L'objectif de cet exercice est de faire une première analyse de ce jeu de données et d'envisager des pistes futures.

1 Préambule sur la forme

On verifie que le fichier/ set de données n'est pas trop corrompu, si il est sain. **df.isnull().values.any() = False** : sur la forme il l'est, pour le fond il faudrait plus de temps pour appréhender la partie métier.

Au passage on note que le **format des dates** ne correspond **pas au standard** d'un objet datetime. On corrige cela.

La plus grosse assomption que je fais c'est que les timestamp des évenements correspondent à départ et arrivée sur un tronçon ; ça ne m'a pas paru clair. Pour s'en assurer, on **vérifie** qu'il y a (presque ?) une **contiguïté chronologique** entre les **tronçons** d'une même route.

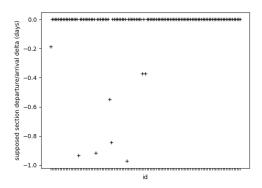


Figure 1: différence temporelle entre fin et début de tronçons consécutifs

On s'aperçoit qu'il y a quelques **incohérences** avec des deltas entre tronçons négatifs mais on considèrera que l'hypothèse est juste.

On note également que certains parcours ne sont pas complets dans le jeu de données. Ici on peut juste soupçonner qu'il a été tronqué.

Peut être opportun de transformer en catégories numériques certaines variables. Tout dépend comment est implémenté l'opérateur d'égalité pour un string. Peut être pour un fichier plus grand.

2 Quelques indicateurs en vrac

(fig4) pour constater la dynamique des temps trajets clients et un argument de plus pour prouver l'intérêt du service proposé. D'autant qu'il y a des passages obligés pour certains clients (fig3.a trajets début-fin pour mais on aurait pu decliner pour chaque tronçon et chaque client). Mais le jeu n'est pas proportionné ; fig2.a et fig3.b pour illustrer ici.

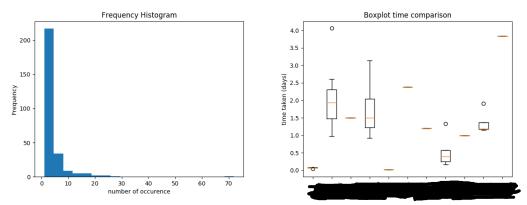
(fig3.b) peut aussi servir en interne à remarquer l'évolution des clients, si c'est bien **equilibré** / **pas** trop **exposé** et sa **dynamique** pour juger de la **confiance** qu'ils font à l'entreprise ; il vaut peut être mieux analyser les facturations plus explicites.

Pour (fig2.b), on pourrait distinger les tronçons intra et inter site. En tous les cas, cet indicateur est important pour les différents opérateurs Ils ont sûrement d'autres sources, mais ils leur manquent peut être la comparaison globalement, avec des plateformes similaires, ou leurs concurrents plus proches géographiquement.

Les api (au moins pour les bateaux) que j'ai trouvé étaient payantes et je n'ai pas fini d'arpenter les données ouvertes européennes mais c'est certainement possible de croiser cet indicateur avec d'autres des systèmes qui font le lien pour distribuer les fautes ou les qualités : l'opérateur du hub / le transporteur (ou encore anticiper des écarts) et ainsi permettre la construction d'un système de **notation** des différents **opérateurs** pour l'interne ou fournir à des tiers.

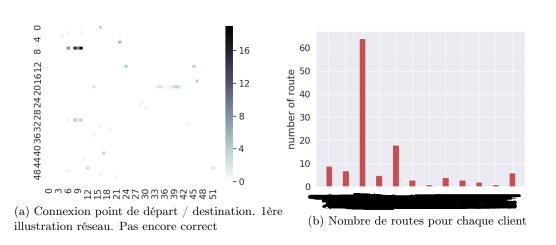
Si s'écarte trop, detection de défaillances du tracker (**fig2.b**) (=> nettoyer jeu de données) On peut imaginer des **fenêtres temporelles** pour suivre l'évolution de tous ces indicateurs.

