Analyse exploratoire

Contents

Introduction

L'objectif de ce chapitre est de présenter la structure de nos données, leur origine et en extraire les statistiques nécessaires à nos futurs choix de paramètres. La bonne connaissance de la nature et du comportement de nos données est essentielle pour la construction de notre infrastructure et pour la validation de notre approche. Elle passe par l'étude des techniques d'optimisation de prix appliquées par les différents sites marchands et par une analyse des effets de ces optimisations sur les courbes de prix.

Les changements de prix dans le domaine aérien suivent des règles régies par des algorithmes de *yield management* ou *revenue management* décrits dans de nombreux ouvrages [2][15] ou [46], laissant entrevoir que certaines de ces règles, communes à tous les marchands, peuvent être apprises pour prévoir leurs évolutions. Le Yield Management est une discipline économique adaptée à des secteurs où la tarification par segments de marché est pratiquée et combinée à une analyse statistique poussée. Cette pratique a pour objectif d'augmenter le revenu de la compagnie par siège disponible. Les paramètres déterminants dans l'optimisation des prix sont donc le taux de remplissage de l'avion et l'évolution de la demande mais d'autres variables peuvent introduire des subtilités dans la maximisation des revenus [51]. Ces informations n'étant pas publiques, nous pouvons uniquement percevoir ces variables cachées par le biais de l'évolution des séries temporelles et de la répartition par destination du trafic de liligo.com.

Nous rappelons que liligo.com est un moteur de recherche de voyages permettant aux utilisateurs, de comparer plus de 250 sites d'agences de voyages et compagnies aériennes. A chaque recherche utilisateur, toutes les informations de la page de résultats sont conservées en base de données représentant une source volumineuse d'informations à traiter. Il est donc nécessaire de faire des choix quant aux vols que nous souhaitons utiliser et quant à l'architecture de notre base de données. Notre base d'apprentissage devra représenter la majorité des comportements existants tout en conservant une taille raisonnable. L'architecture doit permettre de reconstruire les séries temporelles de prix, de comparer les mêmes vols proposés par des sites différents et d'accéder rapidement aux caractéristiques des vols.

Nous introduisons la notion de vol unique, qui décrit un trajet défini par des dates de départ et de retour, des aéroports de départ et d'arrivée ainsi que les codes des vols correspondants (AF5653 par ex.). Ces vols sont vendus par la ou les compagnie(s) qui les affrète(nt) (elles sont plusieurs en cas de vols à escales ou de partage de code) mais aussi la plupart du temps par des agences de voyages. Chaque vol unique possède donc une série temporelle de prix par site marchand. Il est de fait intéressant de constater que les évolutions de prix d'un même vol unique peuvent être similaires ou alors complètement différentes selon le site qui les vend.

La première partie de ce chapitre est consacrée à la description du phénomène à l'origine de ce projet : le yield management. Après en avoir expliqué les tenants et les aboutissants, nous détaillons une des méthodes utilisées par les compagnies aériennes. Dans un second temps, nous décrivons l'origine de nos données (recherches utilisateur, alertes mail) et l'architecture de données qui nous a semblé la plus proche de la réalité du transport aérien. Nous avons notamment pris soin de conserver le lien entre un billet vendu directement par la compagnie aérienne et indirectement par le biais d'une agence de voyages. Nous illustrons ce phénomène

par des exemples, et nous détaillons les différentes approches d'optimisation des prix pratiquées par chacune des parties.

Puis nous expliquons le choix de nos paramètres de construction de la base d'apprentissage par l'analyse statistique du comportement des utilisateurs dans l'achat de leurs billets d'avion. Nous y discutons de la sélection des routes étudiées, de la durée de séjour et de la longueur des séries temporelles ainsi que de leur quantité et de leur qualité.

Enfin nous expliquons les enjeux du service et la pertinence d'un conseil à l'achat du point de vue de l'utilisateur. Si le lieu commun est que les prix ne font qu'augmenter et qu'il faut acheter le plus tôt possible pour avoir le meilleur prix, nous démontrons qu'avec notre service il est possible d'économiser de l'argent quelque soit le jour avant la date de départ. Nous rappelons que notre service ne consiste pas à indiquer le meilleur moment pour acheter son billet mais à fournir à l'utilisateur une indication sur la future évolution de son billet à un instant t.

1.1 Notations

n	:	Nombre de séries temporelles dans la base de données.
i	:	Numéro du vol de la base d'apprentissage $i \in 1, ..., n$
V_i	:	Vecteur d'attributs du vol i.
$p_i(t)$	:	Courbe de prix du vol i.

1.2 Yield Management

Au sein d'une même cabine, la compagnie divise son avion en classes de réservation, ou classes tarifaires, ou encore classes de yield. C'est un découpage purement informatique, invisible pour le passager, et sans conséquence sur le positionnement des voyageurs à l'avant ou l'arrière de l'appareil.

Il ne faut pas confondre ce découpage avec le découpage en classes de transport, que sont la première classe, la classe affaires et la classe économique. Les classes de réservation sont des sous-divisions de l'avion au sein même de ces cabines. Toutes ces classes sont emboîtées à la manière de poupées russes, de la classe la plus basse à la classe la plus haute. Chaque vol est décomposé en 10 à 20 classes de réservations. Elles sont désignées par des lettres de l'alphabet. En général, la première classe de transport contient les classes tarifaires P et F, la classe affaires contient les classes tarifaires J et C, et la classe économique contient le plus de classes tarifaires, dont la Y. L'IATA [1] recommande une certaine codification, mais chaque compagnie a ses propres habitudes.

Ces classes sont emboîtées, au sens où une classe inférieure ne peut pas empiéter sur une classe supérieure, alors qu'une classe tarifaire supérieure peut préempter des sièges prévus pour

[1] Association internationale du transport aérien

une classe inférieure. Les compagnies low-cost[2] appliquent pour la plupart les mêmes principes, mais d'une manière fortement simplifiée. Ainsi, le prix de leurs billets ne varie généralement que suivant deux facteurs : l'achat à l'avance, et l'état de remplissage de l'avion. À un instant donné, il n'existe qu'un seul prix pour le billet d'avion, valable pour tout le monde. Ce système est bien adapté à la clientèle plus homogène (essentiellement loisir) de ces compagnies, et présente également l'avantage d'être bien compris par les passagers, car il se résume par la formule simple "plus on achète tôt, moins c'est cher". Dans la réalité, ce principe est infirmé quotidiennement par les algorithmes de revenue management qui doivent baisser les prix dans diverses situations : annulation de billet, augmentation de la taille de l'avion, retour de places allouées aux agences de voyages, etc. Nous allons d'ailleurs montrer dans la section "Pertinence"(1.5) de ce chapitre que nombre de lieux communs ne sont pas toujours vérifiés.

Les algorithmes utilisés ont pour but essentiel de déterminer quelles classes de réservation seront ouvertes sur un vol, avec quel quota de sièges affecté à chacune. Il s'agit d'un contrôle de l'offre par ajustement des capacités disponibles. Par exemple, il faudra ouvrir beaucoup de sièges dans les basses classes de réservation et n'en garder que peu pour les passagers à haute contribution sur un vol en heure creuse, qui sinon ne sera pas rempli, alors que sur un vol en heures pleines il s'agira de procéder à l'inverse pour obtenir le revenu maximal.

Exemple d'algorithme : Bid-Price Une des méthodes utilisées dans le domaine aérien pour maximiser les revenus est l'optimisation d'un vecteur représentant l'évolution du prix selon le remplissage de l'avion. Ces vecteurs nommés "bid-price vectors" sont des indications de modification de prix par cabine envoyées aux GDS[3] afin qu'ils ajustent les prix annoncés au fur et à mesure du remplissage. Chaque cabine est divisée en classes associées à un tarif. Toutes les classes ayant un tarif inférieur au bid-price seront alors fermées à la vente.

La création de ce vecteur se fait en plusieurs étapes et nécessite un certain nombre de paramètres d'entrée :

1. Les évolutions passées des demandes par cabine
2. Le type d'appareil permettant de connaître la capacité par cabine (première classe, business, économique)
3. Les divisions passées des cabines en classes : nombre de classes, capacité et tarifs de chaque classe
4. L'historique des présences des passagers (Entre 15 et 20% d'absence en moyenne)
5. L'historique des sièges alloués aux agences de voyages/brokers non vendus
6. Le seuil à partir duquel il est préférable de surclasser un passager plutôt que de lui changer son vol

[2]Une compagnie aérienne à bas prix ou compagnie aérienne low cost, est une compagnie aérienne qui s'est positionnée sur le créneau commercial du transport aérien à moindre coût (low cost) en limitant ou en supprimant les services annexes au sol et en vol.

[3]Un système informatisé de réservation, ou global distribution system (GDS) en anglais, est un système informatisé qui centralise les données concernant les vols, les horaires, les places disponibles, les tarifs et les services connexes avec des moyens permettant d'effectuer des réservations et de délivrer des billets.

