Titre: Title:	Modèles de demande et gestion du revenu en transport aérien
Auteur: Author:	Miladin Djurisic
Date:	2007
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Djurisic, M. (2007). Modèles de demande et gestion du revenu en transport aérien [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7975/

UNIVERSITÉ DE MONTRÉAL

MODÈLES DE DEMANDE ET GESTION DU REVENU EN TRANSPORT AÉRIEN

MILADIN DJURISIC
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTE EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(MATHÉMATIQUES APPLIQUÉES)
MAI 2007

Your file Votre référence
ISBN: 978-0-494-29229-7
Our file Notre référence
ISBN: 978-0-494-29229-7

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

MODÈLES DE DEMANDE ET
GESTION DU REVENU EN TRANSPORT AÉRIEN

présenté par : DJURISIC Miladin
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. ADJENGUE Luc, Ph.D., président

M. SAVARD Gilles, Ph.D., membre et directeur de recherche

M. MARCOTTE Patrice, Ph.D., membre et codirecteur de recherche

M. CÔTÉ Jean-Philippe, Ph.D., membre

RÉSUMÉ

Ce mémoire présente trois modèles de choix discret pour analyser la demande des passagers en transport aérien. La caractéristique principale de ces modèles est qu'ils supposent que chaque individu d'une certaine population (*eg.* les passagers voyageant entre deux villes) fait face à un choix d'alternatives et observent l'alternative qu'il choisit. Par la suite, la probabilité de refaire un tel choix simultané est modélisée à l'aide de caractéristiques reliées à l'individu et à l'alternative. Chaque caractéristique a un poids et la somme de ces poids représente la valeur (utilité) de l'alternative. Ainsi, l'alternative avec la plus grande valeur aura la plus grande probabilité d'être choisie aux dépends des autres. En regroupant ces probabilités individuelles, la part de marché de chaque alternative peut être calculée et la demande pour celle-ci est une redistribution de la demande totale selon sa part.

Dans le contexte de l'analyse de la demande de transport, les modèles de choix discret ont joué un rôle important au cours des 25 dernières années. La plupart des applications utilisent les modèles d'utilité aléatoire qui se basent sur les règles déterministes de la théorie économique néo-classique et captent l'incertitude à l'aide de variables aléatoires dans les utilités. Les modèles utilisés dans ce mémoire font partie de cette catégorie.

La démarche suivie est de commencer par un modèle logit multinomial relativement simple et d'en examiner les résultats. Ensuite, dans le but d'améliorer ces résultats, une formulation logit imbriquée, puis logit imbriquée croisée, est présentée où des nids sont créés pour regrouper les alternatives soupçonnées de se concurrencer plus étroitement. L'échantillon de données provient d'un grand transporteur aérien nord-américain et permet de définir des itinéraires (déplacement unidirectionnel entre deux villes), puis de comparer les différentes classes de service offertes aux passagers sur ces itinéraires.

L'approche innovatrice de ce mémoire est le nombre et la gamme d'attributs reliés aux alternatives et aux individus qui sont utilisés dans les modèles. Les attributs sont des variables explicatives dans les modèles et leur nature peut causer certains problèmes numériques au cours de l'estimation. Quelques pseudo-attributs liés aux passagers sont aussi inclus mais provoquent les mêmes résultats et conséquences.

ABSTRACT

This paper uses three discrete choice models to analyze airline passenger demand. The main feature of these models is that they take each individual from a population (*eg*. Passenger flying between two cities), place him in front of set of alternatives and record his choice. At a later stage, the probability of remaking that same choice is modeled as a function of attributes of the decision maker and the alternatives. Each attribute carries a distinct weight in the decision process and the sum of these weights amounts to the value (utility) of the alternative. Hence, the alternative with the highest value will have the highest probability of being selected. By grouping the individual probabilities, the market share of each alternative can be calculated and demand for that particular alternative is a redistribution of the total demand.

In transportation demand analysis, discrete choice models have played major role over the last decades. Random utility models, which combine deterministic rules from neo-classical economic theory and capture uncertainty with random variables added to the utility functions, are used in this paper.

The approach used is to start with a relatively simple multinomial logit model and to examine the results. Then, a nested structure, followed by a cross-nested structure, is included in the mutlinomial logit model to better those results. In other words, nests are created to gather alternatives suspected of being closer substitutes of each other. By capturing this underlying competition, a better demand model will arise. The data is provided by a major North American carrier and allows for passenger itineraries (one-way flights or combination of flights between two cities) to be built. Different levels of service offered to passengers on those itineraries can then be compared.

This paper is innovative in its approach to include many level of service attributes in its models. The attributes enter the models as explanatory variables but their nature can cause numerical difficulties during the estimation process. Some pseudo-

attributes linked to the passengers are also included but with similar results and consequences.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1: INTRODUCTION

Ce mémoire présente trois modèles de choix discret pour analyser la demande de passagers en transport aérien et s'inscrit dans un cadre plus vaste de gestion du revenu. Pour un transporteur aérien qui tente de maximiser les revenus générés par la vente de billets, une modélisation précise de la demande joue un rôle primordial dans la réalisation de cet objectif. Dans l'optique où le billet est un produit qui donne droit à un service de transport, chaque produit est défini par un ensemble d'attributs tels: l'origine, la destination, le prix, le moment du départ, etc. L'approche commune des modèles présentés est de supposer que les passagers accordent un poids différent à chacun des attributs et que l'addition de ces poids résulte en une valeur (ou utilité) globale pour le produit. Cette valeur permet de classer les produits et conséquemment, le produit au premier rang aura la plus haute probabilité d'être choisi. Le transporteur aérien pourra ainsi déterminer la demande pour chaque produit offert en fonction de certains attributs que les passagers considèrent plus ou moins désirables. Or, s'il ne connaissait pas la valeur que les passagers attribueraient à un nouveau service de transport qu'il voudrait offrir, il serait incapable de prédire combien de passagers l'utiliseraient.

L'organisation du mémoire succède les trois sections suivantes qui couvrent brièvement l'industrie du transport aérien, la gestion du revenu et l'optimisation bi-niveau.

1.1 Industrie du transport aérien

Souvent décrite comme un marché très compétitif, l'industrie du transport aérien a subi de maintes réformes au cours des dernières années qui impactent l'expansion de ses services de transport domestiques et internationaux[1]. L'entrée de nouveaux compétiteurs est facilitée par des taux d'intérêt bas ce qui a pour effet de saturer le

[1] http://www.investopedia.com/features/industryhandbook/airline.asp

marché. Les programmes de fidélité et l'image corporative sont généralement suffisants pour attirer et retenir les passagers (même s'ils payent plus cher) mais puisque les marges de profits sont minces, les conséquences peuvent être désastreuses lorsque l'économie ralentit.

Au point de vue des fournisseurs, Boeing et Airbus se partagent le marché de la fabrication d'avions et il est improbable qu'ils s'intègrent verticalement en offrant un service de transport aérien commercial. Au point de vue des transporteurs, ils sont impuissants face aux coûts élevés de nouveaux appareils et se doivent de choisir parmi les quelques modèles et configurations possibles, ce qui les empêche de concurrencer au niveau du service avec les autres transporteurs. Les vols de la compétition sont considérés comme substituts et pour les vols plus courts, d'autres modes de transport en surface viennent concurrencer le service offert par les transporteurs.

La plus large partie des revenus encaissés par un transporteur proviennent de billets vendus aux passagers réguliers (vacanciers) et d'affaires. Les passagers d'affaires sont importants parce qu'ils voyagent plusieurs fois dans l'année et achètent les tarifs les plus chers (moins restreints) qui rapportent plus de revenus au transporteur. Les passagers réguliers sont plus sensibles aux variations dans les prix et achètent, en général, les billets les moins chers (plus restreints). Lorsque l'économie ralentit ou que la confiance de la clientèle est ébranlée, c'est principalement le nombre de passagers réguliers qui baisse.

Les coûts principaux pour le transporteur sont le carburant et la main-d'œuvre. Ils représentent 15% et 40%, respectivement, du coût total d'opération. Finalement, il est important de considérer la zone géographique que le transporteur dessert. Il est désirable pour le transporteur de s'accaparer une part de plus en plus grande d'un marché donné mais, pour survivre à long-terme, il doit desservir plusieurs destinations (diversification).

1.2 Gestion du revenu

La gestion du revenu se définit comme une pratique de gestion qui cherche le meilleur rendement possible pour chaque unité de capacité disponible. Elle est particulièrement efficace dans les industries, comme le transport aérien, où les coûts fixes encourus pour offrir un produit sont considérables alors que les coûts variables encourus lorsque ce produit est vendu sont négligeables. En différenciant son produit, le transporteur segmente son marché de façon à transformer les surplus du consommateur sur chaque segment en revenus additionnels.

Pour obtenir un bon rendement de la gestion du revenu, le produit doit être disponible en quantité limitée et être périssable dans le sens où il existe un seuil au-delà duquel sa valeur résiduelle tombe à zéro. Par conséquent, le produit est vendu à l'avance et consommé au moment où il périt. C'est le cas de billets d'avions, de spectacles, etc.

