Processus d'import de la WAD (WorldWide Aviation Database)

Introduction

Ce document détaille comment transformer le processus actuel de génération de données Segments et O&D en un processus automatisable de génération / correction par import de données, qui intègre différentes sources (aviation civile, clients, etc.)

Processus Actuel :

- Données historiques Segments, O&Ds et Revenus de Sabre
- Données schedule OAG (vols + avions) → capacité
- Données actuelles générées selon historique + capacité
- Processus manuel
- Processus non reproductible
- Incohérences de calcul difficilement détectables

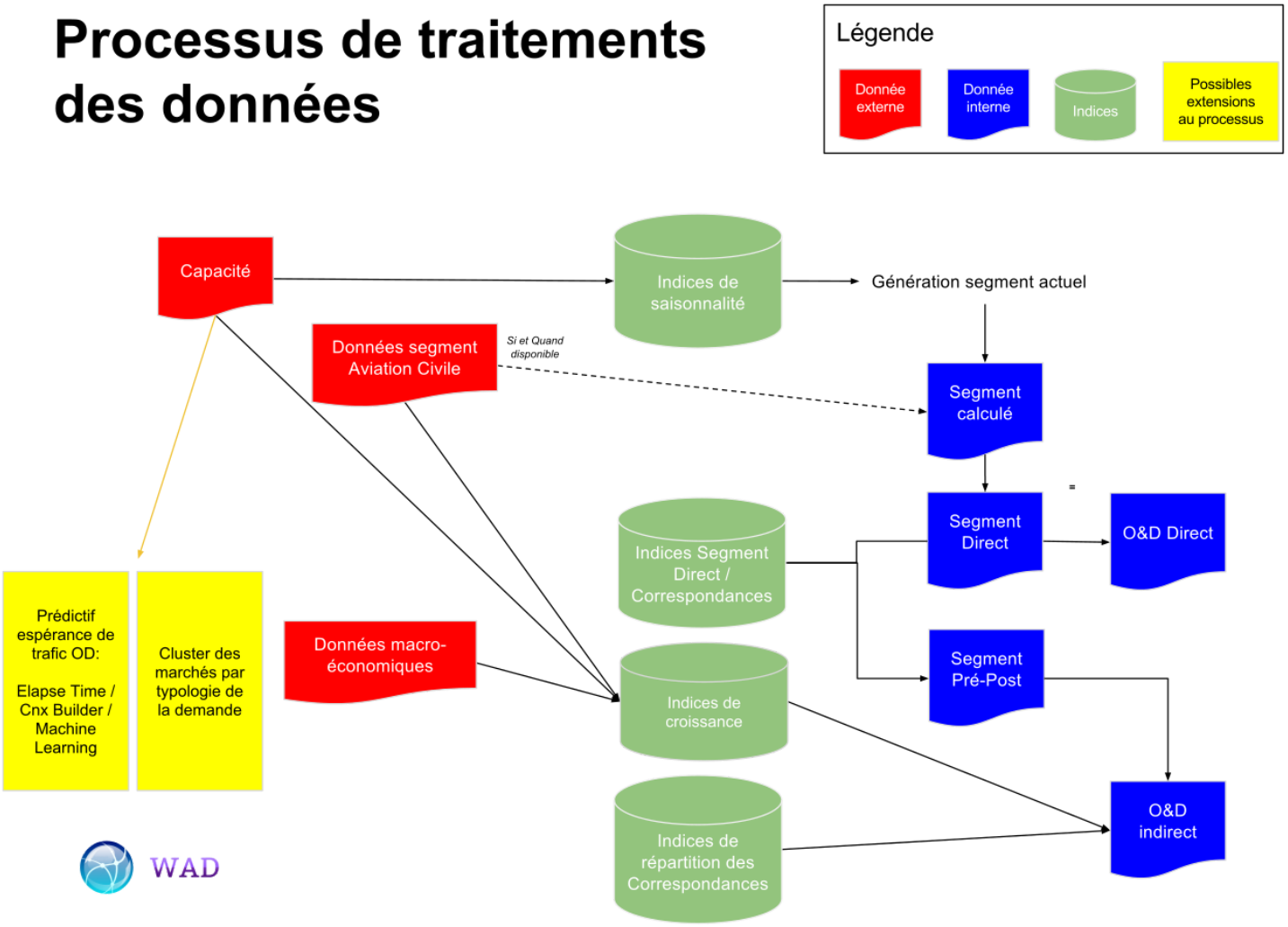
Objectif Final :