Avec tous ces paramètres provenant de bases de données d'historiques, de nouvelles données sont calculées :

1. Le placement du rideau séparant la première de l'économique (pour les plus petits avions)
2. Le choix du nombre de sièges dépassant la capacité de l'appareil (surbooking)
3. La division des cabines en classes (capacité, tarif)
4. La prédiction de la demande sur l'ensemble des jours avant la date de départ

A partir de toutes ces données de sortie, il est alors possible de construire de manière certaine par des algorithmes d'optimisation, un vecteur de bid price optimal qui sera ensuite transmis au GDS pour qu'il puisse appliquer les règles de ventes de tous les vols en question (Figure 1.1).

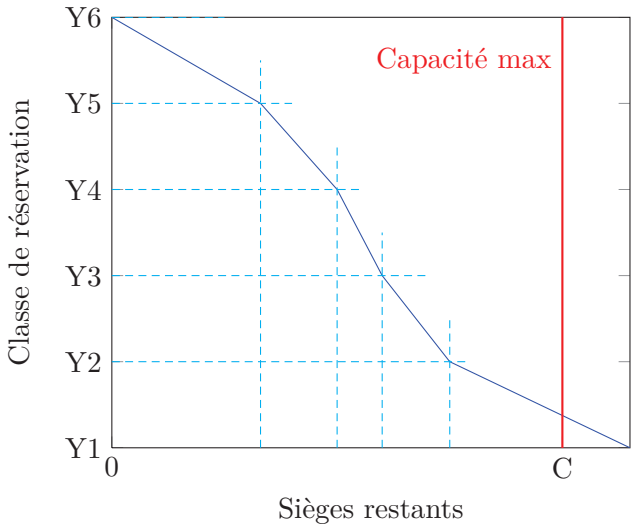

FIGURE 1.1 – Exemple d'évolution de prix en fonction du Bid Price

La modification des prix des classes peut se faire manuellement suite à la détection d'une trop grande différence entre la demande actuelle et la prévision, ou pour coller aux offres de la concurrence. Dans la majorité des cas ce processus d'optimisation est quotidien pour les vols dont la date de départ est proche et moins fréquent à mesure que l'on s'éloigne de celle-ci. La fluctuation des prix aura donc un comportement différent au fur et à mesure du temps et du remplissage. Nous allons voir dans la section suivante quelles sont les conséquences du yield management sur le comportement des séries de prix.

1.3 Structure et description de nos données

Dans cette section nous décrivons l'origine et les propriétés de nos données, et la manière dont elles ont influencé la structure de notre base de données. Cette dernière est composée de

trois tables principales dont chacune aura pour but de faciliter une des étapes de notre travail préparatoire : l'identification d'un vol unique, l'extraction de la ou des série(s) temporelle(s) $p_i(t)$ associée(s) et enfin la création des vecteurs d'attributs V_i correspondants. Ces tables respectent à la fois la notion d'unicité d'un vol et celle d'unicité du billet à vendre.

1.3.1 Description des données

Liligo.com possède une base de données d'historiques de recherche des utilisateurs grâce à laquelle nous avons construit notre base d'apprentissage et de test. Dans cette base de données, un *trajet unique* ou *vol unique*, correspondant à un couple vol aller/vol retour, est défini par 6 attributs : le couple "aéroport de départ" et "aéroport d'arrivée" que nous nommerons "route", la date de départ et la date de retour comprenant les heures et les minutes et enfin, le code transporteur du vol aller et celui du vol retour[4]. Les prix sont collectés depuis des sites marchands tels que les compagnies régulières (*Air France*, *EasyJet*, etc.) et les agences de voyages (*GoVoyages*, *VoyagesSNCF*, etc.) ces dernières pouvant vendre les mêmes vols que les compagnies régulières. Cela implique que pour un même trajet, on peut avoir autant de séries temporelles que de sites proposant le vol. Nous nommerons "*provider*" le site marchand d'où le prix est extrait et "*supplier*" la compagnie aérienne qui affrète l'avion. Si le prix est issu du site de la compagnie aérienne, alors le code *provider* sera identique au code *supplier*. Inversement, si les prix proviennent d'un site tiers tel *GoVoyages* ou *Opodo*, le code *provider* ne sera pas celui du *supplier*. L'association d'un trajet unique et d'un site marchand correspond alors à un résultat de recherche, chaque recherche renvoyant une multitude de résultats.

Les origines de nos données sont diverses : tout d'abord lorsqu'un utilisateur se rend sur le site liligo.com, il lance une recherche avec comme paramètre une ville ou un aéroport de départ et d'arrivée ainsi qu'une date de départ et de retour sans préciser les horaires. Chaque recherche utilisateur renvoie alors en moyenne 300 résultats ayant chacun des paramètres supplémentaires mais différents : le site marchand (*provider*), la compagnie aérienne (*supplier*), les horaires et le tarif. La notion de vol unique est déjà présente sur le site car les vols proposés par plusieurs sites marchands sont regroupés afin de comparer rapidement les offres similaires et d'alléger l'affichage. Cet important volume d'informations doit être correctement stocké et indexé afin de pouvoir y accéder facilement.

Par ailleurs, chaque utilisateur peut définir des alertes : l'alerte est un outil programmable qui effectue une recherche demandée, quotidiennement, à la place de l'internaute. Après avoir renseigné la destination et les dates de vol, le voyageur reçoit chaque jour par email un récapitulatif des meilleurs résultats, classés par prix. Ces alertes peuvent nous donner de longues séries de prix régulièrement échantillonnées mais ne garantissent pas de pouvoir suivre un même vol pendant toute la durée des envois de mails. En effet, seules les cinq meilleures offres sont sélectionnées réduisant les chances d'avoir pendant un mois la même offre dans le top cinq. Cette limitation ayant été mise en lumière, nous pensons faire évoluer le système des alertes pour conserver l'intégralité des résultats. Cependant les alertes peuvent aussi être stoppées à n'importe quel moment par l'utilisateur. Malgré tout, cette source de données nous donne un

[4] ex : AF350 pour un vol sans escale et AF630AF405 pour un vol avec une escale.

grand nombre de séries consistantes. Nous définissons une série consistante comme une série possédant suffisamment de points pour garantir la détection de la majorité des variations de prix.

Les données issues des recherches utilisateurs, sont quant à elles plus éparses lorsque la date de départ est éloignée, et ne nous garantissent pas d'avoir toujours une collecte quotidienne. Heureusement, nous verrons dans le chapitre suivant que la fréquence des changements de prix est en grande majorité supérieure à 24 heures autorisant un échantillonage plus faible. Ces changements de prix par à-coup créent des séries de prix constantes par morceaux, comme le montre la Figure 1.2 avec un faible nombre de plateaux. Pour pouvoir garder des ensembles de séries temporelles consistantes et uniformément échantillonnées, nous divisons notre base de données en 2 sous-ensembles :

1. Un ensemble constitué de séries de 28 jours échantillonnées toutes les 6 heures
2. Un ensemble constitué de séries de 90 jours échantillonnées quotidiennement

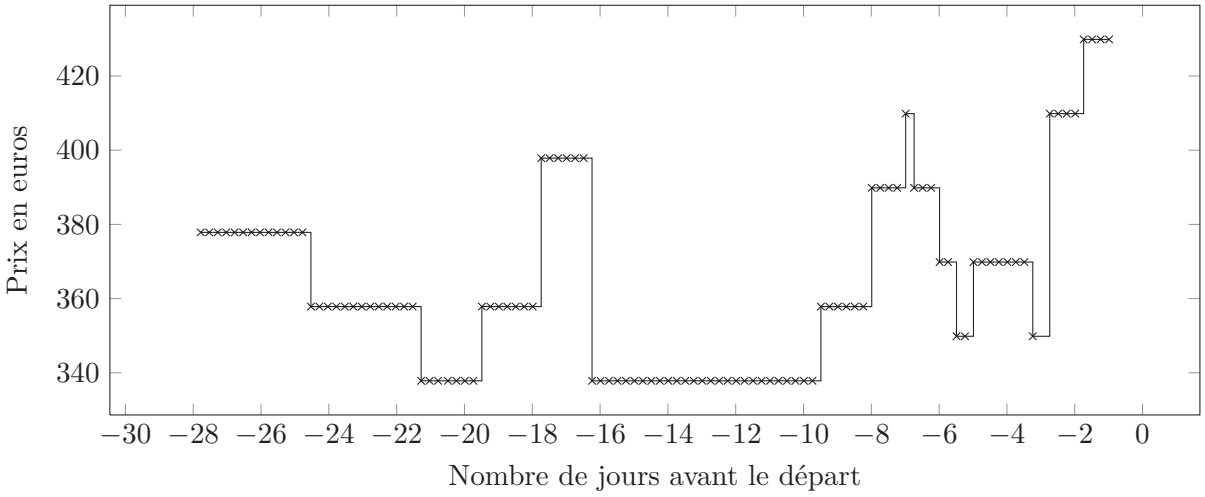

FIGURE 1.2 – Exemple de série de prix pour un Amsterdam-Barcelone du 20/11/2011 au 23/11/2011 proposé par *Austrian Airlines*. Cette série possède un point toutes les 6 heures pendant les 28 derniers jours de sa vente.