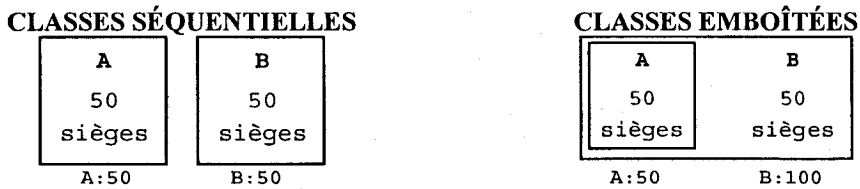
McGill et Van Ryzin (1999) présentent une revue de littérature assez détaillée et complète sur les différents aspects de la gestion du revenu et identifient les six hypothèses principales, notamment:

(1) *La séquence d'arrivés des réservations suit l'ordre croissant des tarifs offerts.* Le tarif correspondant à une classe est le prix du billet et les conditions applicables. Ces conditions sont généralement liées à la durée du séjour, à la période avant le départ où le billet doit être acheté et d'autres attributs. Un tarif contraint est généralement plus abordable puisqu'il offre moins de flexibilité. Un tarif non-contraint est plus dispendieux en retour la flexibilité qu'il offre. L'hypothèse veut que l'inventaire de la classe la moins dispendieuse se liquide en premier, ensuite la classe suivante la moins dispendieuse, etc.;

(2) *Les demandes pour chacune des classes tarifaires sont indépendantes les unes des autres au point de vue statistique.* Les classes tarifaires sont des catégories (ensembles d'attributs) qui regroupent plusieurs tarifs jugés appartenir à ces

catégories. Une demande indépendante signifie que si plus aucun billet n'est disponible dans la catégorie désirée, le passager décidera de pas voyager plutôt que d'acheter un billet disponible d'une classe tarifaire;

(3) *Les classes tarifaires sont séquentielles, i.e., sans emboîtement* (*nesting*). L'inventaire d'un vol est le nombre de sièges disponibles à la vente sur ce vol. Cet inventaire est ensuite réparti parmi les classes tarifaires existantes de sorte qu'un certain nombre de sièges soit alloué à chaque classe. Avec 100 sièges répartis uniformément sur deux classes, la différence entre une structure de classes séquentielles et emboîtées apparaît même avant la vente du premier siège. Dans le cas d'une structure de classes séquentielles, chaque classe a son propre inventaire puisque chaque classes est indépendante. Dans le cas d'une structure de classes emboîtées, les sièges de la classe A sont toujours disponibles à la classe B et sont liquidés en premier. Autrement dit, le dernier siège vendu appartiendra à la classe B et donc, ne sont liquidés que lorsque l'inventaire total tombe à 0. Initialement:



et suite à la vente d'un siège en classe B:

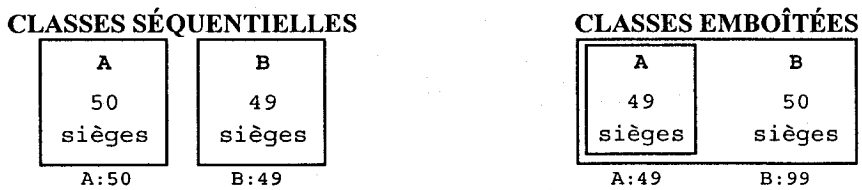


Figure 1.1: Classes imbriquées.

(4) *Il n'y a pas de sur-réservation.* La sur-réservation est la pratique d'accepter plus de réservations que la quantité de sièges disponibles en anticipant que certaines réservations ne se matérialiseront pas. Puisque les annulations et les

passagers ne se présentant pas à l'embarquement (*no-shows*) ne sont pas considérés, toutes les réservations se matérialisent;

(5) *Le vol est sur un seul segment ne considérant pas ainsi les effets-réseaux.* Chaque vol est considéré individuellement sans se soucier des autres vols du transporteur;

(6) *La séquence d'arrivé des réservations ne permet pas les arrivées multiples de réservations* (*batch bookings*). Les réservations arrivent une à la suite de l'autre;

Avec ces hypothèses, Littlewood (1972) établie une règle de protection considérée comme une des premières formes du coût de délogement (*displacement cost*) avec deux classes tarifaires. La problématique étant de déterminer combien de sièges protéger lors d'une action de vente de billets à tarif réduit r. Au temps t, il s'agit de continuer de vendre le billet à tarif réduit tant que:

$$r \geq (1 - P_t)R \tag{1.1}$$

où R est le tarif régulier et P_t est la probabilité de vendre au moins la quantité restante de billets. En prenant R = \$1000, r = \$200, N pour représenter le nombre de sièges protégés pour les passagers volant au tarif régulier R et M pour la quantité de sièges totale, la politique de gestion est de protéger N des M sièges aux passagers payant R en refusant plus de (M - N) passagers payant r jusqu'à ce que:

$$\Pr(\text{ne pas vendre } N \text{ billets à \$1000}) > \frac{r}{R} = \frac{200}{1000} = \frac{1}{5}. \tag{1.2}$$

En 1987, dans sa thèse de doctorat, Belobaba étend la règle de Littlewood à plusieurs classes tarifaires et introduit le terme EMSR (*Estimated Marginal Seat Revenue*) qui essentiellement calcule les coûts de délogement de chaque siège sur l'avion en fonction de données historiques et de la situation courante au point de vue des réservations. McGill et van Ryzin (1999) soulignent que l'approche de Belobaba ne permet de générer des limites de réservations optimales que dans le cas de deux classes tarifaires. Ils continuent en citant des ouvrages où ceci s'étend à plusieurs

classes, plusieurs segments de vol et finalement, aux itinéraires de passagers sur le réseau complet du transporteur.

Les méthodes les plus sophistiquées présentement utilisées sont celles des variables duales (*bid-price methods*) puisqu'elles remplacent les limites de réservations multiples sur les classes tarifaires et la structure d'emboîtement de celles-ci par un seul et unique coût de délogement. Ces techniques utilisent l'information provenant de l'optimisation, notamment la valeur des variables duales associées aux contraintes de flot, afin de calculer le coût de délogement (*bid-price*) sur cet itinéraire et déterminer si une nouvelle réservation doit être acceptée ou non. La somme des valeurs des variables duales associées aux segments que comporte l'itinéraire représente ce coût de déplacement et donc, sert de seuil pour l'acceptation des réservations.

Les limites de ces modèles sont les hypothèses sur lesquelles ils se basent, notamment: la séquence d'arrivée des réservations, la demande indépendante pour chaque classe tarifaire et le fait de négliger les effets-réseaux. En réalité, la séquence d'arrivée des réservations ne suit pas nécessairement l'ordre croissant des tarifs offerts et la demande pour une classe tarifaire n'est pas indépendante des autres classes. Dans le contexte où les conditions rattachées au tarif sont telles qu'elles constituent de vraies barrières entre les classes, l'hypothèse d'indépendance peut encore tenir. Cependant, il faut que l'avion soit suffisamment rempli pour empêcher les passagers dont la volonté de payer est plus haute de profiter de tarifs plus bas encore disponibles. Les maintes réformes qu'a subi l'industrie suite à la déréglementation aux États-Unis dans les années 1970 et au cours des dernières années ne la place pas dans ce contexte. C'est plutôt la tendance vers la réduction du nombre de tarifs offerts et la simplification de leur structure qui prend le dessus avec l'arrivée en force des transporteurs à rabais (*low-cost carriers*). Finalement, l'apparition d'alliances inter-transporteurs révèle l'importance des effets-réseaux.

Du point de vue de la tarification, il est maintenant reconnu qu'elle fait partie intégrante de la gestion du revenu à cause de la dualité qu'elle partage avec le contrôle de l'inventaire. Gallego et van Ryzin (1997) et You (1999) examinent un contexte où la tarification résulte d'un procédé stochastique et les vols peuvent être multi-segments ou avoir plusieurs classes tarifaires. Leurs modèles sont résolus par la programmation dynamique et ils expliquent que si le prix est interprété comme une variable qui peut être contrôlée sur une base continue, il y aura moyen de fermer une classe tarifaire en haussant le prix au-dessus d'un certain seuil. De plus, s'il y a plusieurs classes tarifaires disponibles simultanément, en fermer une peut s'interpréter comme un changement de structure tarifaire auquel fait face le passager potentiel. Outre Côté *et al.* (2003) qui présente un modèle d'optimisation bi-niveau pour la gestion du revenu, peu d'articles présentent une approche qui permet de traiter directement et simultanément la gestion de l'inventaire et la tarification. De plus, lorsque la tarification se fait à l'aide d'un tel modèle, les effets réseaux sont considérés.

1.3 Modèles d'optimisation bi-niveau

L'optimisation bi-niveau est une branche de la programmation mathématique dans laquelle un agent (le meneur) intègre au sein de son problème d'optimisation la réaction d'un autre agent (le suiveur). Une fois que le meneur a fixé ses variables de décision, le suiveur résout, à son tour, un problème d'optimisation où ces variables sont exogènes. Le meneur a donc connaissance du problème d'optimisation du suiveur. Les principaux éléments de ce procédé hiérarchique sont présentés au Tableau 1.1. Au premier niveau, le transporteur veut maximiser ses revenus en fixant ses tarifs de façon optimale mais il est contraint par la réaction des passagers face aux tarifs imposés. Ainsi, le second niveau est aussi un problème d'optimisation où les passagers maximisent l'utilité retirée du voyage.

Tableau 1.1: Optimisation du revenu par procédé hiérarchique.

Transporteur – Problème d'optimisation (Niveau I)
Cherche à maximiser le revenu sur son réseau en entier en tenant compte de:

- Contraintes commerciales (politiques, structures tarifaires, règlements, etc.)
- Contraintes du réseau (horaires, capacités du réseau, etc.)
- Contraintes de disponibilités de ses sièges
- Offre de sièges provenant de compétiteurs
- La réaction du client, i.e., le passager

Passager – Modèle de choix discret (Niveau II)
Cherche à satisfaire son désir de voyager tout en maximisant l'utilité associée avec sa décision d'achat dans laquelle il:

- Considère le prix et autres caractéristiques du produit (tangibles ou abstraites)
- A accès à de l'information détaillée sur l'offre des compétiteurs
- Prend une décision éclairée

Source: Riss *et al.* (2006), p.17.

Ce mémoire comporte sept chapitres. Le chapitre 2 présente une revue de littérature et introduit les aspects théoriques des modèles de choix discret. Le chapitre 3 est axé sur les applications de ces modèles en transport aérien alors que le chapitre 4 documente le traitement des données. Trois modèles de choix discret sont proposés au chapitre 5 pour déterminer la part d'itinéraire et leurs résultats sont analysés. Le chapitre 6 valide les résultats obtenus. Le dernier chapitre résume les aspects théoriques importants et les résultats obtenus dans cette recherche.