- Données de référence: OAG
- 100 % des données clients intégrées
- 100 % des données aviation civile détaillées disponibles intégrées
- Intégration de données générales (aviation civile non détaillée, offices de tourisme, indicateurs économiques, etc.)
- 85 % des traitements automatisés
- Résultats reproductibles
- Alertes automatiques sur cas particuliers à traiter manuellement et sur anomalies
- Indice de fiabilité des données permettant d'influencer l'intervalle de confiance sur les prédictions
- Amélioration progressive des algorithmes de prédiction par apprentissage

Laurent Berder

1. Schéma de processus

Processus de traitements des données



2. Indices

Plusieurs types d'indices servent à recalculer les données.

- Indice de croissance du trafic (d'une année sur l'autre): $moyenne(passagers\ annuels)$
- Indice de saisonnalité du trafic (d'un mois sur l'autre):
  $(pax(ym_{N-3}) / moyenne_{N-3} * 0.2) + (pax(ym_{N-2}) / moyenne_{N-2} * 0.3) + (pax(ym_{N-1}) / moyenne_{N-1} * 0.5)$
- Indice de saisonnalité du yield:

$(yield(ym_{N-3}) / moyenne_{N-3} * 0.2) + (yield(ym_{N-2}) / moyenne_{N-2} * 0.3) + (yield(ym_{N-1}) / moyenne_{N-1} * 0.5)$

- Indice de saisonnalité du load factor:

$(lf(ym_{N-3}) / moyenne_{N-3} * 0.2) + (lf(ym_{N-2}) / moyenne_{N-2} * 0.3) + (lf(ym_{N-1}) / moyenne_{N-1} * 0.5)$

- Indice Segment Direct / Correspondances
  - Par couple Origin/Destination
  - Formule: $pax_{direct} / somme(pax)$
- Indice de clef de répartition des correspondances
  - Par couple Origin/Destination
  - Formule: $pax_{direct} / somme(pax)$

Mis à part l'indice de croissance, tous les autres sont calculés sur une période de 3 ans, avec des poids différents selon les années:

- 50% pour l'année N-1
- 30% pour l'année N-2
- 20% pour l'année N-3

Un dernier indice doit être calculé pour qualifier la fiabilité de la source de donnée: l'indice de confiance, tel que montré dans le tableau ci-dessous (cf. Indices de confiance des sources) :

Source	Définitions	Périmètre	Aller / Retour	Régulier / Non-régulier	Year Month	Destination	Origine	Compagnie	Revenu	Pré / Post	Indice Confiance
India - intl	Non	Int	Séparés	Cumulés	Non	Nom de ville	Nom de ville	Non	Non	Non	11
India - domestic	Non	Nat	Séparés	Cumulés	Oui	Nom de ville	Nom de ville	Non	Non	Non	14
Ireland	Non	Int	Séparés	Cumulés	Oui	Nom de ville	Nom de ville	Non	Non	Non	14
Australia	Oui	Int	Séparés		Oui	Nom de ville	Nom de ville	Non	Non	Non	15
Mexico	Non	Nat + Int séparés	Séparés	Séparés	Oui	Nom de ville	Nom de ville	Non	Non	Non	16
Chile	Oui	Nat + Int séparés	Séparés		Oui	Nom de ville	Nom de ville	Nom airline	Non	Non	18
United Kingdom	Non	Nat + Int séparés	Cumulés		Oui	Nom aéroport	Nom aéroport	Non	Non	Non	19
United States	Oui	Nat + Int mélangés	Séparés		Oui	FAA/IATA	FAA/IATA	IATA	Non	Non	34
European Union	Oui	Nat + Int mélangés	Séparés	Séparés	Oui	ICAO	ICAO	Non	Non	Non	37
Brazil	Oui	Nat + Int mélangés	Séparés	Séparés	Oui	ICAO	ICAO	ICAO	Non	Non	39
Colombia	Oui	Nat + Int mélangés	Séparés		Oui	IATA	IATA	ICAO	Non	Non	45
Gesap - Aeroporto	Non	Nat + Int mélangés	Séparés	Cumulés	Oui	IATA	IATA	IATA	Non	Non	45
ZI – Aigle Azur	Non	Nat + Int mélangés	Séparés	Cumulés	Oui	IATA	IATA	IATA	Oui	Non	50