Nous allons maintenant décrire les trois grandes tables de notre base de données qui nous permettent d'extraire les séries de prix et leurs caractéristiques.

1.3.2 Structure de la base de données

Comme expliqué précédemment, nous avons choisi de diviser notre base de données en trois tables principales, nous permettant chacune d'extraire les informations nécessaires à différentes étapes de notre processus de prédiction. Une première table stockera les paramètres des trajets

uniques et leur associera un identifiant unique. La seconde collectera les prix pour chaque couple vol unique/provider. Enfin une table contiendra l'ensemble des paramètres associés à une série de prix, tels que le type de compagnie, la ville, le pays et le continent de l'aéroport de départ et d'arrivée etc.

Les vols uniques

Dans le domaine du tourisme aérien, un vol opéré par une compagnie aérienne est défini par un aéroport de départ (*departureStation*) et un aéroport d'arrivée (*arrivalStation*) (couple nommé "route"), un horaire de départ (jour, mois, année, heure, minute), et un horaire de retour, un code transporteur aller (*transportCode*) et un code transporteur retour (*transportCodeRet*). La règle de nommage des codes transporteur est opaque mais les codes correspondent souvent au couple d'aéroports et à un horaire variant selon les périodes de l'année. Par exemple le vol 'HV3014' Paris-Marrakech opéré par *Transavia* (Low Cost d'*Air France*) correspond au vol de 6h40 de mars à juin et de 7h00 de juillet à février.

Ce code est systématiquement composé d'un premier groupe de lettre correspondant au code IATA de la compagnie aérienne opérant le vol et est suivi d'un groupe de chiffre choisi par la compagnie. La Figure 1.3 nous montre 3 vols ayant les mêmes codes transporteur aller et retour à différentes dates et nous montre qu'il n'y a aucune corrélation apparente entre les variations de prix, car comme expliqué dans l'exemple de yield management, le paramètre principal de l'évolution des prix est la courbe de demande prédite, différente en fonction du jour de départ. On peut cependant noter une similarité dans l'ordre de prix des vols décollant à la même période : 17 et 21 février 2012.

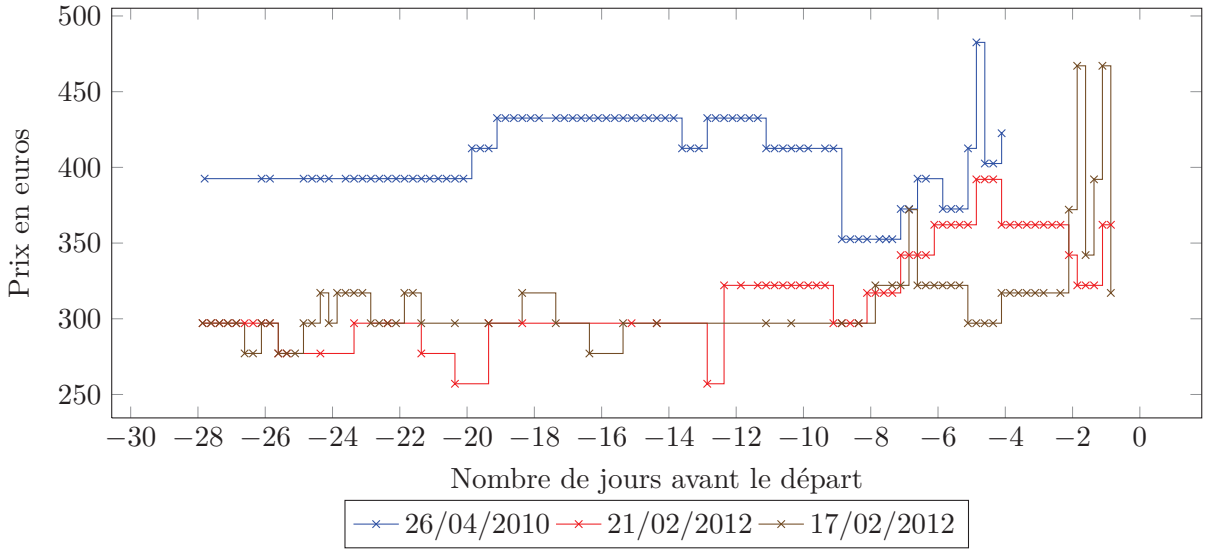

FIGURE 1.3 – Séries temporelles d'un même couple "code transporteur aller" (AT749) et "code transporteur retour" (AT640) (Paris-Marrakech sur *Royal Air Maroc* vendu par *VoyagesSNCF*) à différents moments de l'année.

En ce qui concerne les trajets à escales, nous avons décidé de concaténer les codes transporteur comme pour 'LH4271LH3472' par exemple, qui est un Paris-Budapest opéré par Lufthansa et faisant une escale à Düsseldorf. Pour des problèmes de maintenance de base de données et de volumétrie, nous ne conservons malheureusement aucune information sur les escales. Cependant, dans notre exemple, il est possible d'interroger notre base de données en cherchant le vol "LH4271" afin de retrouver l'aéroport de l'escale. Si le vol n'existe pas, des sites spécialisés peuvent nous fournir cette information, mais elle ne nous est pour l'instant pas nécessaire.

Nous définissons alors un vol unique par le n-uplet suivant : { aéroport de départ, aéroport d'arrivée, date de départ, date de retour, code transporteur aller, code transporteur retour} en excluant la notion de site marchand : ce vol unique représente le trajet indépendamment du vendeur. A chaque n-uplet nous associons un identifiant unique nommé *id_unique_flight* .

La création d'un identifiant unique indépendant du site marchand va nous permettre d'observer des phénomènes de concurrence ou de détecter les changements de prix supplémentaires appliqués par les agences de voyages.

En revanche, dans certains cas, la route et les dates sont identiques mais les codes transporteur différents : nous avons affaire à un vol cobrandé (aussi appelé “partage de code”). Il arrive que certaines compagnies s'accordent à vendre en parallèle des places pour un même vol, chacune sous sa propre marque. Dans ce cas, toutes les caractéristiques du vol sont les mêmes excepté le code du vol (AF312 et IB415 par exemple). Ces accords commerciaux permettent d'augmenter la visibilité des deux compagnies et d'assurer une meilleure rentabilité [10].

Trois types de partage de code existent :

- **En blocs :** la compagnie X achète à la compagnie Y un nombre de sièges fixé à l'avance sur un vol donné, à un prix également fixé à l'avance. Chaque compagnie vend les sièges qui lui reviennent en suivant sa propre politique tarifaire, et au stade de la réservation tout se passe comme si les deux blocs étaient en fait deux avions différents les deux partenaires ne vendent pas forcément leurs billets respectifs au même tarif, car leur clientèle n'est pas la même.

- **Free flow :** dans ce modèle, la répartition des sièges n'est pas fixée à l'avance : la compagnie qui opère l'avion gère un inventaire commun de places, réparti en classes de réservation, et dans lequel les deux compagnies vont puiser. L'autre compagnie paie un prix par billet, qui dépend de la classe de réservation. Il suppose donc une coopération plus poussée, et une certaine harmonisation des conditions tarifaires, si ce n'est des tarifs.

- **Joint venture :** les deux compagnies peuvent décider de mettre en commun toutes leurs ressources sur une ligne ou un ensemble de lignes, et de partager coûts et recettes. La coopération tarifaire est alors totale, et du point de vue tarifaire elle est équivalente à une fusion. Ce type de partage a été mis en place par *Air France* et *Alitalia* sur les lignes reliant la France et l'Italie.

Selon les cas ces vols peuvent être considérés comme des vols indépendants ayant leur propre évolution, ou comme des vols similaires. Dans les faits, nous pensons que chaque compagnie applique son propre yield management et qu'il faut séparer le comportement des deux séries

temporelles. Les séries seront susceptibles de se terminer plus rapidement car chaque compagnie possède moins de sièges et le vol sera plus rapidement complet. Dans l'exemple de la Figure 1.4, on observe deux séries de prix d'un vol à escales dont le deuxième tronçon est cobrandé. Il s'agit d'un vol Amsterdam-Barcelone passant par Londres et vendu par *LastMinute*. Le Londres-Barcelone est à la fois visible sous le code "IB7453" (vol *Iberia*) et sous "BA486" (vol *British Airways*). Quelques recherches nous indiquent que le vol est opéré par *British Airways* mais il nous est impossible de savoir qui opère le vol à partir des informations récoltées sur les sites marchands. Les politiques de prix des deux compagnies sont complètement différentes et on observe des variations beaucoup plus fréquentes de la part d'*Iberia*, alors même que les prix pour un tel trajet sont déjà très élevés. En revanche il est intéressant de remarquer que les plateaux sont quasiment les mêmes, et donc que le mécanisme de changement de prix, comme expliqué précédemment est indépendant de l'étape de définition des prix par classe de réservation.

