université de Caen Basse-Normandie

SINAY
MARITIME DATA SOLUTION

Détection de patterns dans des séries de voyages

Présentation Mi-parcours

Wassim Aya Riad Louheb Ali Rafik

UFR des Sciences
Université de Caen Normandie

27 janvier 2025

Plan

Introduction

Introduction

Qu'est-ce que Sinay ?

Sinay est une entreprise qui propose des solutions de données pour l'industrie maritime, comme le suivi des navires et l'analyse des conditions océaniques.

En quoi consiste ce projet ?

Le projet vise à extraire des motifs et détecter des patterns dans des séries de voyages, ainsi identifier des lignes maritimes régulières (LMR).

Objectifs

Objectifs

Objectif principale

Cette étude vise à identifier des lignes maritimes régulières (**LMR**) dans les historiques de trajets des navires.

Pourquoi extraire les LMR ?

- **Prédiction** : Anticiper les ports à visiter pour une meilleure gestion logistique.
- **Trajets réel et proposé** : Comparer entre les trajets réels et proposés pour identifier des optimisations.

Lignes Maritimes

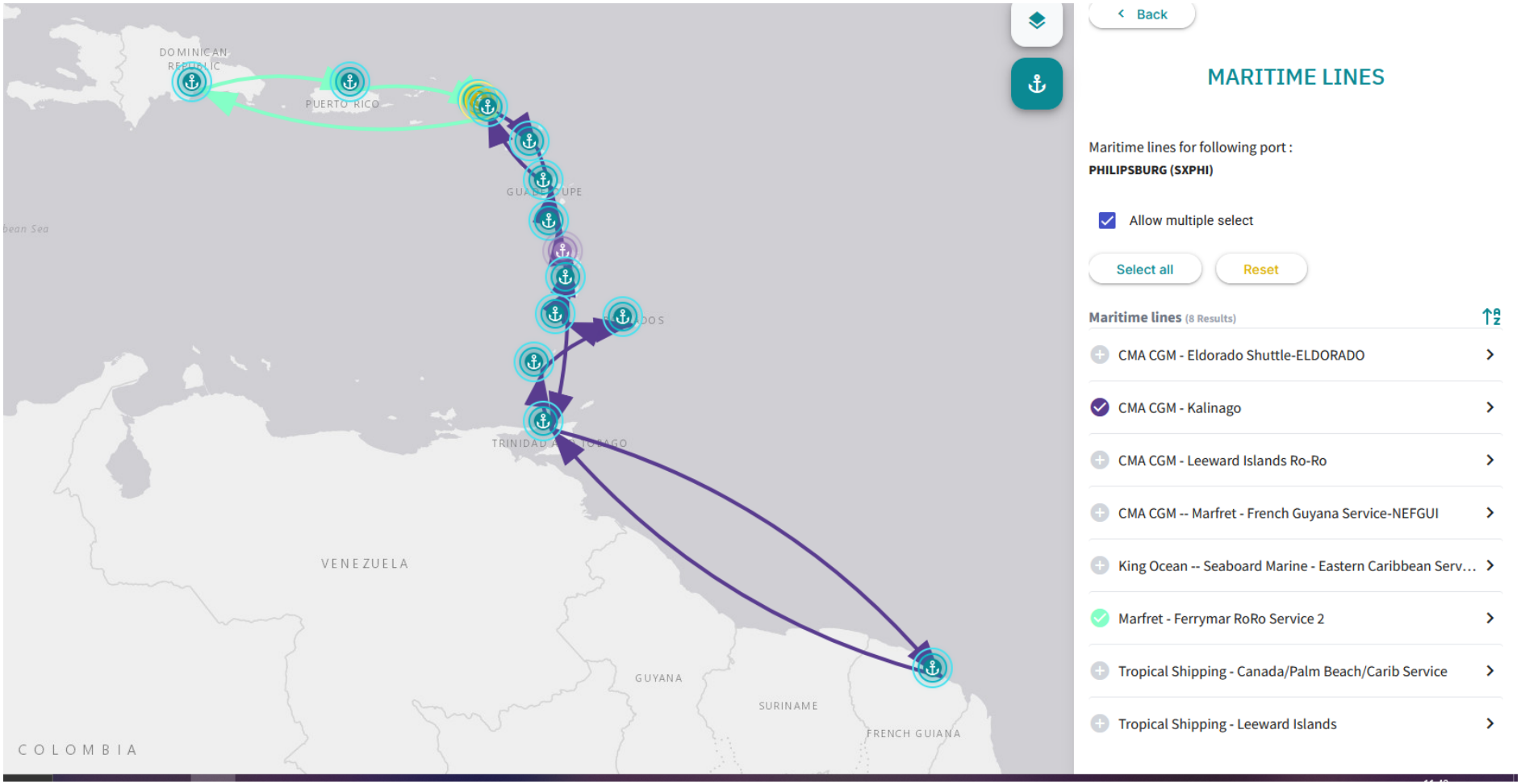


Figure – Lignes maritimes - (Fournit par Sinay)

Comment extraire ces LMR ?