Le calcul de l'indice de confiance dépend de colonnes qui le précèdent (indicatives de la fiabilité de la donnée), et peut être adapté selon l'importance que l'on veut mettre sur certaines variables.

3. Informations 'Provider'

Dans notre base de données, il faut garder une table 'Provider' qui recense toutes les sources de données de segments.
Les champs de cette table sont:

- Provider (chaîne)
- import_process (booléen)
- latest_ym_available (chaîne: YYYY-MM)
- index:
    - ym_start (chaîne: le year_month à partir duquel l'indice de confiance est valable)
    - confidence (entier: l'indice de confiance calculé)

C'est en se basant sur cette table que l'on pourra automatiser le téléchargement des données (en comparant le latest_ym_available et le dernier year_month qui a été téléchargé).
Pour mettre à jour le latest_ym_available des providers, il faut lancer le programme: "latest_available_year_months_per_source.py"

4. Processus étape par étape :

- Génération des données
- Formatage des sources à importer
- Traitement des données
- Répartition segments et O&D

Etape 1 : Génération des données sur un Year Month

On conserve le processus actuel de génération de données. Les données sont générées pour toutes les routes en application des indices, avant d'importer des données d'aviation civile ou de clients.

Etape 2 : Formatage des sources à importer

Format :
Selon la source, une ligne par route ou cumul de routes, avec les champs suivants.

Champ	Format	Commentaire
Year_month	Liste	
Provider	Chaîne	
Origin	Liste	code ou « régions » (type ville) ?
Destination	Liste	code ou « régions » (type ville) ?
Airline	Liste	« * » pour « toutes » dans les cas où la compagnie n'est pas précisée
Airline_Ref_Code	Liste	« * » pour « toutes » dans les cas où la compagnie n'est pas précisée
Passagers	Entier	
Revenue	Entier	Exprimé en USD
Both_ways	Booléen	
Revenue	Entier	Si disponible
Frequency	Entier	Nombre de vols, si disponible
URL	Chaîne	

From_file	Chaîne	
From_line	Entier	
Raw_rec	Dictionnaire	Contient les informations complètes de la ligne du fichier
Inserted	Date	

Index :

- year_month + provider
- origin
- destination
- airline

Etape 3 : Traitement des données du Year Month

Processus :

- Lister les providers qui matchent le year_month
- Générer les lignes multiples en dé-listant year_month, origin, destination et airline
  - Pour les cas d'airline manquante (« * ») : remplir par toutes les airlines des schedules sur les routes concernées
  - Répartir la somme de passagers selon la capacité (cf. Chevauchement)
- Déterminer les périmètres qui se chevauchent entre les providers :
  - chevauchement direct : ne conserver que l'indice de confiance maximum
  - chevauchement indirect : répartir la masse (cf. Chevauchement)
- Ajouter un code de traitement pour identifier le processus qui a été fait. A terme, on utilisera ce code pour limiter les re-calculs sur les périmètres non impactés.

Format :

Une ligne de données par tuple leg_origin / leg_destination / trip_origin / trip_destination / airline / year_month / class, contenant un nombre de passagers et un revenue, stockées dans la table New_Segment_Initial_Data.

Etape 4 : Segments et O&D

Processus :

- Appliquer les indices pour mettre à jour les enregistrement dans les bases des Segments Direct, Segments Pré/Post, O&D Directs, O&D Indirects.
- Calculer les nouveaux indices pour les prochains mois.