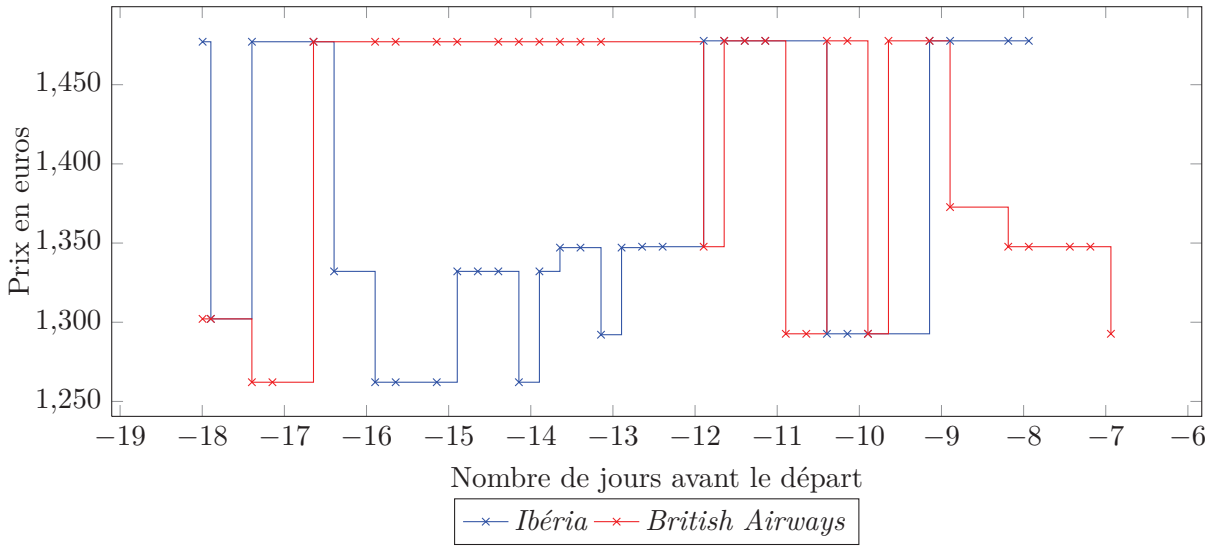

FIGURE 1.4 – Exemple de vols cobrandés proposés par *Ibéria* et *British Airways* (Amsterdam-Barcelone)

Les séries de prix

Chaque vol est ensuite vendu indépendamment, ou non, par un ou plusieurs sites marchands que nous appelons "*provider*". Le couple {*id_unique_flight* , *provider*} (vol unique, site marchand) possède lui aussi un identifiant unique nommé *id_flight* . L'ensemble des prix proposés sous l'identifiant *id_flight* constituera une série temporelle de prix uniques que nous stockons dans une table dédiée. Chaque ligne de cette table correspond à un vol unique proposé par un marchand, à un instant t avec un prix p créant ainsi la série de prix $p_i(t)$. Lorsqu'une recherche utilisateur est effectuée, nous ajoutons donc pour chaque résultat une entrée dans cette table.

La Figure 1.5 nous montre l'utilité d'avoir créé un identifiant différent pour un vol et pour les billets associés : on y observe pour un même *id_unique_flight* , les différentes séries temporelles

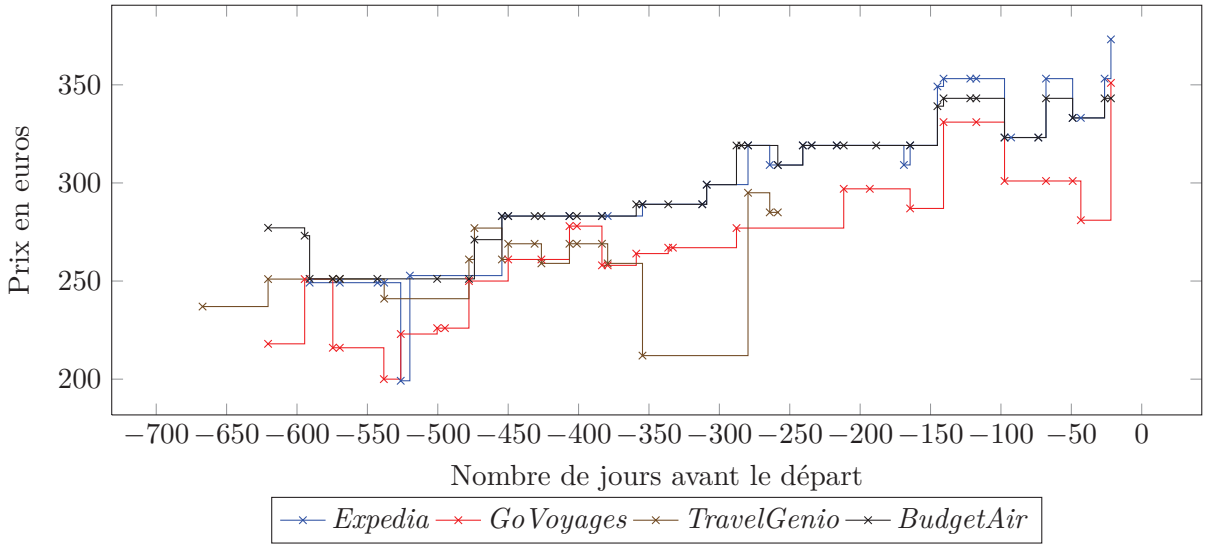

FIGURE 1.5 – Yield Management de différentes agences de voyages pour un même billet d'avion. L'*id_unique_flight* est identique mais les *id_flight* sont différents. Paris-Marrakech opéré par *Royal Air Maroc* du 26/02/2013 au 04/03/2013.

identifiées par un *id_flight* . Le même vol Paris-Marrakech opéré par *Royal Air Maroc* est vendu à la fois par *Expedia*, *GoVoyages*, *TravelGenio* et *BudgetAir*, et possède donc quatre séries bien distinctes. Pour pouvoir identifier ces séries, il suffit de rechercher tous les *id_flight* associés à l'*id_unique_flight* correspondant.

Il aussi intéressant d'observer l'évolution parallèle d'un billet vendu par la compagnie qui affrète le vol et de celui vendu par l'agence de voyage ayant des accords commerciaux avec ladite compagnie.

Observons que les agences de voyages, en plus des variations du billet souvent liées à la variation du billet vendu par la compagnie régulière, ajoutent leur propre couche de yield management afin d'augmenter davantage leurs revenus. Ils optimisent alors leur prix jusqu'à plusieurs fois par jour. Sur la Figure 1.6, nous pouvons voir l'évolution de prix d'un même billet Paris-Bangkok vendu par la compagnie régulière (*Qatar*) et l'agence de voyages (*GoVoyages*). Selon l'heure de la journée certaines agences peuvent ajouter des frais supplémentaires et dans le cas présent, *GoVoyages* applique la tarification du Tableau 1.1.

Horaire	0-2	2-4	4-6	6-8	8-19	19-23	23-24
Supplément(€)	7	6	0/6[5]	4	8	9	11

TABLE 1.1 – Évolution du supplément sur le site *GoVoyages* en fonction de l'heure d'achat (le 27 novembre 2012).

Qatar fait vendre une partie de ses sièges par *GoVoyages* qui propose des tarifs plus avanta-

[5]En fonction des compagnies

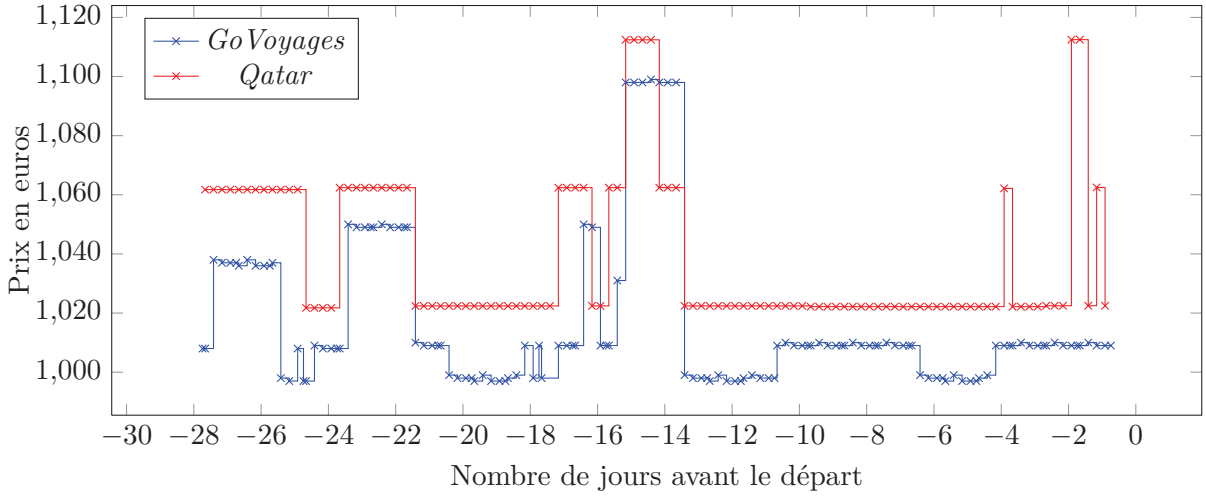

FIGURE 1.6 – Vol Paris-Bangkok opéré par *Qatar*, vendu par *GoVoyages* et *Qatar*

geux en échange d'une visibilité plus importante du vol. *GoVoyages* applique alors ses propres optimisations de prix tout en suivant les variations importantes de *Qatar*. Bien que nous n'utilisons pas cette information, elle pourrait nous permettre de prédire un saut de prix de manière presque certaine.