Recherche de motifs séquentiels : Problématique

Défis de l'Extraction de Motifs Séquentiels

- Les motifs séquentiels étendent les règles d'association en ajoutant la dimension temporelle.
- Une **base séquentielle** contient des séquences ordonnées de transactions.
- L'intégration de la temporalité nécessite des calculs supplémentaires.

Motifs Séquentiels : Définitions Clés

Définitions Clés

- **Item** : Élément observé (ex. produit, événement)
- **Transaction** : Items achetés par un client à un instant donné
- **Itemset** : Ensemble d'items
- **Séquence** : Liste ordonnée d'itemsets
- **Support** : Pourcentage des clients qui supportent une séquence donnée
- **Motif Fréquent** : Séquence dont le support dépasse un seuil minimum

Motifs séquentiels : exemple

Exemple de motifs séquentiels

Après avoir défini les motifs séquentiels, prenons l'exemple du motif {B, I}, qui est un sous-ensemble récurrent dans les séquences, respectant l'ordre temporel. Sa fréquence dans la table des séquences est de 50%.

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

Figure – Motifs séquentiels

Nettoyage des données

Données de l'entreprise

Échantillion de données :


```
{"id": 228283
 "mmsi": 229338000
 "imo": 9625906
 "departure_port": "AUBTB"
 "arrival_port": "AUPOR",
"arrival_date":
"2022-07-16T13:27:32" ,...}
```


- **id** : Identifiant unique du voyage.
- **mmsi** : Numéro d'identification du navire.
- **imo** : Identifiant permanent du navire.
- **departure_port** : Nom du port de départ.
- **arrival_port** : Nom du port d'arrivée.

Nettoyage des données

- **Pré-traitement :**

Etapes de nettoyage

- Sélectionner les ports de départ et d'arrivée.
- Créer des séquences ordonnées par date d'arrivée.
- Exclure les anomalies :
  - Ports non renseignés (null).
  - Duplication succesive du même port dans les séquences.

Exemple de séquence avant nettoyage

$Séq : \langle AUBTB \rightarrow AUPOR \rightarrow AUPOR \rightarrow VNSGN \rightarrow NULL \ldots \rangle$

Exemple de séquence après nettoyage

$Séq : \langle AUBTB \rightarrow AUPOR \rightarrow VNSGN \ldots \rangle$

Adaptation

Adaptation du format de données :

1. **Dictionnaire de conversion :** Association des ports à des entiers uniques.
2. **Séparation des éléments :**
   - Ports séparés par la valeur -1.
   - Fin de séquence marquée par la valeur -2.

Exemple de séquence transformée

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

Post-traitement des motifs

Objectif du post-traitement

Assurer que seuls les motifs pertinents, susceptible d'être des **lignes maritimes régulières**, soient conservés après l'exécution de l'algorithme `PrefixSpan`.

1. **Exclusion des motifs de taille 1.**
2. **Suppression des ports consécutifs identiques.**
3. **Exclusion des motifs avec un port unique répété.**

Exemple de motif après post-traitement

Motif : $\langle AUPOR \rightarrow VNSGN \rightarrow CNE76 \rightarrow HKHKG \rangle$

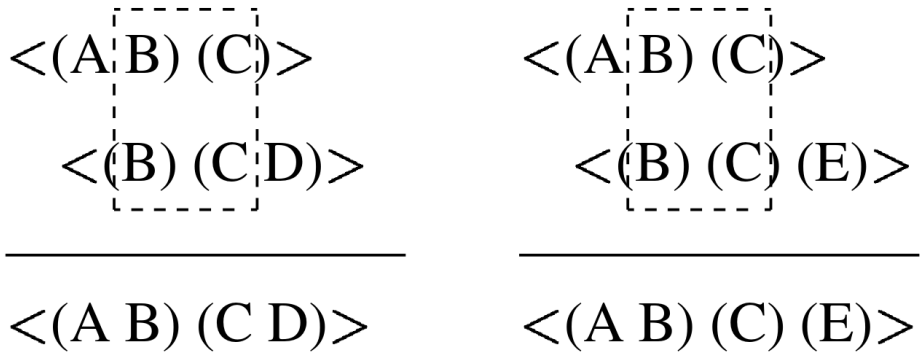
Algorithmes

Algorithme GSP

Principe

Extraction des motifs séquentiels fréquents via une approche itérative.