5. Chevauchement

Selon les sources, le détail peut être faible, et pour une ligne de données, on peut avoir plusieurs tuples de origin/destination/airline, qu'il faut répartir.
De même, on peut avoir plusieurs sources externes qui contiennent les mêmes routes, et dont il faut répartir aussi les données.

Pour chaque route, on adopte la formule suivante pour répartir les données:

$$\frac{s}{r} * p$$

avec:

s : nombre de passagers dans la source dont le niveau de confiance est le plus élevé

r : somme des passagers sur toutes les routes concernées par le périmètre de la source

p : nombre de passagers actuels sur la route

Exemple:

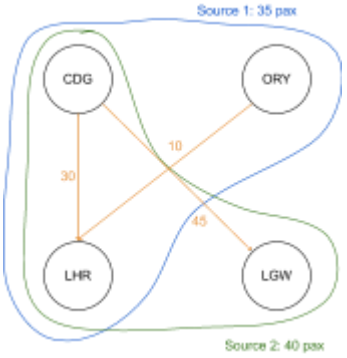
Prenons l'exemple des routes entre Paris et Londres

Sources externes:

Dans notre exemple, nous obtenons des infos sur ces routes de la part des aviations civiles française et anglaise

Source 1 : [CDG, ORY] -> [LHR]: 35 pax

Source 2 : [CDG] -> [LHR, LGW]: 40 pax



En base:

Dans optimode, nous avions déjà les données ci-dessous

CDG -> LHR: 30

CDG -> LGW: 45

ORY -> LHR: 10

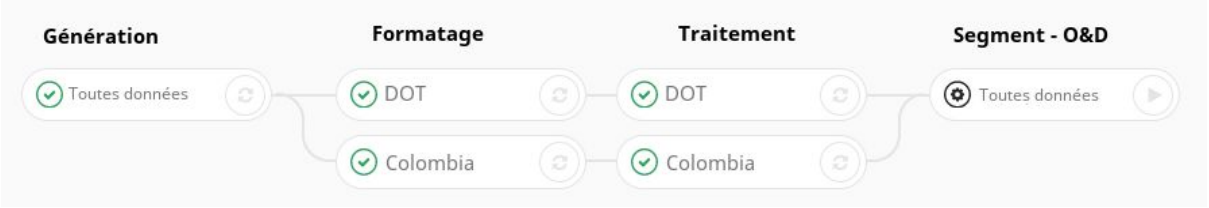
Répartition:

Le calcul de répartition se fait tel que décrit dans ce tableau

Route	*En base*	*Périmètre*	*Calcul*	*Répartition*
CDG -> LHR	*30*	*Source 1 + Source 2*	*Si confiance(source 2) > confiance(source 1):* $\frac{40}{(30+45)} * 30$ *Si confiance(source 1) > confiance(source 2):* $\frac{35}{(30+10)} * 30$	*16* *26 (26.25)*
CDG -> LGW	*45*	*Source 2*	$\frac{40}{(30+45)} * 45$	*24*
ORY -> LHR	*10*	*Source 2*	$\frac{35}{(30+10)} * 10$	*9 (8.75)*

6. Automatisation

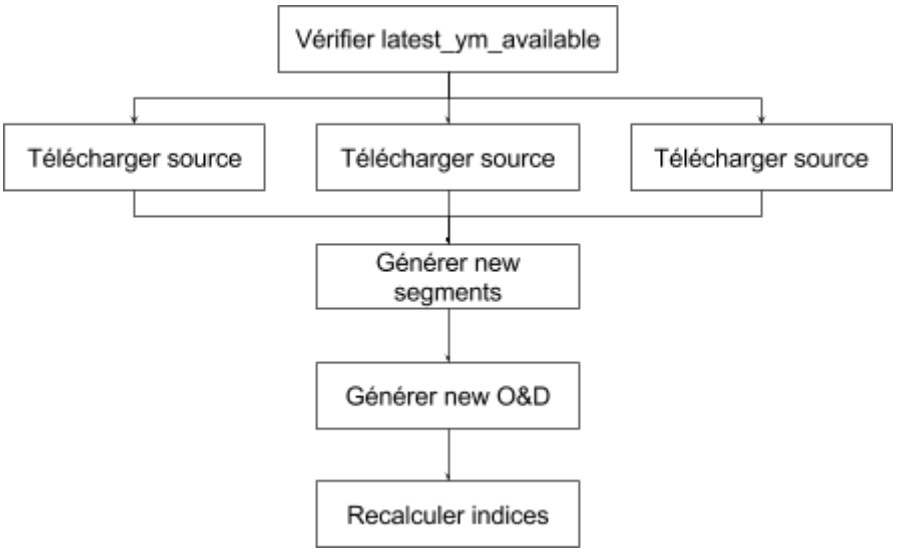
Le processus entier de génération, téléchargement, sauvegarde et traitement des données peut être automatisé par **Gitlab Pipeline**. Cela apporte l'avantage de permettre un suivi des données et même des versions d'algorithmes utilisées (pour améliorer la traçabilité et la reproductibilité), et de paralléliser certaines des étapes afin de réduire le temps d'exécution.