CHAPITRE 2: REVUE DE LA LITTÉRATURE

Dans ce chapitre, les approches pour modéliser le choix d'un consommateur face à un éventail d'alternatives sont présentées. Ces modèles sont appelés modèles de choix discret puisqu'ils se basent sur le choix révélé d'un individu face à un choix d'alternatives distinctes. Ils permettent d'incorporer les caractéristiques de l'alternative et de l'individu pour calculer la part de marché de chaque alternative.

Par exemple, un individu voulant voyager entre deux villes à une certaine période se verra offrir plusieurs tarifs (alternatives) sur plusieurs vols offerts par plusieurs transporteurs. Il choisira l'alternative qu'il juge la meilleure, i.e., celle qui lui apporte le plus grand surplus, aux dépends des autres[2]. Conséquemment, plus une alternative est choisie souvent, plus elle s'accaparera une grande part du marché. En définissant un marché comme étant le nombre de passagers voulant voler entre deux villes, le transporteur peut calculer sa demande ainsi: s'il connaît la part du marché que son tarif s'accaparera et la taille de ce marché, il n'aura qu'à multiplier ces deux quantités pour connaître le nombre de passagers qu'il transportera.

Du même coup, le transporteur pourra simuler plusieurs scénarios courants ou futurs en changeant les attributs des alternatives et la taille du marché. Le degré de précision de ces modèles devrait être très grand puisqu'ils partent d'observations individuelles puis les regroupent et additionnent, selon le besoin, pour représenter un certain ensemble. Pour cette raison, les modèles de choix discret sont dits désagrégés. Leur seule hypothèse à l'égard du comportement de l'individu est qu'il soit constant et rationnel dans son choix. L'importance qu'il accorde aux différents attributs est fixe durant la période à l'étude.

Le second attrait principal se situe au niveau de la décision elle-même. Une seule alternative est choisie une fois que toutes les alternatives ont été examinées et

[2] Si aucune des alternatives ne lui apporte un surplus du consommateur positif, l'individu choisira l'alternative de ne pas voler.

évaluées. Il n'y a pas d'arborescence dans le procédé de décision donc, pas d'évaluation séquentielle, ni de hiérarchie d'attributs. Les modèles de choix discret sont donc simultanés. Dans la plupart des situations en transport, le passager a le choix de la fréquence[3], du moment de la journée, de la destination, du mode et du chemin pour effectuer un déplacement. Dans un modèle simultané, le passager prend une décision pour chaque composante dans un même choix 'joint'. Autrement dit, tous les attributs sont considérés simultanément pour évaluer l'attrait d'une alternative et la classer par rapport aux autres. À première vue, l'inconvénient majeur de ces modèles est qu'ils requièrent énormément de données.

Warner (1962) a été un des premiers à utiliser des données désagrégées lorsqu'il a développé un modèle probabiliste de choix-modal. Depuis, plusieurs études ont appliqué ces modèles de choix-modal au transport urbain (ex. de Donnea, 1971; Reichman et Stopher, 1971; Charles River Associates, 1972; Peat, Marwick, Mitchell et cie, 1972; Ben-Akiva, 1973; Watson, 1974) montrant ainsi que la modélisation de la demande de transport au niveau désagrégé est réalisable. D'autres modèles simultanés désagrégés ont aussi été développés par Ben-Akiva (1973) et Cambridge Systematics (1973).

2.1 Propriétés théoriques du modèle de choix discret

Tel que mentionné précédemment, le décideur dans ces modèles désagrégés est l'individu. Ce concept d'individu peut facilement s'étendre pour représenter un petit groupe homogène tel une famille ou groupe de passagers voyageant ensemble. Dans ces cas, les décisions internes du groupe sont ignorées et seule la décision représentant le groupe en entier est considérée. L'ensemble d'alternatives contient un nombre fini d'alternatives qui s'énumèrent sous forme de liste et chaque alternative de cet ensemble doit être caractérisée par un ensemble fini d'attributs. Un attribut n'est pas nécessairement une quantité observable. Il peut être une fonction de données

[3] Le nombre de fois que le voyage est entrepris.

disponibles tel un ratio ou une transformation. Par exemple, au lieu de prendre directement le temps en transit comme attribut, le logarithme du temps en transit peut être utilisé.

La notion d'utilité provient de la théorie économique néo-classique qui suppose que chaque décideur est capable de comparer deux alternatives a et b appartenant au même ensemble en utilisant l'opérateur préference-indifférence $\succeq$. Par exemple, si $a \succeq b$ alors le décideur préfère a à b ou est indifférent. Cet opérateur est présumé posséder certaines propriétés désirables[4] et donc, l'utiliser pour faire un choix est équivalent à assigner une valeur, appelée utilité, à chaque alternative et de choisir l'alternative associée à la plus grande utilité.

Le concept d'utilité associé aux alternatives est important dans le contexte des modèles de choix discret. Cependant, les hypothèses néo-classiques à l'égard de l'opérateur $\succeq$ posent certains problèmes en pratique. Notamment, la complexité du comportement humain suggère qu'un modèle de choix discret doive tenir compte d'un certain niveau d'incertitude. La théorie économique néo-classique ne le fait pas.

L'incertitude peut venir de multiples sources. Par exemple, la théorie de l'utilité aléatoire suppose que le décideur a l'information complète au moment de son choix alors que l'analyste n'a que de l'information partielle et doit tenir compte de l'incertitude. Autrement dit, les individus considèrent toutes les alternatives qui leur sont offertes alors que l'analyste observe seulement les alternatives choisies par les individus et ajoute un terme d'erreur pour tenir compte des alternatives non-choisies. Les modèles d'utilité aléatoire se basent sur les règles déterministes de la théorie économique néo-classique et captent l'incertitude à l'aide de variables aléatoires dans les utilités. Ils sont utilisés dans ce mémoire et dans la plupart les applications en transport.

---

[4] Ces conditions sont expliquées par Bierlaire (1997).

Les sous-sections suivantes décrivent les hypothèses relatives aux composantes déterministes et aléatoires de ces modèles pour en arriver à la présentation d'une formulation logit multinomial. Les propriétés que les alternatives doivent avoir dans ces modèles de choix discret sont aussi présentées. Finalement, le concept d'élasticité est introduit pour faire ressortir l'interaction entre les attributs et les probabilités de sélection des alternatives. D'ailleurs, un des avantages de ces modèles est la simplicité avec laquelle différents scénarios peuvent être comparés et analysés.

2.1.1 représentation de l'utilité dans le choix probabiliste

En notant U_a^i l'utilité de l'alternative a pour le consommateur i et C_i est l'ensemble d'alternatives qui lui sont disponibles. La probabilité que le consommateur i choisisse l'alternative a parmi C_i est:

$$\Pr_{C_i}^i(a) = \Pr\left[U_a^i = \max_{b \in C_i} U_b^i\right] = \Pr\left(U_a^i \geq U_b^i, \forall b \in C_i\right). \tag{2.1}$$

L'ensemble C_i est, à la fois, mutuellement exclusif et exhaustif de sorte qu'une seule alternative est choisie, i.e, celle avec la plus grande probabilité).

L'utilité U_a^i est fonction de variables, notée $V_a^i = V_a^i\left(x_a^i\right)$, où x_a^i est un vecteur d'attributs qui caractérisent l'alternative a et le consommateur i. Cette composante de l'utilité est déterministe et, généralement, présumée linéaire (additive) dans ses paramètres. Si K attributs sont considérés, elle s'écrit:

$$V_a^i = \sum_{k=1}^{K} \beta_k x_a^i(k) = \beta^{\mathrm{T}} x \tag{2.2}$$

où $\beta = (\beta_1, \beta_2, \ldots, \beta_K)$ est un vecteur $K \times 1$ de coefficients à estimer pour chaque modèle.

L'hypothèse de linéarité de V_a^i est justifiée si les utilités sont considérées être des combinaisons de coûts équivalents pour les différentes alternatives. Dans cette optique, les coefficients β_K reflètent les facteurs de conversion d'unités d'un certain

paramètre en unités monétaires. De plus, elle est nécessaire puisque les procédures d'estimation disponibles requièrent la linéarité dans les paramètres. Finalement, en plus de simplifier la formulation et l'estimation du modèle, elle n'est pas aussi contraignante qu'elle puisse le paraître. Des effets non-linéaires peuvent tout de même être captés dans les attributs puisque, au besoin, ils seront définis comme des fonctions ou des transformations d'autres attributs disponibles dans les données.

L'utilité U_a^i est aussi fonction d'une composante aléatoire pour capter l'incertitude. En supposant que la composante aléatoire de l'utilité peut être exprimée comme un terme d'erreur ajouté, l'utilité s'écrit:

$$U_a^i = V_a^i + \varepsilon_a^i \tag{2.3}$$

où ε_a^i capte l'incertitude et représente la partie stochastique (aléatoire) de l'équation. Conséquemment, l'utilité est elle-même une variable aléatoire puisqu'elle renferme la composante aléatoire ε_a^i. En substituant (2.3) dans l'équation (2.1) et en réarrangeant les termes, l'équation suivante est obtenue:

$$\Pr_{C_i}^i(a) = \Pr\left(V_a^i - V_b^i \geq \varepsilon_b^i - \varepsilon_a^i\right). \tag{2.4}$$

Cette expression implique que la forme mathématique du modèle de choix discret est déterminée à partir de l'hypothèse faite sur la distribution jointe des éléments aléatoires. En pratique, l'espérance mathématique des termes d'erreur ε_a^i est présumée égaler 0. Bierlaire (1997) démontre que cette condition ne provoque aucune perte de généralité lorsqu'une constante spécifique à l'alternative (*ASC*) est incluse dans le modèle.