Les attributs

A chaque *id_flight* est associé un vecteur d'attributs $V_i(1), \ldots, V_i(p)$ qui regroupe toutes les informations possibles sur le billet vendu. Ces attributs sont très utiles pour extraire des statistiques sur notre base d'apprentissage. Ainsi il est facile de trouver la destination la plus populaire en Asie, ou bien le prix minimum observé depuis Paris vers n'importe quelle ville allemande opéré par une low cost. C'est en grande partie grâce à ces attributs que nous comptons prédire l'évolution de la trajectoire des vols, c'est pourquoi il est important de conserver l'intégralité des informations fournies par le site marchand et de créer de nouveaux attributs dérivés.

Les attributs peuvent être divisés en quatre catégories : tout d'abord nous utilisons les attributs définissant un vol unique (Tableau 1.2).

Nom	Type	Description
day	$\in 1-31$	Jour du mois de départ
month	$\in 1-12$	Mois de départ
year	$\mathbb{N}$	Année de départ
departureHour	$\in 0-23$	Heure de départ
departureMinute	$\in 0-59$	Minute de départ
transportCode	Code	transporteur du vol aller
departureStation	Code	Code aéroport de départ
arrivalStation	Code	Code aéroport d'arrivée

TABLE 1.2 – Caractéristiques du vol unique. Les trajets étant des allers-retours, ces attributs sont tous doublés.

Viennent ensuite les attributs dérivés de ces dernières caractéristiques (Tableau 1.3) qui détaillent les attributs temporelles (saison, départ le week end, jour de l'année, période vacances scolaires, etc.), les attributs géographiques (ville, pays, continent) et les attributs liés au trajet (long-courrier, nombre d'escales, etc.).

Nom	Type	Description
season	$\in 1,2,3,4$	Saison (1=printemps,...)
length_of_stay	$\mathbb{N}$	Durée du séjour
day_of_year	$\in 1-365$	Jour de l'année
day_of_week	$\in 1-7$	Jour de la semaine (1=lundi)
dep_on_we_dep	boolean	Départ le week end
dep_periode	$\in 1,2,3,4$	Période de la journée (1=très tot, 2=tot,3=après midi,4=soir)
holiday	$\in 0,1,2$	0=hors vacances, 1=vacances, 2=grandes vacances
stops	$\mathbb{N}$	Nombre d'escales
distance	$\mathbb{R}^+$	Saison
haul	$\in 0,1,2$	0=court-courrier, 1=moyen-courrier, 2=long-courrier
departureCity	Code	Code ville de départ
departureCountry	Code	Code pays de départ
departureContinent	Code	Code continent de départ
arrivalCity	Code	Code ville d'arrivée
arrivalCountry	Code	Code pays d'arrivée
arrivalContinent	Code	Code continent d'arrivée

TABLE 1.3 – Attributs dérivés des caractéristiques du vol unique

Nous avons aussi des attributs liés au site marchand (Tableau 1.4) et enfin les attributs contextuels qui évoluent avec les points de la série temporelle (Tableau 1.5) : à chaque instant t, nous calculons par exemple le nombre de sauts observés précédemment ou la somme des demandes faites pour ce vol. L'utilisation des ces derniers attributs est différente des autres, c'est pourquoi nous stockerons ces informations dans une table à part où la clef sera à la fois

l'*id_flight* et l'instant t.

Nom	Type	Description
provider	Code	Code du site marchand
type	$\in 0, 1, 2$	1=Agence de voyages, 2=Compagnie régulière, 3=Low cost,
directSeller	boolean	Vendeur direct
train	boolean	Trajet en train

Table 1.4 – Attributs liés au site marchand

Nom	Type	Description
volatility	$\mathbb{N}$	Nombre de sauts passés
volatility_hausse	$\mathbb{N}$	Nombre de sauts à la hausse passé
volatility_baisse	$\mathbb{N}$	Nombre de sauts à la baisse passé
demand	$\mathbb{N}$	Nombre de recherches utilisateur passé

Table 1.5 – Attributs contextuels évoluant avec la série temporelle

Nous sommes maintenant capables de construire facilement nos séries temporelles, d'accéder aux attributs d'un vol ou même d'afficher simultanément les séries de vols uniques vendus par des sites différents. Notre seule difficulté est de mettre en évidence les vols co-brandés : il faudrait pour cela créer un identifiant unique supplémentaire composé des attributs d'un vol unique sans les codes transporteurs mais avec les horaires des atterrissages. Cet ajout de complexité dans l'architecture de notre base de données diminuerait les performances pour une utilité limitée, c'est pourquoi nous avons préféré ne pas prendre en compte ce phénomène.

Une fois notre base de données construite nous pouvons commencer à l'explorer pour en extraire les premières informations sur les comportements des usagers et des séries temporelles. Ces comportements influeront sur les paramètres de notre modélisation des séries temporelles.

1.4 Statistiques expliquant le choix des paramètres

L'architecture de notre base de données nous permet maintenant d'extraire de nombreuses informations essentielles à la construction de notre modèle de prédiction : les séries de prix, leurs attributs et toutes les statistiques dérivées possibles. La collecte de nos données ayant débuté il y a plus de 6 ans, l'importante masse d'informations à traiter nécessite des choix dans la sélection des séries à étudier. Cette sélection doit à la fois refléter l'attente des utilisateurs sur un service de prédiction (fiabilité sur l'ensemble des routes, prédiction disponible le plus tôt possible, etc.) et prendre en compte les spécificités du marché du voyage (différences de comportements entre vols touristiques et business, suivant les durées de séjour, etc.).

Il est important de pouvoir proposer un service de prédiction à un maximum d'utilisateurs sans détériorer les performances. Il faut donc utiliser un panel représentatif des comportements de prix avec un échantillonage suffisant et choisir des longueurs de séries couvrant le plus large spectre de recherches utilisateurs.

Nous allons donc dans un premier temps décrire les paramètres des vols étudiés (route, durée de séjour, site marchand), puis nous expliquerons le choix de la longueur des séries temporelles.

1.4.1 Les trajets

Afin d'éviter la manipulation d'un trop grand nombre de séries, nous nous sommes limités dans un premier temps à un ensemble de routes représentatives (Tableau 1.6). Nous avons choisi des destinations touristiques (Paris-Bangkok) et professionnelles (Paris-Budapest) pour des durées de séjours (*Lenght of stay*) touristiques (7 et 14 jours) et professionnelles (3 jours). Nous avons par la même occasion sélectionné des vols long-courriers, moyen-courriers et court-courriers (Paris-Toulouse). Les résultats des agences de voyages nous permettent de récupérer des vols sur un large panel de compagnies aériennes et les sites des compagnies nous garantissent des prix stables et plus fiables. Les deux sites peuvent vendre le même billet mais les agences modifient fréquemment le prix pour ajuster leur marge dynamiquement. L'objectif est de collecter le plus de comportements différents tout en minimisant la taille des données.

Lowcost			
From	*To*	*Provider*	*Length of stay*
Paris (ORY)	Budapest	EasyJet (U2)	3,7
Paris (ORY)	Toulouse	EasyJet (U2)	3,7
Paris (ORY)	Marrakech	Transavia (HV)	3,7
Paris (XGB)	Toulouse	IDTGV (TGV)	3,7
Agences de Voyages			
From	*To*	*Provider*	*Length of stay*
Paris (CDG)	Budapest	VoyagesSNCF (VOY)	3,7
Paris (ORY)	Marrakech	VoyagesSNCF (VOY)	3,7
Paris (CDG)	Toulouse	VoyagesSNCF (VOY)	3,7
Paris (PAR)	Bangkok	GoVoyages (GOV)	7,14
Amsterdam	Barcelone	GoVoyages (GOV)	3,7
Paris (CDG)	New York	GoVoyages (GOV)	7
Compagnies Régulières			
From	*To*	*Provider*	*Length of stay*
Paris (PAR)	Bangkok	Quatar (QR)	7,14
Amsterdam	Barcelone	Austrian (OS)	3,7

TABLE 1.6 – Routes de la base d'apprentissage

Il faut par ailleurs sélectionner des trajets fréquemment recherchés pour pouvoir construire des séries consistantes, c'est pourquoi nous avons choisi les durées de séjour les plus demandées (Figures 1.7 et 1.8).