En cas d'échec d'un des composants du processus d'import/génération, on peut le reprendre là où il s'est arrêté.

En cliquant sur un des composants, on peut voir l'historique de tâche, et l'annuler ou refaire si besoin est.

Pour que le processus soit automatisé, il faut que le recueil des données externe le soit aussi. Il faut donc que l'on puisse détecter les données quand elles sont mises à disposition. Dans le cas des sites d'aviation civile, on peut sniffer leur site régulièrement pour repérer les fichiers qui n'existaient pas précédemment. Dans le cas des clients, les données sont déjà envoyées à une adresse email spécifique qui permet l'automatisation de l'import de nouveaux fichiers. L'idéal serait de demander aux clients une API, ou un moyen d'obtenir la donnée de manière automatique et régulière.



7. Alertes

Rapport d'alerte

Pour éviter de produire des données incohérentes, il faut définir des alertes (non bloquantes) pour chaque round de génération/altération de données.
Ces alertes peuvent porter sur :

- Pourcentage d'augmentation/baisse du nombre de passagers (en dehors de la saisonnalité habituelle) supérieur à 15%
- Valeurs extrêmes (revenu, yield à 0)
- Routes abandonnées (passées à 0 passagers) avec capacité existante
- Routes sans capacité avec traffic
- Déséquilibres allers-retours (qui ne se retrouvent pas d'un mois sur l'autre)
- Conflits entre sources externes supérieur au seuil
- Aéroports ou compagnies non-reconnues

Il faudra créer un rapport de fin de processus, sous forme visuelle, à base de drill-down par type d'alertes, et qui permette de prendre action pour chaque alerte (confirmer la donnée, ou revenir au précédent résultat de calcul si disponible). Ces actions rapides n'empêchent pas de pouvoir prendre des actions de correction plus poussées si besoin est.

MILANAMOS

Rapport de fin de processus

- Temps de processus: 8h43
- Nombre de sources importées: 8
- Nombre de routes modifiées: 2465
- Routes générées 84% / modifiées 14% / créées 2%

Alertes importantes
- 16 revenus à 0
- 12 yields à 0
- 2 sources contradictoires

Alertes subsidiaires
- 21 augmentations inhabituelles
- 413 déséquilibres A/R
- 24 routes abandonnées

Cf. Rapport de fin de processus d'import

Ce rapport devra être généré pour l'import de données, mais aussi avant ça au moment de la génération des données

Actions sur alerte

Chaque alerte n'ayant pas le même niveau d'importance, les actions à prendre ne sont pas les mêmes pour toutes.