2.1.2 formulation logit multinomiale

En supposant que les éléments aléatoires sont distribués de façon indépendante et identique suivant une distribution de Gumbel (Gumbel, 1958) avec paramètres de localisation 0 et d'échelle 1, il est démontrable (Ben-Akiva, 1973; Charles River

Associates, 1972) que le modèle de choix discret prend la forme du modèle logit multinomial suivant:

$$\Pr_{C_i}^{i}(a) = \frac{e^{V_a^i}}{\sum_{b \in C_i} e^{V_b^i}}. \tag{2.5}$$

Les coefficients β_k sont estimés à partir d'observations de choix actuels et de l'éventail d'alternatives disponibles mais non-choisies. Ainsi, la variable observée prend la valeur 1 lorsque l'alternative est choisie et 0 sinon. Les prévisions vont donc donner la probabilité de choisir chacune des alternatives et leur somme donne 1.

Les hypothèses faites sur les distributions des éléments aléatoires sont questionnables dans bien des cas de modèles de choix discret mais, en revanche, la simplicité des calculs qu'elles procurent sont à la base de la popularité des modèles logits[5]. De plus, différentes hypothèses sur les composantes aléatoires et déterministes de l'utilité produiront différents modèles de sorte que chaque modèle est spécifique aux hypothèses sous-jacentes. Tableau 2.1 et Tableau 2.2 donnent quelques exemples. En substituant $Y_a = e^{V(a)}$ dans le modèle logit (2.5), le modèle développé par Luce (1959) est obtenu. Ce modèle a permis au modèle à utilité aléatoire de se développer et est un des premiers modèles à affecter une probabilité de sélection à une alternative au lieu d'identifier une alternative comme l'option choisie. Sans contredit, une caractéristique importante des modèles qui tiennent compte de l'incertitude.

2.1.3 indépendance des alternatives non-pertinentes (IIA)

Lorsqu'une alternative a une probabilité de 0 (ou très près de 0) d'être choisie, son inclusion ou exclusion de l'ensemble d'alternatives aura un effet négligeable sur les résultats d'estimation et de prévision du modèle. Il est donc possible d'attribuer des alternatives peu attrayantes à n'importe quel individu et le modèle prédira qu'elles auront une très faible probabilité de sélection. Inversement, l'ensemble des

[5] Certains cas sont présentés dans Luce et Suppes (1965) et Manski (1973).

alternatives possibles peut être restreint à celles réellement pertinentes et les autres peuvent être omises sans impact significatif sur le modèle.

Cette réduction à l'ensemble pertinent, i.e., les alternatives dont la probabilité de sélection est non-négligeable, a des avantages pratiques puisqu'elle permet d'importantes économies au niveau de la collection de données et en temps de calcul. Or, déterminer les alternatives pertinentes crée certains problèmes. Une solution est de tenter d'identifier les alternatives apparentes à partir de prises de décisions de voyages. Une autre est d'identifier, durant la collecte des données, les alternatives disponibles non-choisies en plus de celles qui ont été choisies.

L'indépendance des alternatives non-pertinentes (Luce, 1959) implique que la comparaison entre deux alternatives n'est pas influencée par les autres alternatives disponibles. En termes de probabilités, l'idée est que le ratio de la probabilité de choisir une alternative versus la probabilité d'en choisir une autre est indépendant de l'ensemble d'alternatives disponibles. Si deux sous-ensembles CA_i et CB_i où $CA_i \subseteq CB_i \subseteq C_i$ alors pour deux alternatives a et b appartenant à CA_i, le ratio suivant est obtenu:

$$\frac{\Pr_{CA_i}(a)}{\Pr_{CA_i}(b)} = \frac{\Pr_{CB_i}(a)}{\Pr_{CB_i}(b)} \tag{2.6}$$

où le ratio des probabilités est uniquement fonction des caractéristiques de deux alternatives. Bien que cette propriété soit désirable en général, dans certains cas, des difficultés d'application pour fin de prévision peuvent survenir[6]. Cette limitation est souvent illustré par le paradoxe autobus rouge/autobus bleu dans le contexte du choix modal (Ben-Akiva et Lerman, 1985). Bierlaire (1997) considère l'exemple suivant:

[6] Les ensembles d'alternatives pour fins d'estimation et de prévision ne doivent pas nécessairement être identiques mais, il semble souhaitable qu'une certaine correspondance existe entre les deux ensembles.

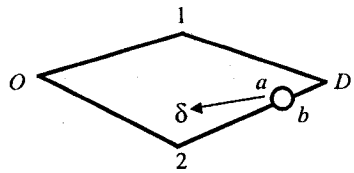


Figure 2.1: Exemple de chemins IIA.

où un voyageur i doit choisir un chemin pour se déplacer de l'origine O à la destination D. Il a le choix parmi trois chemins, i.e., $C_i = \{1,2a,2b\}$. Le seul attribut considéré est le temps de trajet (TT) et il est supposé le même pour chaque alternative, i.e., $V_1 = V_2 = V_3 = TT$. Finalement, le temps de trajet δ sur les courtes sections a et b est significativement plus petit que TT. Intuitivement, le résultat espéré est que la probabilité de choisir le chemin 1 ou 2 soit autour de 50%, indépendamment du choix entre a et b. Or, les probabilités fournies par le modèle pour cet exemple sont:

$$\Pr_{C_i}(1) = \Pr_{C_i}(2) = \Pr_{C_i}(3) = \frac{e^{TT}}{\sum_{b \in C_i} e^{TT}} = \frac{1}{3}. \tag{2.7}$$

Ce résultat ne concorde pas avec le résultat intuitif. Cette situation apparaît souvent dans les problèmes de choix où les alternatives sont corrélées, comme dans ce cas-ci. Les alternatives $2a$ et $2b$ sont si similaires que leurs utilités partagent plusieurs attributs inobservés du chemin et donc, l'hypothèse d'indépendance de la partie stochastique de ces utilités n'est pas valide dans ce contexte. Donc contrairement à toute attente, l'introduction de nouvelles alternatives peut réduire la part des alternatives existantes de proportion égale. Le problème peut être plus prononcé si la nouvelle alternative n'est pas une alternative entière indépendante (Charles River Associates, 1972). En pratique, par contre, la plupart de ces difficultés peuvent être surmontées en éliminant *a priori* (ou en combinant) certaines des alternatives si elles ne sont pas réellement indépendantes ou en segmentant le marché. Le modèle logit imbriqué corrige en partie cette limitation du modèle multinomial logit ordinaire.

2.1.4 formulation imbriquée (*nested*)

Le modèle logit imbriqué est une extension du logit multinomial proposé par Ben-Akiva (1973, 1974) pour détecter les corrélations entre les alternatives. Il est basé sur la division de l'ensemble de choix C_i en M nids C_{mi} tel que:

$$C_i = \bigcup_{m=1}^{M} C_{mi} \text{ et } C_{mi} \cap C_{li} = \phi \,; \forall m \neq l \,. \tag{2.8}$$

La fonction d'utilité de chaque alternative est composée d'un terme spécifique à l'alternative et un terme associé au nid. Si $a \in C_{mi}$,

$$U_a^i = V_a^i + \varepsilon_a^i + V_{C_{mi}}^i + \varepsilon_{C_{mi}}^i \,. \tag{2.9}$$

Les termes d'erreur ε_a^i et $\varepsilon_{C_{mi}}^i$ sont indépendants. Comme pour le modèle multinomial logit, les ε_a^i sont indépendantes et distribuées de façon identique selon une distribution de Gumbel avec paramètre d'échelle μ_m (qui peut être différente pour chaque nid). La distribution de $\varepsilon_{C_{mi}}^i$ est telle que la variable aléatoire $\max_{a \in C_i} U_a^i$ suit la distribution de Gumbel avec paramètre d'échelle μ. Chaque nid m à l'intérieur d'un ensemble C_i est associé à une utilité composée (ou log-somme), $W_{C_{mi}}^i$, qui s'écrit:

$$W_{C_{mi}}^i = V_{C_{mi}}^i + \frac{1}{\mu_m} \ln \sum_{b \in C_{mi}} e^{\mu_m V_b^i} \,. \tag{2.10}$$

La probabilité d'un individu i de choisir l'alternative a au sein du nid C_{mi} est:

$$\Pr_{C_i}^i(a) = \Pr_{C_i}^i(C_{mi}) \Pr_{C_{mi}}^i(a) = \frac{e^{\mu W_{C_{mi}}^i}}{\sum_{l=1}^{M} e^{\mu W_{C_{li}}^i}} \frac{e^{\mu_m V_a^i}}{\sum_{b \in C_{mi}} e^{\mu_m V_b^i}} \,. \tag{2.11}$$

Les paramètres μ et μ_m reflètent la corrélation parmi les alternatives d'un même nid C_{mi}. La covariance entre l'utilité de deux alternatives a et b au sein du nid C_{mi} est:

$$\operatorname{cov}\left(U_a^i, U_b^i\right) = \begin{cases} \operatorname{var}\left(\varepsilon_{C_{mi}}^i\right), & \text{si } a,b \in C_{mi} \\ 0, & \text{sinon} \end{cases} \tag{2.12}$$

et la corrélation est:

$$\operatorname{corr}\left(U_a^i, U_b^i\right) = \begin{cases} 1 - \mu^2/\mu_m^2, & \text{si } a,b \in C_{mi} \\ 0, & \text{sinon} \end{cases}. \tag{2.13}$$

Puisque la corrélation est non-négative, $0 \le \mu/\mu_m \le 1$ et $\mu/\mu_m = 1 \Leftrightarrow \operatorname{corr}(U_a^i, U_b^i) = 0$. Les paramètres μ et μ_m sont fortement liés dans le modèle mais, au fait, seul leur ratio est important car il est impossible de les identifier séparément. En pratique, un de ces paramètres est fixé à 1.