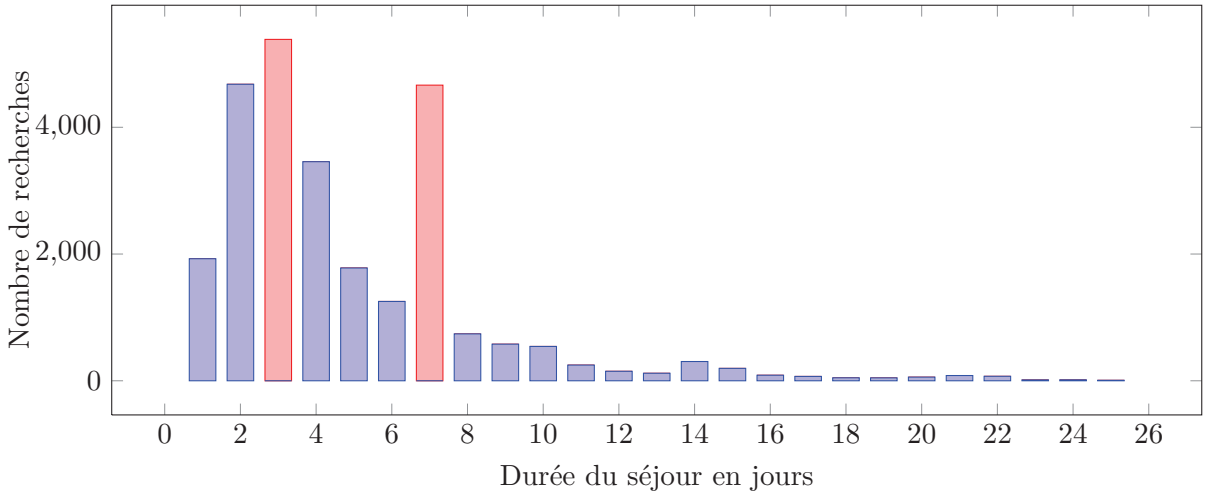

FIGURE 1.7 – Nombre de recherches par durée de séjour pour un Paris-Budapest

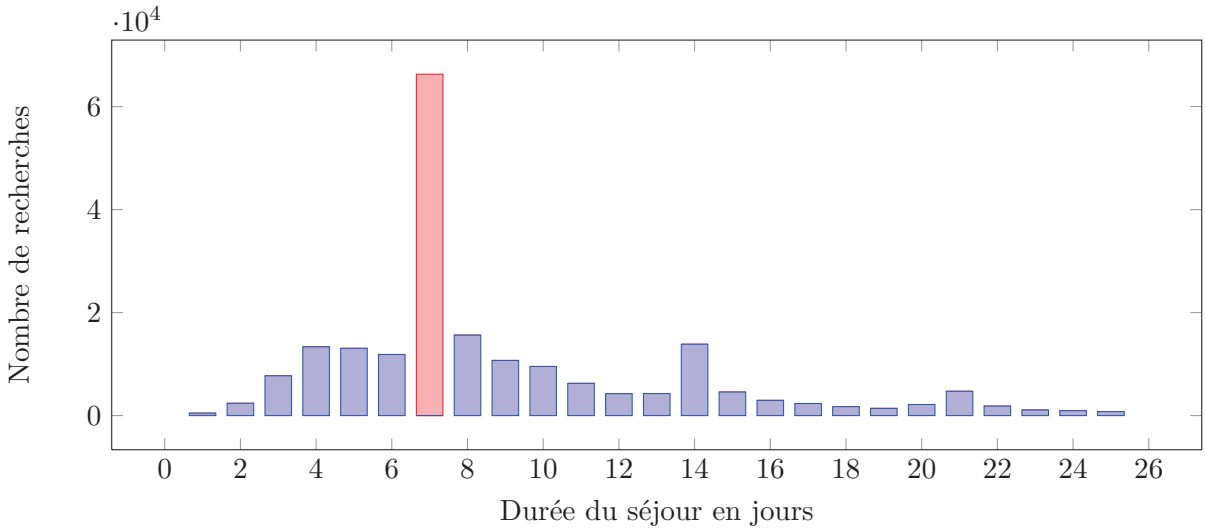

FIGURE 1.8 – Nombre de recherches par durée de séjour pour un Paris-New York

Pour étudier correctement le comportement des séries temporelles, il est souhaitable de détecter le maximum de changements de prix. Nous devons alors sélectionner les vols les plus recherchés pour garantir des séries de prix fidèles à la réalité. Les recherches sur les routes sélectionnées sont fréquentes mais pas nécessairement régulières. Plus la date de départ approche et plus les recherches sont nombreuses, c'est pourquoi nous voulons dans un premier temps étudier les derniers jours de nos séries pour nous assurer une base consistante. Cette réduction de la taille de nos séries implique une réduction du pourcentage d'utilisateurs susceptibles d'obtenir des prédictions. Inversement, en augmentant le nombre de jours étudiés nous augmenterions le spectre d'utilisateurs bénéficiant du service mais les performances et la fiabilité du service

diminueraient. Nous allons donc étudier le comportement des voyageurs pour trouver la durée optimale des séries temporelles à choisir.

1.4.2 La longueur des séries temporelles

Par l'évolution de la demande nous estimons la période à laquelle commencer l'étude des séries de prix. Le but est de couvrir le plus de demandes tout en conservant un jeu de données raisonnable et des séries consistantes. La ventilation représente l'évolution en pourcentage de la demande par rapport au nombre final de recherche. La Figure 1.9 représente les ventilations d'allers-retours pour différents types de destinations et différentes durées de séjours. Pour les long-courriers, la première moitié des recherches se fait entre un an et trois mois avant le départ et entre un an et 1 mois avant le départ pour les moyen-courriers. Dans notre exemple, nous couvrons avec les 28 derniers jours des séries temporelles 15% des utilisateurs pour les long-courriers et 50% pour les moyen-courriers, mais au global un tiers des usagers de liligo.com avec un nombre de recherches suffisant pour avoir des séries correctement échantillonnées.

Par ailleurs, le mois précédant la date de départ correspond à une forte augmentation des sauts comme le montre la Figure 1.10 et donc de la volatilité des prix. C'est alors un moment crucial pour l'utilisateur qui voit les prix changer jusqu'à plusieurs fois par jour et où le conseil à l'achat s'avère déterminant.

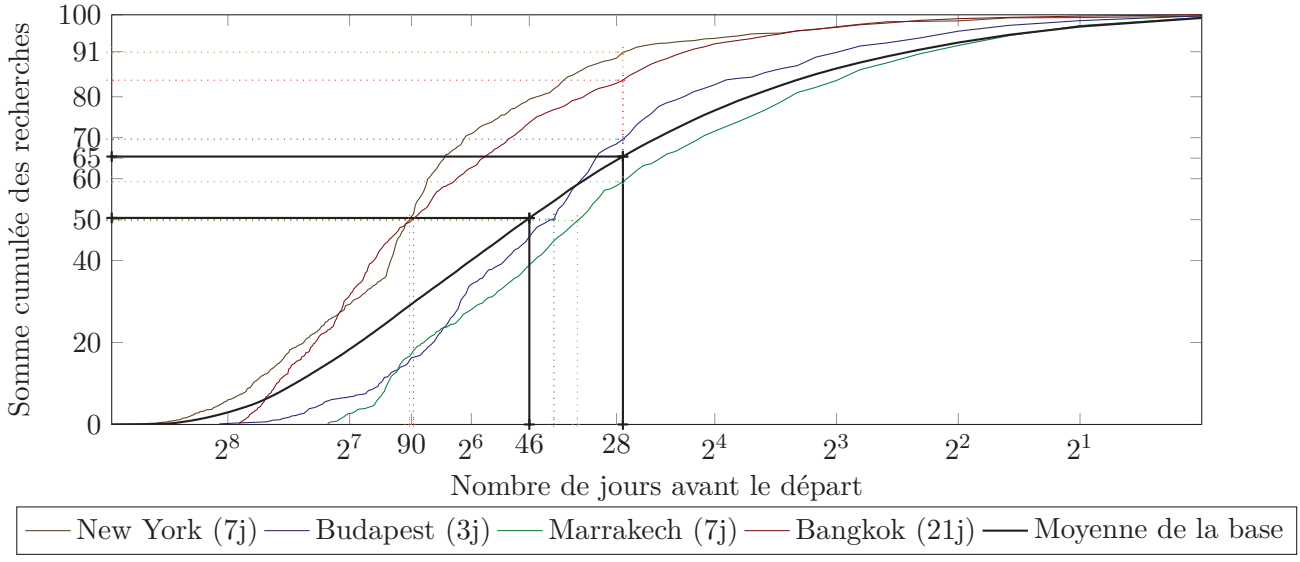

FIGURE 1.9 – Ventilation pour un vol Paris-New York du 14-04-2012 au 21-04-2012, un vol Paris-Bangkok du 28-07-2012 au 18-08-2012, un vol Paris-Marrakech du 21-04-2012 au 28-04-2012 et un vol Paris-Budapest du 01-11-2012 au 04-11-2012