1. **Génération de candidats :**
   - Identification des motifs fréquents de taille 1 (*1-fréquents*).
   - Génération des séquences potentielles (*candidats*) en combinant les motifs fréquents précédents.
2. **Évaluation du support :**
   - Calcul du support pour chaque candidat.
   - Conservation des séquences atteignant un seuil minimal.



Algorithme PREFIXSPAN

Principe

Réduction de l'espace de recherche grâce à la projection de bases.

1. **Items fréquents :**
   - Identification des items fréquents (*1-fréquents*).
2. **Projection :**
   - Division de la base en sous-ensembles selon les préfixes fréquents.
   - Utilisation des suffixes pour réduire la recherche.
3. **Exploration récursive :**
   - Détection de nouveaux préfixes fréquents.
   - Extension des motifs jusqu'à épuisement.

Exemple illustratif de PREFIXSPAN

- Préfixe <a> : extraction des bases projetées (*suffixes*).
- Génération des motifs fréquents à partir des bases projetées.
- Support minimum : 50% (2 séquences)

Client	Séquence
10	< (a) (a b c) (a c) (d) (c f) >
20	< (a d) (c) (b c) (a e) >
30	< (e f) (a b) (d f) (c) (b) >
40	< (e) (g) (a f) (c) (b) (c) >

Préfixe	base projetée (suffixes)	motifs séquentiels
<a>	<(abc)(ac)(d)(cf)>, <(_d)(c)(bc)(ae)>, <(_b)(df)(c)(b)>, <(_f)(c)(b)(c)>	<a>, <(a)(a)>, <(a)(b)>, <(a)(bc)>, <(a)(bc)(a)>, <(a)(b)(a)>, <(a)(b)(c)>, <(ab)>, <(ab)(c)>, <(ab)(d)>, <(ab)(f)>, <(ab)(d)(c)>, <(a)(c)(a)>, <(a)(c)(b)>, <(a)(c)(c)>, <(a)(d)>, <(a)(d)(c)>, <(a)(f)>
<b>	<(_c)(ac)(d)(cf)>, <(_c)(ae)>, <(df)(c)(b)>, <c>	<b>, <(b)(a)>, <(b)(c)>, <(bc)>, <(bc)(a)>, <(b)(d)>, <(b)(d)(c)>, <(b)(f)>
⋮	⋮	⋮

Algorithme CLOSPAN

Principe

Extraction de motifs séquentiels fermés pour réduire la redondance.

1. **Ordre lexicographique :**
   - Exploration structurée en triant les motifs.
2. **Relations entre items :**
   - Identification des relations fixes (*"A précède toujours B"*).
3. **Optimisation :**
   - Basé sur PREFIXSPAN pour la projection des bases.
   - Réduction de la redondance en extrayant uniquement les motifs séquentiels fermés.

Expérimentations et analyse

Protocole expérimental

Algorithme choisi : `PrefixSpan`

Données utilisées : 1552 navires

Paramètre variable : Support minimum (`minsupp`) (1% à 20%)

Métriques observées :

- Temps d'exécution
- Nombre de motifs extraits
- Taille moyenne des motifs
- Écart-type de la taille des motifs

Expérimentations

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
2%	31	2021.33	25696	5.66	2.00
3%	47	768.38	4345	4.92	2.14
4%	62	569.71	1539	4.47	2.10
5%	78	504.18	741	4.29	2.04
6%	93	420.92	411	3.91	1.77
7%	109	437.76	225	3.29	1.39
8%	124	363.63	133	3.06	1.19
9%	140	357.14	74	2.59	0.80
10%	155	347.67	51	2.37	0.62
11%	171	341.53	29	2.28	0.52
12%	186	345.66	17	2.18	0.38
13%	202	337.82	8	2.00	0.00
14%	217	339.09	5	2.00	0.00
15%	233	329.06	1	2.00	0.00
16%	248	334.06	1	2.00	0.00
17%	264	335.52	0	nan	nan