Figure 2



 $\hbox{(a) Repartition nombre d'occurences par tronçon (b) Exemple de temps pris pour quelques tronçons}$

Figure 3



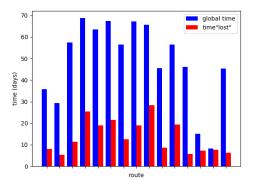
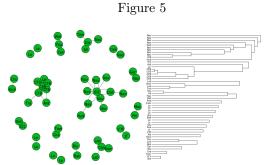


Figure 4: Quelques exemples de routes temps global / temps passé "inerte"

3 Dessine moi un réseau

C'est surtout là que le **champs des possibilités est large**. Et il faut bien des spécialistes notamment de la "supply chain" pour donner des pistes, il me semble que vous tirez parti de votre **hauteur de vue** ; les **trackers** accrochés **traversent différents systèmes** et on s'évite les tracas de différences de "culture" entre ceux ci.

Vous avez le choix de l'échelle alors que les différents **opérateurs** sont **cantonnés** à leur **environnement**. Bref c'est du pain béni pour les chercheurs entre autres. On peut analyser ces réseaux sous quantité d'angles différents, on **pare au plus simple ici** pour l'instant.



(a) Connexion point de(b) Distinction en comdépart / destination.munautés, noms trongénéré avec igraph qués.



Figure 6: Vue un peu plus artistique avec Gephi

Anticiper une déstabilisation du réseau : permettre quelques tronçons en avance de prévoir du retard ou même changer de parcours en profitant du passage d'autres conteneurs déjà passés. (suivre des conteneurs qui passent par le même tronçon qu'un autre qui y passera 1 ou plusieurs tronçons plus tard pour anticiper dans l'absolu et comparer avec les temps historiques mais le jeu de données n'en valait pas la chandelle ici).

Simuler une déstabilisation du réseau : un relai inacessible / cours circuité (le port de bayrouth qui explose) ; comment les reports se feraient et comment ça influe sur les chaines globales ou pour un trajet particulier (croiser avec d'autres données, établir des lois liant les différents indicateurs pour anticiper les effets) pour juger de la résilience globale (beaucoup d'analogies avec reseaux télécommunication/eau/énergie... beaucoup de recherche déjà produite) ou la resilience d'un client pour savoir les risques auquels il est exposé et dans quelle mesure il peut les contourner.

Intérêts analyse communautés, de leur dynamique : pour fournir des indicateurs macro économiques aux institutions financières/ décideurs locaux (juger du repli / de l'ouverture des régions) Eviter que les changements se fassent à l'aveugle (analogie avec l'automatique : boucle de retour du plus rapide et plus précise), d'autant qu'au moins en Europe des reports modaux s'annoncent ; ferroutage encouragé ou en tout cas vélléités, canal seine nord et gros projets d'infrastructure chinois, des grandes entreprises qui vont devoir tenir des comptes carbone.

Ceci dit si les **différents tronçons sont choisis/guidés avec un certain dessein** (optimiser le temps par exemple ou le bilan carbone avec d'autres intérêts propre au client qui s'y mêle) on entrevoit une **réalité biaisée** du réseau (dépend aussi de la couverture). On imagine facilement pour un **large réseau** que la **complexité** va exploser et on devra se poser la question du **choix** (conception?) **d'heuristiques** d'optimisation (couvrir la littérature sur les algorithmes d'optimisation quantique pour pareil réseau, ca plairait peut être au chercheur mais pas forcément à l'entrepreneur.)

PS: Si je devais tenter un peu d'exotisme : ça peut servir les intérêts de l'intelligence economique, on l'a dit comme de l'intelligence securitaire : analyser les dynamiques liées aux réseaux criminels et vérifier si il existe une corrélation avec la dynamique des réseaux d'échange de conteneurs. Je ne crois pas que ces données soient déjà publiques.

En tous les cas à mesure que j'y réfléchi, le **sujet m'interesse**. J'espère n'avoir pas trop divagué, avoir placé cet aperçu à la hauteur que vous attendiez. Et si d'aventures vous souhaiteriez continuer avec moi, je serais content de connaître les pistes que vous envisagez et de lire la littérature qui s'y rapporte avant de commencer.

qui m'a permis de partir en Chine est je crois proche de ce domaine de recherche. Il a essaimé puisque un de ses doctorants chinois enseigne au HIT (où j'étais donc) des fois que vous envisagiez nouer des liens en Chine. Top 10 université chinoises quand même, encore mieux si on ne considère que l'ingénierie. Un des centres intellectuels de la province du Dongbei qui a pris une large part dans le développement du chemin de fer chinois avec son industrie lourde. HIT possède des antennes à Weihai et Shenzhen villes portuaires, stratégique pour la dernière. Il y a bien une culture du génie industriel.