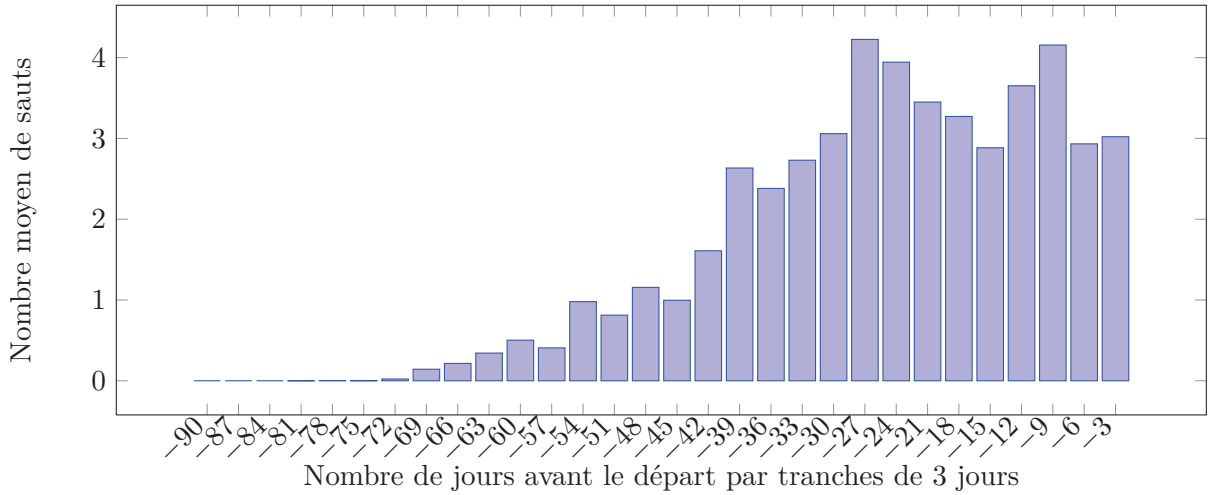

FIGURE 1.10 – Histogramme de l'évolution du nombre de sauts avant la date de départ par tranches de 3 jours.

En accord avec les contraintes d'échantillonnage évoquées dans le chapitre précédent et au vu de la Figure 1.9, nous avons donc décidé de nous focaliser sur les 28 derniers jours des séries de prix, mais nous conservons toutes les informations précédant le dernier mois pour une application étendue à 90 jours de notre méthode. De la même manière, nous avons dans un premier temps restreint notre corpus de vols, mais pour une future extension de notre service, nous continuons d'enregistrer tous les résultats des recherches utilisateurs.

1.5 Pertinence

L'initiative de ce projet est venue de la constatation que le voyageur est volontairement tenu dans l'ignorance des évolutions de prix pour entretenir une confusion quant au meilleur moment pour acheter son billet. Certaines règles sont alors communément utilisées telles que : "Plus on achète tôt, moins on paye cher" et son corollaire "Rien ne sert d'attendre, les prix ne font qu'augmenter".

Tout d'abord, il faut rappeler que nous ne tentons pas d'aider l'utilisateur à choisir la meilleure date pour partir, ni le meilleur moment de l'année pour acheter son billet.

Pour chaque recherche utilisateur, plusieurs offres à différents horaires et proposées par différentes compagnies sont présentées. Dans ces choix l'utilisateur va identifier le vol qu'il souhaite prendre et c'est à ce moment qu'il lui sera conseillé d'acheter son billet immédiatement ou de reporter son achat pour économiser de l'argent.

Sur d'autres sites[6], seule l'évolution du plus bas prix observé quotidiennement est étudiée. Lorsque la prédiction d'une baisse de prix est prodiguée, l'utilisateur n'a pas la garantie que les caractéristiques du vol au plus bas prix à 7 jours corresponderont à celles du vol actuellement le moins cher. La compagnie aérienne affrétant le vol et les horaires étant des critères souvent

[6]Bing.com ou Kayak.com par exemple

très importants pour l'utilisateur, nous avons choisi de faire une prédiction pour chaque billet retourné par la recherche.

Nous évitons ainsi le risque de détériorer la qualité du vol en question en passant, par exemple, d'un vol direct à un vol à multiples escales d'une durée beaucoup plus importante que nécessaire.

Nous montrons donc dans cette section la pertinence de notre service dans l'aide à la décision du voyageur.

1.5.1 Meilleur moment pour acheter

Le comportement le plus courant chez un voyageur est d'acheter ses billets d'avion en avance pour s'assurer un prix raisonnable et éviter les hausses successives de prix. Des études ont montré que les prix n'étaient pas strictement croissants et qu'une période située aux alentours de 8 semaines avant le départ était optimale [37]. Sur la Figure 1.11 représentant l'évolution des prix des billets de l'intégralité de notre base de données, il existe bien une période où les prix sont généralement au plus bas et qui correspond à environ 50 jours avant la date de départ (soit 7 ou 8 semaines avant le départ).

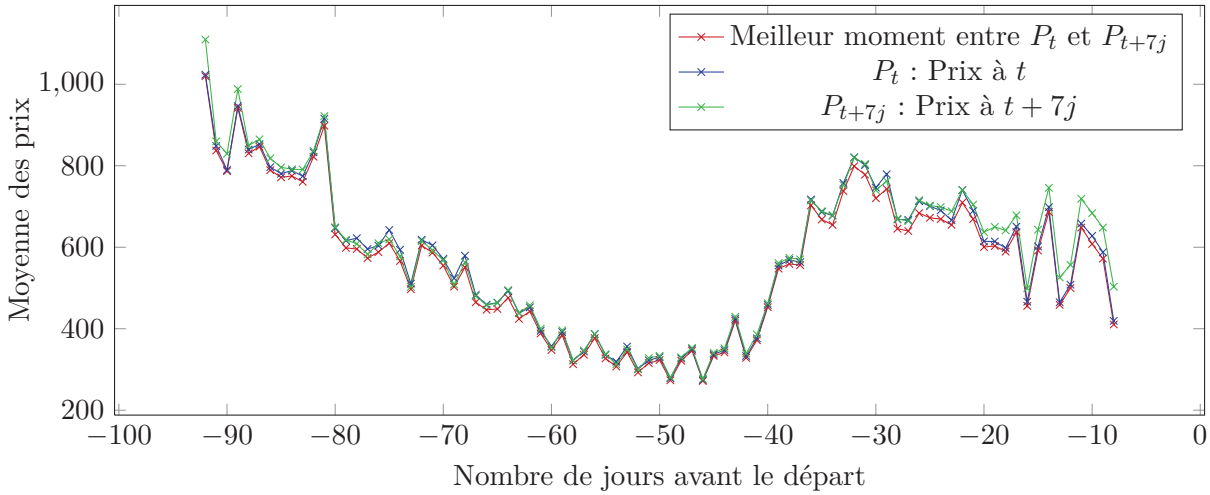

FIGURE 1.11 – Évolution de la moyenne des prix en fonction du moment d'achat

Cependant, comme nous l'avons vu sur la Figure 1.9, la moitié des usagers du site font leurs recherches au delà du 46e jour avant le départ, ce qui correspond à la période où les prix commencent à augmenter. Et c'est à cette période qu'il est crucial de conseiller au voyageur d'acheter son billet avant que la hausse ne soit trop importante ou bien d'attendre, car malgré le comportement globalement haussier de l'ensemble des séries, il est probable qu'une baisse apparaisse.