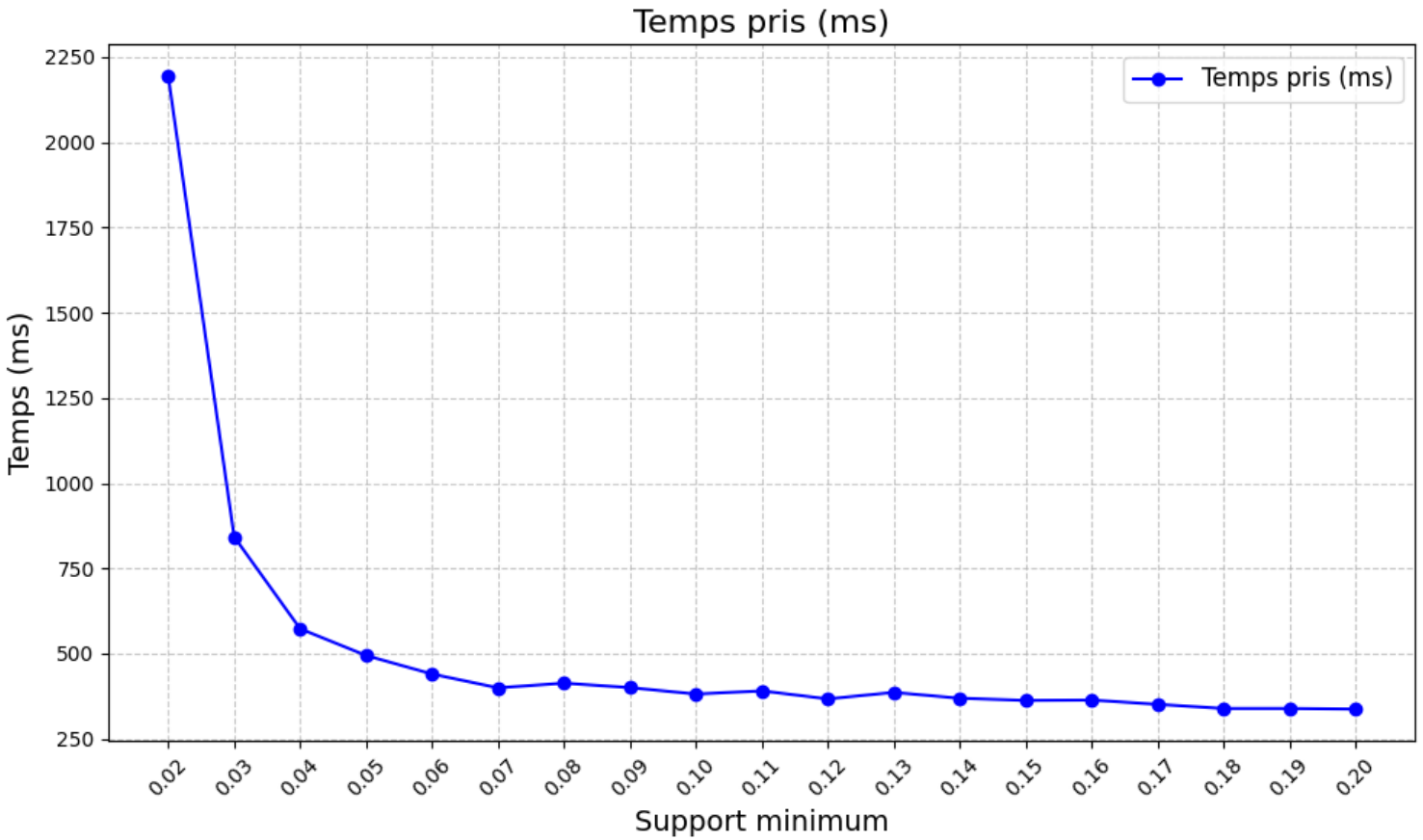
Figure – Résultats des expérimentations

Analyse des résultats expérimentaux

Tendances principales observées :

- **Impact du support minimum :**
  - Plus le seuil de support augmente, moins il y a de motifs.
  - *Exemple :* 6M motifs (support 0.01) $\rightarrow$ 0 motif (support 0.17).
- **Taille des motifs :**
  - Séquences longues moins fréquentes avec un support élevé.
- **Temps d'exécution :**
  - Support faible (0.01) : $>$ 131s
  - Support élevé ($>$ 0.10) : $\approx$ 350ms

Visualisation du temps d'exécution



Explication des motifs extraits

Un échantillion de motifs fréquents extraits :

- `PATBG → PACTB → PAROD → PACTB → NLMSV → PAROD #SUP : 16`
- `GBLGP → BEANR → NLMSV #SUP : 63`

Exemple d'interprétation

Le motif `GBLGP → BEANR → NLMSV` (#SUP : 63) signifie que 63 navires ont suivi ce trajet, passant par ces ports dans cet ordre.

NB

- Entre deux ports, des trajets intermédiaires sont possibles.
- Le trajet peut aussi être direct.

Conclusion