En reprenant l'exemple précédent des trois chemins, l'ensemble $C_i = \{1,2a,2b\}$ est divisé en $C_{1i} = \{1\}$ et $C_{2i} = \{2a,2b\}$. Les composantes déterministes des utilités sont:

$$V_{C_{1i}}^i = TT \ , \ V_1^i = 0 \ , \ V_{C_{1i}}^i = TT - \delta \text{ et } V_{2a}^i = V_{2b}^i = \delta \, . \tag{2.14}$$

Les utilités composées pour chaque nid sont: $W_{C_{1i}}^i = V_{C_{1i}}^i = TT$ et

$$W_{C_{2i}}^i = TT - \delta + \frac{1}{\mu_2} \ln\left(e^{\mu_2 \delta} + e^{\mu_2 \delta}\right) = TT - \frac{1}{\mu_2} \ln(2). \tag{2.15}$$

La probabilité de choisir chacun des nids est:

$$\Pr_{C_i}^i\left(C_{1i}\right) = \frac{e^{W_{C_{1i}}^i}}{e^{W_{C_{1i}}^i} + e^{W_{C_{2i}}^i}} = \frac{e^{TT}}{e^{TT} + e^{TT - \frac{\ln 2}{\mu_2}}} = \frac{1}{1 + e^{\frac{\ln 2}{\mu_2}}} = \frac{1}{1 + 2^{1/\mu_2}} \tag{2.16}$$

et

$$\Pr_{C_i}^i\left(C_{2i}\right) = 1 - \Pr_{C_i}^i\left(C_{1i}\right) = 1 - \frac{1}{1 + e^{1/\mu_2}} = \frac{2^{1/\mu_2}}{1 + 2^{1/\mu_2}} \tag{2.17}$$

où, sans perte de généralité, la valeur de μ a été fixée à 1. Ainsi, $0 \le 1/\mu_2 \le 1$ et la probabilité de chaque alternative est:

$$\Pr_{C_i}^i(1) = \Pr_{C_i}^i\left(C_{1i}\right) = \frac{1}{1 + e^{1/\mu_2}} \tag{2.18}$$

et

$$\Pr_{C_i}^{i}(2a) = \Pr_{C_i}^{i}(2b) = \frac{e^{\mu_2 \delta}}{e^{\mu_2 \delta} + e^{\mu_2 \delta}} \Pr_{C_i}^{i}(C_{2i}) = \frac{1}{2}\left(\frac{2^{1/\mu_2}}{1+2^{1/\mu_2}}\right). \tag{2.19}$$

Lorsque $1/\mu_2 = 1$, le modèle logit imbriqué produit le même résultat que le logit multinomial et les probabilités sont 1/3. Par contre, lorsque μ_2 tend vers l'infini, $1/\mu_2$ tend vers 0 et la probabilité de chaque nid se rapproche de 1/2.

L'extension directe du modèle imbriqué consiste à diviser quelques nids (ou tous) en sous-nids qui, à leur tour, peuvent être divisés en sous-nids de façon à former un arbre. Le modèle est valide à chaque niveau et est généré de façon récursive. La corrélation entre les nids n'est pas détectée par le modèle et lorsque les alternatives ne peuvent être classées par nids, ce modèle n'est pas approprié.

2.1.5 formulation imbriquée croisée (*cross nested*)

Le modèle logit imbriqué croisé est une extension du logit imbriqué proposé par McFadden (1978) où chaque alternative peut appartenir à plus d'un nid. Il est aussi basé sur la division de l'ensemble de choix C_i en M nids C_{mi} mais pour chaque alternative a et pour chaque nid m, des paramètres α_{am} $(0 \le \alpha_{am} \le 1)$ représentant le degré d'appartenance de l'alternative a au nid m sont définis. L'utilité de l'alternative i est:

$$U_{am}^{i} = V_a^i + \varepsilon_a^i + V_{C_{mi}}^i + \varepsilon_{C_{mi}}^i + \ln \alpha_{am}. \tag{2.20}$$

Les termes d'erreur ε_a^i et $\varepsilon_{C_{mi}}^i$ sont indépendants. Les ε_a^i sont indépendantes et distribuées de façon identique selon une distribution de Gumbel avec paramètre d'échelle μ_m. Généralement, μ_m prend la valeur 1 pour faciliter l'écriture du modèle. La distribution de $\varepsilon_{C_{mi}}^i$ est telle que la variable aléatoire $\max_{a \in C_i} U_a^i$ suit la distribution de Gumbel avec paramètre d'échelle μ. Chaque nid m à l'intérieur d'un ensemble C_i est associé à une utilité composée (ou log-somme), $W_{C_{mi}}^i$, qui s'écrit:

$$W_{C_{mi}}^{i} = V_{C_{mi}}^{i} + \ln \sum_{b \in C_{mi}} \alpha_{bm} e^{V_b^i} . \tag{2.21}$$

La probabilité d'un individu i de choisir l'alternative a est:

$$\Pr_{C_i}^{i}(a) = \sum_{l=1}^{M} \Pr_{C_i}^{i}(C_{mi}) \Pr_{C_{mi}}^{i}(a) = \sum_{l=1}^{M} \frac{e^{\mu W_{C_{mi}}^{i}}}{\sum_{l=1}^{M} e^{\mu W_{C_{li}}^{i}}} \frac{\alpha_{am} e^{V_a^i}}{\sum_{b \in C_{mi}} \alpha_{bm} e^{V_b^i}} . \tag{2.22}$$

La prochaine section explique comment estimer ces modèles. Dans les trois cas de modèles, la même procédure est utilisée.

2.2 Méthode d'estimation

L'estimation se fait à partir d'un échantillon fini d'individus faisant un choix et ce sont les choix actuels qui sont observés, non les probabilités. Avec ces données désagrégées, la variable dépendante représente le choix et prend la valeur associée à l'alternative[7] dans C_i. Par exemple, si $C_i = \{1,2a,2b\}$ et le consommateur i a choisi $2a$, la variable dépendante prendra la valeur 2 pour cette observation (et 3 pour $2b$). Les variables indépendantes, quant à elles, peuvent être continues, discrètes ou catégoriques.

2.2.1 maximum de vraisemblance

La méthode du maximum de vraisemblance est utilisée pour estimer les coefficients du modèle. L'estimateur des β_k dans un modèle logit, tel que celui décrit à l'équation (2.5), et ses propriétés sont expliqués par McFadden (1968). L'interprétation de ces β_k se fait en relation avec une alternative de base (normalisée) lorsque leurs attributs associés caractérisent l'individu et pour ASC qui est la constante spécifique à l'alternative. Dans le cas de β_k reliés aux attributs qui caractérisent l'alternative, ils s'interprètent comme des poids influençant la valeur de d'utilité de cette alternative.
La fonction de vraisemblance s'écrit:

[7] La variable dépendante prend la valeur 0 ou 1 dans le cas du modèle logit binomial.

$$L = \prod_{i=1}^{N} \prod_{b \in C_i} \Pr_{C_i}^{i}(b)^{g_b^i} \tag{2.23}$$

où N est le nombre d'observations et g_b^i égale 1 si l'alternative b a été choisie dans l'observation i et 0 sinon. Puisque la fonction vraisemblance est le produit de probabilités, sa valeur se situe aussi entre 0 et 1 et le logarithme naturel de la fonction vraisemblance doit toujours être négatif. En appliquant le logarithme des deux côtés de (2.23),

$$\ln L = L^* = \sum_{i=1}^{N} \sum_{b \in C_i} g_b^i \ln \Pr_{C_i}^{i}(b) \tag{2.24}$$

et le problème de maximisation devient:

$$\max_{\beta} L^*. \tag{2.25}$$

En substituant $\Pr_{C_i}^{i}(b)$ de l'équation (2.5), les conditions de première ordre[8] de L^* sont:

$$\frac{\partial L^*}{\partial \beta_k} = \sum_{i=1}^{N} \sum_{b \in C_i} \left(g_b^i - \Pr_{C_i}^{i}(b)\right) x_b^i(k) = 0 \; ; \quad \forall k \in K. \tag{2.26}$$

Les K équations en (2.26) sont non-linéaires et leur solution requiert une méthode itérative, telle Newton-Raphson. McFadden (1968) démontre que, sauf lorsque certaines conditions spécifiques existent dans les données, le maximum de L^* obtenu de (2.26) est unique et noté $L^*(\hat{\beta})$. Cet estimateur a des propriétés asymptotiques optimales. Lorsque le nombre d'observations augmente, il devient possible de montrer que l'estimateur s'approche de l'optimalité, i.e., approche la variance minimale dans l'estimation de chaque coefficient. La matrice asymptotique de variance-covariance est l'inverse de la matrice de dérivées secondes de L^* multiplié par -1 (Theil, 1971).

[8] Ces conditions sont expliquées plus en détail par McFadden (1968).

2.2.2 qualité du modèle

Puisque les probabilités ne sont pas observées directement, il serait trompeur de comparer les probabilités calculées avec les variables g_b^i. Donc, une mesure de qualité du modèle (*goodness of fit*) basée sur les erreurs de régression (*estimation residuals*), telle le R^2 dans les moindres-carrés ordinaires, n'est pas appropriée. De plus, une comparaison de la somme des probabilités pour une alternative, $\sum_{i=1}^{N} \Pr_{C_i}^i(b)$, versus le nombre total d'observations où elle a été choisie, $\sum_{i=1}^{N} g_b^i$, est aussi trompeuse puisque si l'ensemble des variables inclut une constante spécifique à l'alternative, i.e.,

$$ASC = \begin{cases} \text{constante}, & a = b \\ 0, & a \neq b \end{cases} \tag{2.27}$$

alors, à partir ses conditions de premier-ordre à l'équation (2.26), l'égalité suivante est toujours respectée:

$$\sum_{i=1}^{N} g_b^i = \sum_{i=1}^{N} \Pr_{C_i}^i(b). \tag{2.28}$$

Puisqu'il est impossible d'estimer les erreurs de régression dans un modèle logit multinomial, il n'existe pas de statistique R^2 qui indiquerait à quel point le modèle se conforme aux données. Il est cependant possible de définir une mesure similaire au R^2 pour comparer différents modèles. Cette mesure, notée ρ^2, est basée sur la valeur de la fonction log-vraisemblance et s'écrit:

$$\rho^2 = 1 - \frac{L^*(\hat{\beta})}{L^*(0)} \tag{2.29}$$

où $L^*(\hat{\beta})$ est la valeur de L^* pour le vecteur de coefficients estimés et $L^*(0)$ est la valeur L^* lorsque $\beta = 0$. Puisque le procédé de maximiser la vraisemblance se résume à augmenter $L^*(\hat{\beta})$ à partir d'un nombre initial très négatif, $L^*(0)$, à un nombre se rapprochant le plus possible de 0, ρ^2 est égal au ratio de la log-vraisemblance expliquée sur la log-vraisemblance totale. La valeur de ρ^2 se situe entre 0 et 1. Cette mesure a, en fait, les mêmes lacunes que R^2. Parmi une de celles-ci, elle ne tient pas

compte des degrés de liberté (nombre de paramètres à estimer) que comporte le modèle. En les incluant, (2.29) devient:

$$\bar{\rho}^2 = 1 - \frac{L^*(\hat{\beta}) - K}{L^*(0)} \tag{2.30}$$

où K est le nombre total de variables spécifiées.