Toujours sur la Figure 1.11, on observe l'évolution du prix à l'instant t et celui 7 jours après et l'évolution du meilleur des deux prix (optimum). Ainsi on constate que la période la plus propice au conseil d'achat, c'est-à-dire lorsque la courbe optimale s'éloigne le plus des autres

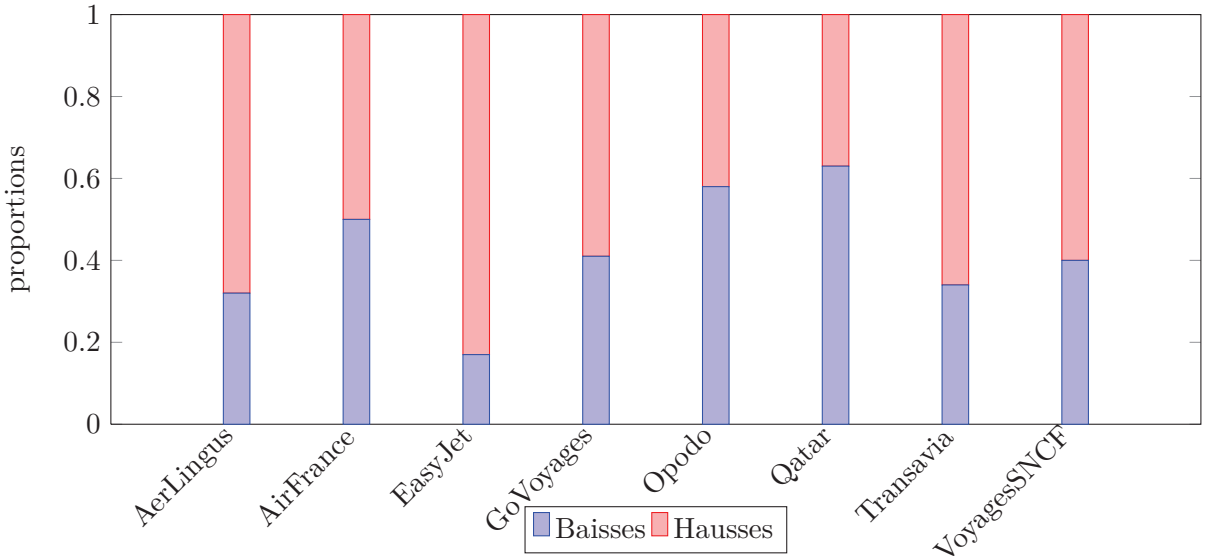

FIGURE 1.12 – Proportion de hausses et de baisses par provider

courbes, se trouve entre 35 et 20 jours avant la date de départ. Dans cette période le bon conseil d'achat ferait économiser de l'argent à l'utilisateur. En ce qui concerne les derniers jours, le conseil est aussi primordial mais évitera surtout au voyageur de perdre de l'argent. La hausse des prix est alors beaucoup plus importante et il donc plus risqué d'attendre une baisse de tarif.

On peut enfin remarquer les très faibles différences de prix entre 60 et 40 jours avant le départ. Les sauts de prix sont alors moins fréquents ce qui prouve encore que c'est une période propice à l'achat de billet d'avion.

1.5.2 Proportion des baisses

Nous insistons sur la nécessité de notre service en montrant que la proportion des baisses de prix est beaucoup plus importante qu'on ne le pense. Sur la Figure 1.12, on constate par exemple qu'*Air France* et *Qatar* pratiquent autant voire plus de baisses de prix que de hausses. En revanche, les compagnies low cost pratiquant des optimisations de prix plus agressives et ne vendant généralement pas de billets remboursables ou échangeables, affichent peu de baisses de prix. Certains attributs (ici la compagnie opérante le vol, *supplier*) sont donc plus discriminants que d'autres dans la prédiction des évolutions de prix. Ainsi un billet *EasyJet* aura plus de probabilité d'augmenter dans les 7 jours qu'un vol d'une autre compagnie.

Nous allons maintenant montrer qu'il est possible d'économiser de l'argent quelle que soit la date avant le départ en prenant les bonnes décisions.

1.5.3 Gain optimal

Nous définissons un gain ou une perte comme valeur absolue de la différence entre le prix initial et le prix à 7 jours qui sera additionnée si la prédiction est bonne et soustraite dans le cas contraire. L'optimal ou oracle est le prédicteur qui sait dans tous les cas quelle décision prendre. Sur la Figure 1.13, nous calculons l'évolution de l'écart relatif à l'oracle en fonction de la date de prédiction pour une personne qui prendrait systématiquement le billet le jour de la recherche, pour une personne qui prendrait systématiquement le billet 7 jours après la recherche et enfin pour une personne qui choisirait au hasard.

Trois périodes se dégagent de ce graphique : avant 80 jours avant la date de départ il est conseillé d'acheter tout de suite. Puis de -80 jours à -50 jours, les prix ont tendances à baisser pour atteindre le prix minimal. Enfin après les 50 jours avant la date de départ il est de nouveau conseillé d'acheter ses billets le jour de la recherche.

La zone rouge représente le gain que nous pourrions apporter à l'utilisateur en lui donnant un conseil plus précis que le précédent et représente donc l'objectif à atteindre.

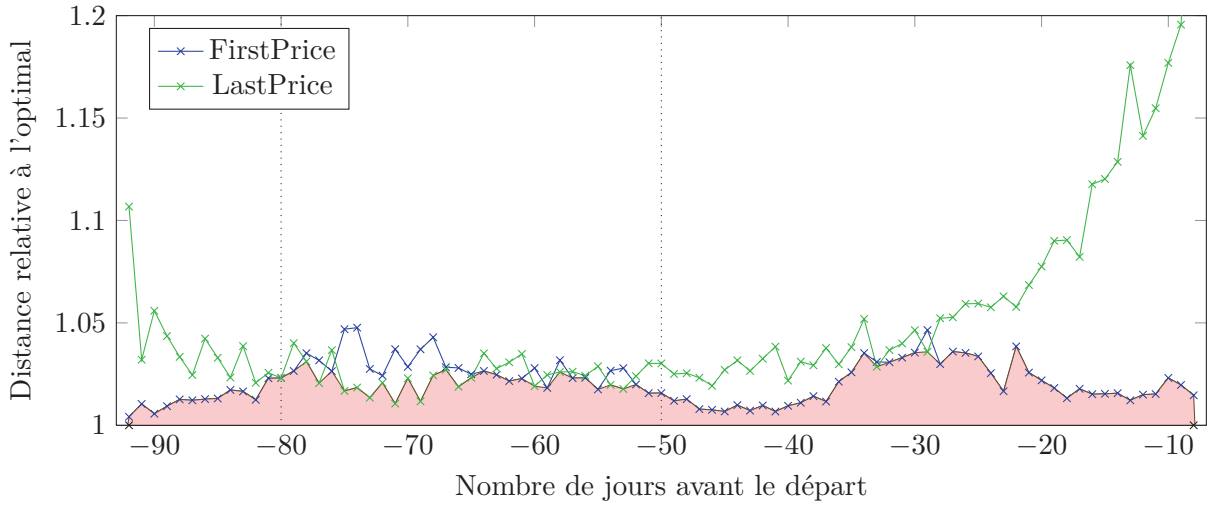

FIGURE 1.13 – Distance relative à l'optimal

1.6 Conclusion et perspectives

Dans ce chapitre nous avons tenté de décrire l'environnement dans lequel nous travaillons, et le contexte économique du monde du voyage. Le "revenue management", technique qui consiste à optimiser le prix de chaque siège d'un avion grâce à la prédiction de la demande, rend le processus d'achat du voyageur complexe et obscur. Le consommateur est maintenu dans l'ignorance et l'incertitude, parfois encouragé à réserver longtemps à l'avance, parfois exhorté à réserver précipitamment à la dernière minute pour bénéficier d'offres présentées comme dégriffées.

Dans un second temps nous avons décrit la structure de notre base de données contenant les recherches utilisateurs et les alertes mails. Elle nous permet d'extraire facilement les séries de

prix, de mettre en parallèle les mêmes billets, vendus par différents sites marchands et d'afficher les attributs correspondants aux vols.

Nous avons tenté d'avoir une base représentative de la demande des utilisateurs ainsi que des différents types de voyageurs, d'avoir des séries de prix régulièrement échantillonnées, tout en limitant la taille de nos données. Pour cela nous avons construit une base sur des destinations et des longueurs de séjour limités, et nous nous sommes focalisés sur les 28 derniers jours avant le départ, garantissant ainsi un nombre de points suffisant pour détecter la grande majorité des sauts de prix.

Enfin nous avons montré qu'il était nécessaire d'apporter une information sur l'évolution du prix aux usagers de liligo.com qui ne possèdent pas la vision globale du marché du tourisme. Il est courant de penser que les prix dans le milieu du tourisme aérien ne font qu'augmenter, mais nous avons démontré qu'il en était autrement et qu'avec notre service nous pourrions détecter une baisse de prix dans une courbe au comportement globalement haussier.

Avec l'accroissement du trafic de liligo.com, il sera possible d'élargir le nombre de destination et la durée des séries temporelles, permettant ainsi de proposer le service à un plus grand nombre d'utilisateurs dans des périodes durant lesquelles ils en ont le plus besoin. Il sera notamment intéressant d'élargir le système d'alerte pour enregistrer l'intégralité des résultats de recherche afin de créer des séries temporelles complètes.

Dans le chapitre suivant nous allons aborder en détail la modélisation des séries temporelles par des processus ponctuels poissonniens dans le but de regrouper les séries de même comportements. Une analyse des comportements des séries temporelles validera notre choix de transformation et nous permettra de choisir les paramètres optimaux de notre nouvelle représentation.