront d'approfondir les problématiques abordées, en combinant des approches analytiques et probabilistes, dans un objectif d'optimisation et de prédiction au service de la logistique maritime.

Références

[1] *Wikipedia GSP Algorithm, The Free Encyclopedia.* `https://en.wikipedia.org/wiki/GSP_algorithm`

[2] *Motifs séquentiels.* `https://www.lirmm.fr/~poncelet/publications/papers/motifs_sequentiels.pdf`

[3] NUMPY. *Fundamental package for scientific computing with Python.* `https://www.numpy.org`.

[4] PANDAS. *Python data analysis library.* `https://www.pandas.pydata.org`.

[5] Matplotlib. `https://matplotlib.org/`.

[6] *SPMF*, `https://www.philippe-fournier-viger.com/spmf/`

[7] *Markov Chains - Wikipedia.* `https://en.wikipedia.org/wiki/Markov_chain`

[8] *Cross-Validation – Scikit-learn Documentation.* `https://scikit-learn.org/stable/modules/cross_validation.html`

Détection de motifs dans les séries de trajectoires maritimes

Projet réalisé par :
Riad Oulmahdi – Wassim Djeha – Aya Kheyar
Ali Azou – Louheb Kaced – Rafik Halit

Projet encadré par :
Abdelkader Ouali – Jacques Everwyn

Résumé

Ce projet s'inscrit dans le cadre de l'analyse des trajectoires maritimes à partir des données AIS (Automatic Identification System). L'objectif principal est de détecter des motifs récurrents dans les parcours des navires, notamment les lignes maritimes régulières, à l'aide de mesures de similarité et d'algorithmes de clustering. Une attention particulière est portée à l'évaluation de la conformité entre les trajets effectués et les itinéraires théoriques. Une seconde phase du projet est consacrée à la prédiction du port de destination à l'aide de modèles probabilistes fondés sur les chaînes de Markov. Ce travail a été réalisé dans le cadre d'un stage de Master 2 et se poursuivra par un approfondissement en entreprise, orienté vers l'intégration des résultats dans un système opérationnel d'aide à la décision maritime.

MOTS-CLÉS : AIS, trajectoires maritimes, motifs récurrents, similarité, lignes régulières, chaînes de Markov, prédiction,validation croisé

Abstract

This project focuses on the analysis of maritime trajectories based on AIS (Automatic Identification System) data. The main goal is to detect recurrent patterns in ship trajectories—particularly regular shipping lines—using similarity measures and clustering algorithms. Special attention is given to assessing the compliance of actual routes with planned itineraries. A second part of the project addresses the prediction of the destination port using probabilistic models based on Markov chains. This work was carried out as part of a Master's internship and will continue with a second phase in a company, aimed at integrating the findings into a real-world maritime decision support system.

KEYWORDS : AIS, maritime trajectories, recurrent patterns, similarity, regular lines, Markov chains, prediction,cross validation