Poser $\beta = 0$ dans l'équation (2.5) revient à faire l'hypothèse que les alternatives sont équiprobables. Lorsque toutes les fonctions d'utilité tendent vers 0 et les C_i sont les mêmes pour chaque individu, la probabilité de n'importe quelle alternative tend vers $1/C_i$ puisque $e^0 = 1$. Dans ce cas, $\bar{\rho}^2$ tend aussi vers 0 et le modèle n'arrive pas à capter l'impact des attributs sur les choix probabilistes des individus de l'échantillon.

2.2.3 élasticité

L'élasticité directe est généralement définie comme étant le changement en pourcentage que subit la variable dépendante lorsque que la variable indépendante varie d'un pour cent. Dans ce cas-ci, c'est le changement en pourcentage de la probabilité de sélection de l'alternative a suite à un changement de un pour cent de la valeur d'un des attributs de sa propre fonction d'utilité $x_a^i(k)$. L'élasticité[9] directe s'écrit:

$$E_{x_a^i(k)}^{\Pr_{C_i}^i(a)} = \frac{\partial \Pr_{C_i}^i(a) / \Pr_{C_i}^i(a)}{\partial x_a^i(k) / x_a^i(k)} = \frac{\partial \Pr_{C_i}^i(a)}{\partial x_a^i(k)} \frac{x_a^i(k)}{\Pr_{C_i}^i(a)}. \tag{2.31}$$

Elle varie de 0 lorsque la probabilité du choix est 1, à $\beta_k x_a^i(k)$ lorsque la probabilité du choix est 0. Lorsque la variable entre dans le modèle sous forme logarithmique, la limite supérieure de l'élasticité sera une constante prenant la valeur du coefficient β_k. Si $x_a^i(k)$ est une variable de prix, par exemple, une valeur négative de l'élasticité directe est prévue et β_k devrait être négatif.

---

[9] Si la variable $x_a^i(k)$ entre dans le modèle sous-forme logarithmique, $\ln x_a^i(k)$, alors les élasticités sont telles que définies ci-haut, divisées par $x_a^i(k)$.

Puisque la probabilité de choisir une alternative i est fonction de l'utilité de chacune des autres alternatives, il est aussi possible de calculer le changement en % dans la probabilité de sélection de l'alternative b résultant d'un changement de un pour cent de la valeur d'un des attributs d'une autre alternative, $x_b^i(k)$. Cette élasticité croisée s'écrit:

$$E_{x_b^i(k)}^{\Pr_{C_i}^i(a)} = \frac{\partial \Pr_{C_i}^i(b)/\Pr_{C_i}^i(b)}{\partial x_b^i(k)/x_b^i(k)} = \frac{\partial \Pr_{C_i}^i(b)}{\partial x_b^i(k)} \frac{x_b^i(k)}{\Pr_{C_i}^i(b)}. \tag{2.32}$$

Puisque β_k devrait être négatif pour une variable de prix, l'élasticité croisée sera positive. L'élasticité croisée dépend uniquement des valeurs reliées à l'alternative b, ce qui veut dire que les élasticités croisées de toutes les alternatives par rapport à un attribut de l'alternative b sont les mêmes. Dans le modèle logit multinomial (2.5), la relation suivante est obtenue à partir des équations (2.31) et (2.32):

$$E_{x_a^i(k)}^{\Pr_{C_i}^i(a)} = \left(1 - \Pr_{C_i}^i(a)\right)\beta_k x_a^i(k) \text{ et } E_{x_b^i(k)}^{\Pr_{C_i}^i(b)} = -\Pr_{C_i}^i(b)\beta_k x_b^i(k) \tag{2.33}$$

et, de façon générale,

$$E_{x_b^i(k)}^{\Pr_{C_i}^i(b)} = \left(\delta - \Pr_{C_i}^i(b)\right)\beta_k x_b^i(k) \text{ où } \delta = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}. \tag{2.34}$$

En conclusion, les élasticités du modèle logit multinomial permet d'extraire beaucoup information sur l'interaction des attributs entre-eux et leur impact sur la probabilité de sélection des alternatives. Elles permettent de comparer plusieurs scénarios et de quantifier la sensibilité des consommateurs aux changements de valeurs des attributs. La prochaine section présente quelques extensions de ces modèles.

2.3 Extensions des modèles de choix discret

Cette section présente quelques extensions de modèles de choix discret. Les définitions et la classification des principaux modèles sont présentées. Ce mémoire utilise les modèles logit multinomial, logit multinomial imbriqué et logit multinomial imbriqué croisé. Les autres modèles apparaissent uniquement pour fins de références.

Tableau 2.1: Classes de modèles de choix-discret (probits).

PROBIT: basé sur la distribution normale issue du théorème central-limite.

- Habileté à détecter toutes les corrélations qui existent entre les alternatives.
- Formulation hautement complexe ce qui en fait un modèle peu appliqué.
- Les erreurs suivent la distribution normale.

Generalized Factor Analytic Specification of Random Utility (GFASRUM)

- Modèle général dont la spécification permet de tenir compte de cas spéciaux tels:
  - hétéroscédasticité, i.e., les variances varient entre les alternatives;
  - attribut analytique, i.e., les écarts-types sont spécifiques à l'attribut;
  - procédé général autoregressif, i.e., les erreurs sont autocorréllées.

Logit Mixte (hybrid)

- Réduit l'écart entre les familles Probit et Logit en prenant les avantages de chacun.
- Peut être combiné avec la représentation d'attributs analytiques pour pouvoir être estimé à l'aide du maximum de vraisemblance simulé (Ben-Akiva et Bolduc, 1996).

Choix de Classe Latente: détecte l'hétérogénéité qui n'est pas observée.

- Cette hétérogénéité résulte de constructions discrètes non-observables et représentées par des classes latentes (Swait et Ben-Akiva, 1987).

Source: Ben Akiva et Bierlaire (1999), pp.18-23.

Tableau 2.2: Classes de modèles de choix-discret (logits).

LOGIT: basés sur la fonction de distribution de probabilité du maximum d'une série de variables aléatoires (Gumbel, 1958).

- Forte tractabilité mais impose des restrictions sur la structure de la covariance. Les autres modèles dans cette famille (qui en découlent) tentent de préserver cette tractabilité tout en relaxant certaines restrictions.
- Introduit dans un contexte de choix binaire où la distribution logistique est utilisée.

Logit Multinomial

- Généralisation du modèle logit à plus de deux alternatives.
- L'hypothèse sur les erreurs des fonctions d'utilité est qu'elles sont indépendantes et distribuées de façon identique suivant la distribution de Gumbel.
- Souffre de la propriété d'indépendance des alternatives non-pertinentes qui limite son utilisation dans certaines applications pratiques.

Logit Multinomial Imbriqué (*nested*)

- Extension du logit multinomial proposé par Ben-Akiva (1973, 1974) pour détecter les corrélations entre les alternatives.
- Basé sur la division de l'ensemble de choix C_i en M nids C_{mi} tel que:

$$C_i = \bigcup_{m=1}^{M} C_{mi} \text{ et } C_{mi} \cap C_{m'i} = \phi \ ; \ \forall m \neq m'.$$

- L'extension directe du modèle imbriqué consiste à diviser quelques nids (ou tous) en sous-nids qui, à leur tour, peuvent être divisés en sous-nids de façon à former un arbre. Le modèle est valide à chaque niveau et est généré de façon récursive.
- La corrélation entre les nids n'est pas détectée par le modèle et lorsque les alternatives ne peuvent être classées par nids, ce modèle n'est pas approprié.

Logit Multinomial Imbriqué Croisé (*cross-nested*)

- Extension du modèle imbriqué où chaque alternative peut appartenir à plusieurs nids.
- En plus de C_{mi} pour chaque nid m, les paramètres α_{am} ($0 \leq \alpha_{am} \leq 1$) représentant le degré d'appartenance de l'alternative a au nid m sont définis.
- McFadden (1978) est le premier à présenter ce modèle comme cas particulier du modèle de valeur extrême généralisée.

Valeur Extrême Généralisée (*GEV*)

- Obtenu à partir du modèle d'utilité aléatoire développé par McFadden (1978).
- Modèle général au sens où il regroupe les modèles décrits ci-haut.

Source: Ben Akiva et Bierlaire (1999), pp.10-18.

CHAPITRE 3: TRANSPORT AÉRIEN

Plusieurs études récentes ont utilisé les modèles de choix discret et les ont appliqués à différents domaines du transport aérien. Dans ce chapitre, les grandes idées de huit d'entre elles sont résumées. Elles sont classées selon l'attribut principal dans le choix probabiliste des passagers. À la section 3.1, la décision du passager consiste généralement à choisir un aéroport de départ, un transporteur ou un vol. Parfois, des nids sont construits pour représenter des combinaisons d'éléments et le passager choisi donc plusieurs éléments à la fois. À la section 3.2, le passager choisit parmi un ensemble d'itinéraires.