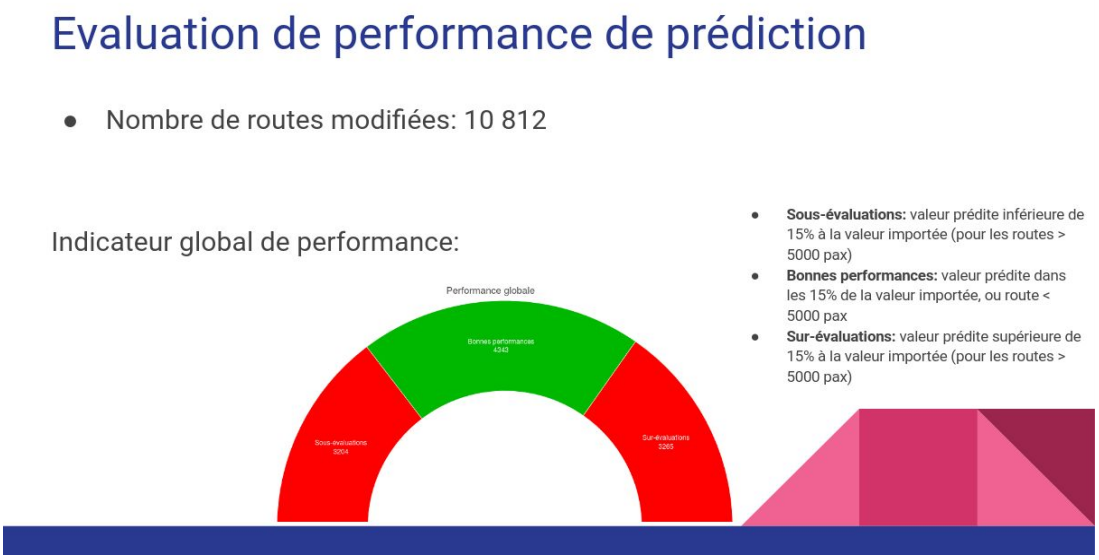
- Les alertes importantes devront être corrigées immédiatement.
- Les alertes subsidiaires devront être vérifiées, et éventuellement corrigées si nécessaire.
- Les aéroports et compagnies aériennes qui ne sont pas reconnues lors des imports sont stockées, et devront être revue trimestriellement (selon l'impact sur les données que leurs informations représentent)

8. Evaluation des performances

Lorsque nous récupérons de nouvelles données externes, elles sont intégrées en remplacement des données que nous avions générées pour ce year month et périmètre.
Il est intéressant d'évaluer la performance de l'algorithme de prédiction qui avait généré ces données en comparaison avec les valeurs réelles (de manière à améliorer l'algorithme si nécessaire).

Tous les mois, lorsque les données externes sont importées, il faut remplir un tableau dans lequel on stocke chaque route importée, avec comme colonne additionnelle la valeur remplacée.

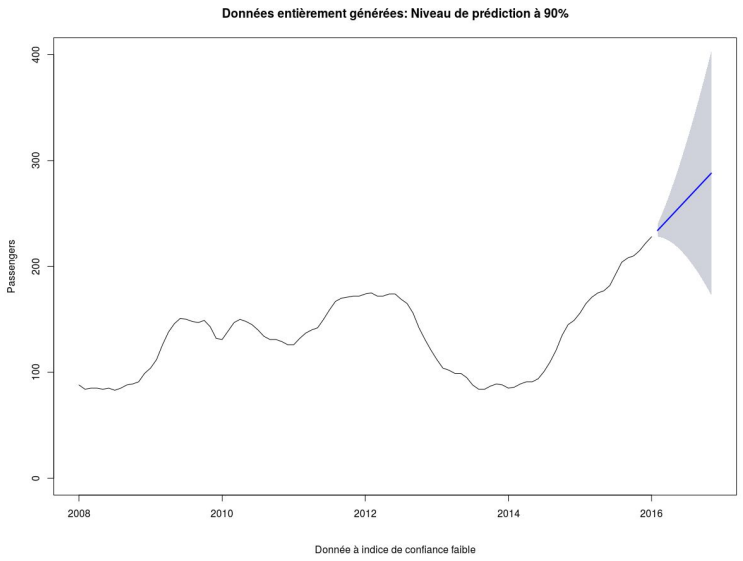
Un autre rapport en fin de processus permet de visualiser la synthèse de performances. Cela nous permettra, si nous faisons évoluer nos processus de génération de données, d'estimer si le nouveau processus est plus efficace que le précédent.