3.1 Contextes variés

3.1.1 moment de la journée

Dans la littérature sur le transport en général, la demande selon le moment de la journée joue un rôle primordial sur l'ordonnancement des systèmes de transports urbains à horaires-fixes tels l'autobus et le train de banlieue. Similairement, une certaine valeur est attribuée aux horaires de vols dans le sens où les vols qui décollent (ou atterrissent) à l'intérieur de certains créneaux horaires sont préférés à d'autres. Pour un passager, la désutilité liée aux perturbations à l'horaire en transport urbain versus aérien n'est pas comparable. Si, suite à une annulation, un passager doit attendre une demi-heure pour le prochain autobus ou jusqu'au lendemain pour le prochain avion, l'impact en terme de désutilité ne sera pas le même. De plus, si le moment de la journée où devait s'effectuer un itinéraire est déplacé hors du créneau idéal, la désutilité augmente de façon non-linéaire et asymétrique. L'effet risque d'être plus prononcé chez les passagers d'affaires.

Walker et Parker (2006) modélisent la demande en fonction de l'heure des départs des vols à l'aide d'un modèle logit mixte. La fonction d'utilité est représentée en temps continu, ce qui offre plus de flexibilité que d'autres méthodes. À partir d'une

heure de départ idéale pour le passager (τ_i) versus offerte par le transporteur (t_a), ils formulent la probabilité du passager i de choisir le vol a, $a \in C_i$. L'intervalle [0, r] est un intervalle de temps (un jour ou une semaine), donc:

$$\Pr_{C_i}^{i}(a) = \int_0^r \frac{e^{V^i_{a|\tau_i=t}}}{\sum_{b\in C_i} e^{V^i_{b|\tau_i=t}}} f_i(t)dt = \int_0^r \frac{e^{\beta_t G_i(t-t_a)}}{\sum_{b\in C_i} e^{\beta_t G_i(t-t_b)}} f_i(t)dt \tag{3.1}$$

où β_t sont les coefficients à estimer et la fonction de désutilité G est définie en 3.2. Cette distribution logit mixte est constituée d'un modèle logit avec le paramètre τ_i qui est associé avec la fonction de densité $f_i(t)$. Donc, si τ n'est pas observable directement chez le passager i mais la fonction de distribution $f_i(t)$ l'est alors, τ_i est une variable latente.

L'effet de cette formulation est qu'elle sépare le problème en deux parties: la première étant de trouver la fonction f appropriée pour la distribution du moment de la journée idéal et la deuxième étant de trouver une fonction appropriée G pour la désutilité de la différence entre l'heure de départ idéale et actuelle. Les auteurs utilisent:

$$G(\tau_i - t_a) = \begin{cases} (\lambda_E)^{-1}\beta_E(t_a - \tau_i)^{\lambda_E}, & \tau_i - t_a < 0 \\ 0, & \tau_i = t_a \\ (\lambda_L)^{-1}\beta_L(\tau_i - t_a)^{\lambda_L}, & \tau_i - t_a > 0 \end{cases} \tag{3.2}$$

où deux paramètres (λ_E, λ_L) d'une transformation Box-Cox sont estimés pour chacune des fonctions, i.e., les départs en avance (E) ou en retard (L).

Leurs données proviennent de sondages faits par Boeing à l'automne 2005 et considèrent les itinéraires allers-retours seulement. Des données démographiques sont utilisées, notamment: l'âge, le sexe, le revenu (*income*), la profession (*occupation*), le niveau d'éducation et le code postal (*zip code*). De plus, les variables de circonstance (*situational variables*) identifiées sont les suivantes: (1) les dates et heures du voyage, (2) l'importance en faveur l'heure de départ ou d'arrivé et les heures idéales de ceux-

ci, (3) le but du voyage (affaires ou vacances), (4) qui paye pour le voyage (le passager, l'employeur qui rembourse, l'employeur directement) et (5) la taille du groupe voyageant ensemble. Finalement, parmi les variables indépendantes, les auteurs utilisent: le transporteur, le type d'avion, le tarif, le nombre d'escales, la durée du voyage, l'espace pour les jambes (*legroom*) et les différences entre les heures de départ et d'arrivée, idéales et actuelles (*schedule delay*).

3.1.2 région à aéroports multiples

Brooke *et al.* (1994) développent une méthodologie pour générer des voyages dans la région des Midlands (en Grande-Bretagne) et les distribuer parmi plusieurs aéroports. Leur méthodologie est basée sur la propension à voler d'une population segmentée géographiquement et par la nature de leurs déplacements. En estimant le taux de croissance de cette propension, des prévisions pour l'année en cours sont obtenues. La distribution se fait au moyen d'un modèle logit multinomial où la probabilité d'un individu i de choisir l'aéroport a pour se rendre à destination d dépend du temps d'accès à l'aéroport depuis son domicile, $x(1)$, de la fonction d'utilité de la fréquence, $x(2)$, du tarif aller-retour, $x(3)$, et d'une variable binaire désignant le type d'avion, $x(4)$. Ainsi:

$$\Pr_{C_i}^{i}(a) = \frac{e^{V_a^i}}{\sum_{b \in C_i} e^{V_b^i}} = \frac{e^{\beta_1 x_a^i(1) + \beta_2 x_a^i(2) + \beta_3 x_a^i(3) + \beta_4 x_a^i(4)}}{\sum_{b \in C_i} e^{\beta_1 x_b^i(1) + \beta_2 x_b^i(2) + \beta_3 x_b^i(3) + \beta_4 x_b^i(4)}}. \tag{3.3}$$

La fonction d'utilité de la fréquence est basée sur le délai à l'horaire et sur l'hypothèse de demande constante pour tous les vols au courant de la journée. La fréquence contribue le plus à l'utilité et s'écrit: $(x - 0.5)/x$ où x est le nombre de vols par jour de la semaine.

Pels *et al.* (1998) appliquent le modèle logit imbriqué (3.5) pour analyser la compétition entre aéroports et transporteurs qui y opèrent. Ils cherchent un équilibre unique où les tarifs, les taxes d'aéroport et les fréquences des vols sont optimaux. Ils trouvent que lorsque l'élasticité de la demande par rapport à la fréquence est

inférieure à 1, un équilibre tarif-fréquence et taxe d'aéroport unique existe. Le transporteur a, $a \in C_{mi}$, maximise ses profits par route (π_a) à partir de l'aéroport m, $m \in M$, en fonction du tarif optimal (p_a) et de la fréquence optimale sur cette route (f_a), i.e.,

$$\pi_a = (p_a - c_a) N \Pr^i_{C_{mi}}(a) - k_a f_a - K_a, m \in M, a \in C_{mi} \tag{3.4}$$

où K_a, c_a et k_a sont des coûts fixes et marginaux, N est le nombre total de passagers sur la route et $\Pr^i_{C_{mi}}(a)$ est la probabilité de choisir la combinaison (aéroport m, transporteur i). Le modèle logit imbriqué s'écrit:

$$\Pr^i_{C_i}(a) = \Pr^i_{C_i}(C_{mi}) \Pr^i_{C_{mi}}(a) \tag{3.5}$$

où

$$\Pr^i_{C_{mi}}(a) = e^{\frac{\alpha_a - \alpha_p p_a + \alpha_f \ln(f_a)}{\mu}} \left(\sum_{a \in C_{mi}} e^{\frac{\alpha_{a'} - \alpha_p p_a + \alpha_a \ln(f_a)}{\mu}} \right)^{-1} \tag{3.6}$$

$$\Pr^i_{C_i}(C_{mi}) = e^{\frac{\beta_m - \beta_t tax_m + \beta_r \ln(t_m) + \mu \ln W^i_{C_{mi}}}{\theta}} \left(\sum_{m \in M} e^{\frac{\beta_m - \beta_t tax_{m'} + \beta_r \ln(t_m) + \mu \ln W^i_{C_{mi}}}{\theta}} \right)^{-1} \tag{3.7}$$

et

$$W^i_{C_{mi}} = \sum_{a \in C_{mi}} \left(\frac{\alpha_{a'} - \alpha_p p_a + \alpha_f \ln(f_a)}{\mu} \right). \tag{3.8}$$

Les paramètres α_p, α_f, β_t, et β_r sont plus grands que 0; α_a, $\alpha_{a'}$ et β_m sont libres. Le paramètre μ représente le degré d'hétérogénéité parmi les transporteurs (vols) à partir d'un aéroport alors que θ est celui pour les aéroports. Plus μ se rapproche de 0, plus le degré de substitution entre les transporteurs est élevé. Ainsi, il se doit que $\theta > \mu$ (Anderson *et al.*,1996) pour que les transporteurs opérant d'un même aéroport soient des substituts plus rapprochés que ceux opérant à partir d'aéroports différents. Les conditions de premier-ordre de (3.4) par rapport au tarif et à la fréquence s'écrivent:

$$p_a = c_a + \frac{\theta\mu}{\alpha_p\left(\theta\left(1-\Pr^i_{C_{mi}}(a)\right)+\mu\Pr^i_{C_{mi}}(a)\left(1-\Pr^i_{C_i}(C_{mi})\right)\right)} \tag{3.9}$$

et

$$f_a = \frac{\alpha_p N \Pr^i_{C_i}(a)(p_a - c_a)}{k}\left(\frac{\theta\left(1-\Pr^i_{C_{mi}}(a)\right)+\mu\Pr^i_{C_{mi}}(a)\left(1-\Pr^i_{C_i}(C_{mi})\right)}{\theta\mu}\right). \tag{3.10}$$