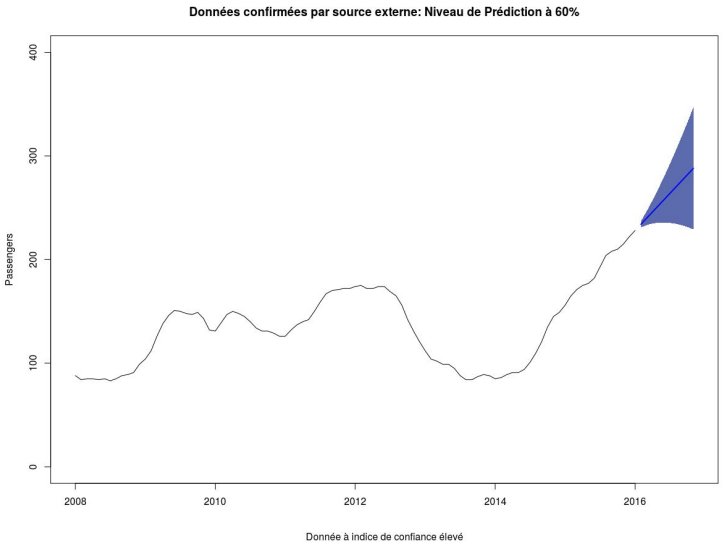


cf. Evaluation de performance de prédiction

9. Intervalles de prédiction

Plus une donnée est confirmée par des sources externes, et plus elle peut être considérée comme fiable. Comme en plus les sources sont chacune estimées par un indice de confiance, chaque information peut être qualifiée par un score.
Ce score (nul pour les cas de données générées non confirmée par une source externe) intervient alors dans les prédictions pour réduire/augmenter les intervalles de prédiction : plus le score est élevé, moins le niveau de prédiction exigé est important, ce qui réduit l'incertitude.





10. Données macro-économiques

Nous pouvons utiliser plusieurs sources afin de télécharger régulièrement des données macro-économiques pertinentes pour les prévisions et calculs d'indices.
Ces données doivent être détaillées par pays, et permettre un suivi annuel dans le passé.
Certaines sources offrent aussi des prédictions sur plusieurs années, ce qui permet d'utiliser ces données pour générer les forecasts à long terme.

Indicateurs à télécharger:

- PNB
- Population
- Population urbaine
- Vols effectués
- Voyageurs aériens
- Voyageurs par motif de voyage
- Taux de change

Indicateurs à calculer:

- Nombre moyen de passagers par vol: $somme_{pax} / somme_{vols}$
- Nombre moyen de passagers par population: $somme_{pax} / population$
- PNB par habitant: $PNB / population$
- Corrélation population/passagers:
  $covariance(population,\ somme_{pax}) / (écart\_type_{population} * écart\_type_{pax})$
- Corrélation population urbaine / passagers:
  $covariance(population\ urbaine,\ somme_{pax}) / (écart\_type_{population\ urbaine} * écart\_type_{pax})$
- Corrélation PNB / passagers:
  $covariance(PNB,\ somme_{pax}) / (écart\_type_{PNB} * écart\_type_{pax})$
- Taux passagers tourisme $somme_{pax\_tourisme} / somme_{pax}$
- Taux passagers business $somme_{pax\_business} / somme_{pax}$

- Corrélation évolution du taux de change/passagers:
  $covariance(taux_{paysA} - taux_{payxB},\ somme_{pax})$ / $(écart\_type_{taux_{paysA} - taux_{paysB}} * écart\_type_{pax})$

Sources:

- FMI: https://goo.gl/N6iJjn (forecasts N+4)
- Banque Mondiale: http://databank.worldbank.org/data/download/WDI_excel.zip (N-1)
- ONU statistiques frontières: http://data.un.org/Data.aspx?d=POP&f=tableCode:401 (N-1)

A développer...

Une fois que l'import de données externes spécifiques est pérennisé, on pourra procéder à l'intégration de données externes moins spécifiques, comme celles qui ne renseignent que sur les destinations sans les origines, les agrégations annuelles sans détail de year_month, les parts de marché de compagnies aériennes sans détail d'origine ou destination, données de tourisme…
Ces données ne permettent pas de retravailler de manière précise nos données, mais de donner des poids pour faire coller de plus en plus à la réalité notre connaissance du marché.