L'aéroport maximise le bien-être social (*social welfare*) en fonction de la taxe d'aéroport (tax_m) sous une condition de recouvrement des coûts, i.e.,

$$\begin{aligned} \max_{tax_m} \quad & \int_{tax_m} \Pr^i_{C_{mi}}(a)dtax + (tax_m - mc_m)N\Pr^i_{C_{mi}}(a) - rK_m - g\left(\frac{N\Pr^i_{C_{mi}}(a)}{K_m}\right) \\ \text{s.c.} \quad & \Pi_m \equiv (tax_m - mc_m)N\Pr^i_{C_{mi}}(a) - rK_m > 0 \end{aligned} \tag{3.11}$$

où rK_m et mc_m sont des coûts fixes et marginaux alors que $g\left(N\Pr^i_{C_{mi}}(a)K_m^{-1}\right)$ est une fonction de coût externe; $\frac{\partial g\left(N\Pr^i_{C_{mi}}(a)K_m^{-1}\right)}{\partial \Pr^i_{C_{mi}}(a)} > 0$. Les conditions de premier-ordre de (3.11) par rapport à la taxe d'aéroport donnent:

$$tax_m = mc_m + \frac{rK_m}{N\Pr^i_{C_{mi}}(a)}. \tag{3.12}$$

En partant du principe que pour rendre un aéroport plus accessible, la taxe d'aéroport doit être réduite, la hausse de l'achalandage mène à une redistribution des passagers parmi les transporteurs. Si la hausse de tarifs du transporteur dominant surpassait la hausse de fréquences de vols et la baisse de tarifs d'un compétiteur, l'accès plus facile à un aéroport pourrait même mener à une chute de la part de marché de cet aéroport. Le transporteur ayant haussé ces tarifs profiterait ainsi de la hausse de l'achalandage et de l'accessibilité.

Plus récemment, Hess et Polak (2006) utilisent un modèle logit imbriqué croisé qui permet une représentation jointe de la corrélation entre les alternatives selon trois dimensions de choix, i.e., l'aéroport m, $m \in M$, le transporteur a, $a \in C$, et le mode d'accès o, $o \in O$, évitant ainsi les contraintes associées aux modèles dont la structure

est multi-étages. Par exemple, dans le cas de d+1 dimensions, le modèle logit imbriqué à d+1 niveaux ne peut être utilisé que pour analyser la corrélation le long d'au plus d des d+1 dimensions de choix[10]. En ajoutant un niveau, les nids au dernier niveau ne contiennent qu'une seule alternative et les paramètres structurels de ceux-ci s'annuleraient lorsque additionnés ensemble. Autrement dit, le dernier niveau d'emboîtement devient redondant. L'autre contrainte qui lui est associée vient du fait que la pleine étendue de la corrélation peut uniquement être prise en compte le long d'une dimension. Une quantité limitée de la corrélation peut être prise en compte le long de la seconde dimension.

Leur modèle logit imbriqué croisé est spécifié en définissant trois groupes de nids, i.e., M nids selon l'aéroport, C nids selon le transporteur et O nids selon le mode d'accès, et en permettant à chaque alternative d'appartenir à exactement un nid dans chacun de ces groupes. λ_m, π_a, et ψ_o sont les paramètres structurels mais les paramètres d'allocation, gouvernant la proportion selon laquelle une alternative appartient à chacun des trois ensembles de nids, ne sont pas montrés.

Les données utilisées proviennent d'un sondage sur les passagers mené par le *Civil Aviation Authority* britannique en 1996 et sont divisées en quatre sous-groupes: résidents-d'affaires, résidents-vacanciers, visiteurs-d'affaires et visiteurs-vacanciers. Dans chaque sous-groupe, un sous-échantillon de 95% est utilisé pour calibrer le modèle et le 5% restant est utilisé pour valider le modèle obtenu. L'intérêt porte sur le comportement des passagers quittant à bord de vols directs (*departing*) et ce, uniquement pour les destinations desservies par un seul des aéroports majeurs à l'étude. Au point de vue des caractéristiques de service, ils utilisent les fréquences de vol, les tarifs, les blocs-heures (*block times*) pour incorporer la congestion des aéroports, les types d'avions utilisés, la capacité de sièges disponibles à bord ainsi que la ponctualité (*on-time performance*) du transporteur et de l'aéroport. Une analyse préliminaire vise à déterminer les variables explicatives qui bénéficieraient

[10] Le premier niveau étant la racine.

d'une spécification non-linéaire et une transformée logarithmique leur est appliquée, au besoin.

Les résultats des auteurs révèlent que le comportement du passager est grandement influencé par le temps et le coût d'accès à l'aéroport ainsi que la fréquence et l'horaire de vols. De plus, une comparaison structurelle des modèles montre que le logit imbriqué croisé constitue une amélioration par rapport au logit imbriqué qui surpasse le logit multinomial utilisé à la base.

3.1.3 comportement des passagers

Deux études de Garrow et Koppelman (2004a, 2004b) portent sur le comportement des passagers en attente (*standby*) et de ceux qui ne se présentent pas au vol (*no-show*). Dans la première étude, ils utilisent des données désagrégées au niveau du passager et de son itinéraire directionnel, i.e., vol-sortant (*outbound*) et vol-entrant (*inbound)*, dans un modèle logit multinomial. Ils montrent qu'en distinguant les itinéraires entrants des itinéraires sortants, ils améliorent leurs prévisions. Dans la seconde étude, ils introduisent un modèle logit imbriqué qui tient compte des opportunités des passagers en attente arrivant en avance aux aéroports d'être placés sur d'autres vols du même transporteur. L'information sur le niveau actuel des réservations sur d'autres vols ayant le même jour de départ est incorporée et les résultats démontrent l'importance de distinguer les vols-entrants des vols-sortants dans l'implémentation de leur modèle.

Puisque ce ne sont pas tous les passagers ayant une réservation qui voyagent au moment venu, les transporteurs surréservent pour réduire le nombre anticipé de sièges vides au décollage sur un vol où il y a de la demande pour ces sièges. Il est utile de se rappeler que même une amélioration minime de la prévision de no-shows peut se traduire par des millions de dollars en revenus annuels additionnels pour un transporteur majeur américain. Selon Garrow et Koppelman (2004a), la pratique courante des transporteurs pour prédire les taux de no-shows est d'utiliser des séries

chronologiques basées sur les taux historiques d'une classe tarifaire ou d'un compartiment. Ces modèles considèrent donc les différences entre les taux dues aux attributs spécifiques du vol tels l'heure de départ, le jour de la semaine, le mois, la capacité, l'origine, la destination, etc.

Les auteurs s'intéressent au comportement des passagers vis-à-vis un réordonnancement (*schedule change*) le jour du départ de la part du transporteur. Leurs variables proviennent de données corporatives d'un grand transporteur américain et couvrent les itinéraires domestiques pour mars 2001 et mars 2002. Elles sont basées sur les réservations actuelles du transporteur, les billets, l'horaire de vol, le type de membre dans le programme de fidélité et l'information à l'enregistrement. Quatre sous-groupes sont définis à partir d'itinéraires allers-retours sur deux périodes. Des tests de ratios de vraisemblance justifient cette démarche. Un modèle contraint combinant les données sortantes (vol-aller) et entrantes (vol-retour) pour mars 2001 est comparé à deux modèles non-contraints utilisant les données d'un seul des deux sous-groupes basés sur l'itinéraire. Le test de ratio de vraisemblance s'écrit:

$$LLR = -2\left(L^*\left(\hat{\beta}^{03/01}_{\text{tous}}\right) - L^*\left(\hat{\beta}^{03/01}_{\text{allers}}\right) - L^*\left(\hat{\beta}^{03/01}_{\text{retours}}\right)\right). \tag{3.13}$$

Cette statistique est distribuée χ^2 avec le nombre de degrés de liberté égal au nombre de contraintes. La même procédure est appliquée pour la période de mars 2002.

Garrow et Koppelman utilisent des modèles logit multinomials pour estimer la probabilité que chaque passager se présente, soit un no-show ou se place en attente sur un vol antérieur (ou ultérieur) du même transporteur. Le niveau de service doit être égal ou supérieur à celui de l'itinéraire initial pour considérer l'option d'un autre vol. Les résultats démontrent que les débarquements (*denied boardings*) sont particulièrement dispendieux parce qu'un itinéraire ultérieur n'est pas nécessairement disponible et que le transporteur doit payer l'hôtel ou un de ses compétiteurs, généralement au tarif le plus élevé, pour accommoder le passager débarqué. De plus, les coefficients de durée du vol deviennent de plus en plus négatif lorsque la durée du

vol augmente indiquant ainsi que les passagers sont moins enclins à se placer en attente pour un itinéraire antérieur lorsque le vol est long.

3.1.4 programmes de fidélité

Lederman (2004) examine les programmes de fidélité et l'impact marginal qu'ils ont sur un transporteur dominant à l'aide d'un modèle logit imbriqué. Le gain marginal accru qui se retrouve dans le système de points de ces programmes incite les passagers à toujours voler avec le même transporteur domestique (ou un de ses partenaires) plutôt que de baser leur choix un vol à la fois. Ils préfèrent le transporteur dominant à leur aéroport local car celui-ci offre généralement les meilleures chances de ramasser et de réclamer des points qui constituent des mesures de valeur. Une fois que le passager adhère à un tel programme, chaque vol entrepris avec un autre transporteur représente un certain nombre de points auxquels il renonce. Ainsi, pour induire les passagers à acheter ses billets, un transporteur secondaire doit compenser ceux-ci en leur offrant des tarifs plus bas ou un meilleur niveau de service en retour des points auxquels ils renoncent en volant avec lui.

Les améliorations au programme de fidélité d'un transporteur sont associées avec des hausses dans la valeur de ses vols. L'effet est plus grand sur les routes dont le départ se fait à partir d'un aéroport où le transporteur est dominant alors qu'aucun effet n'a été détecté dans les autres cas. Pour y arriver, elle utilise le nombre de vols et de destinations offertes par les partenaires étrangers sur lesquels des points peuvent être réclamés ou ramassés pour établir la portée du programme de fidélité du transporteur domestique. L'utilité du passager i se procurant le produit a, sur une certaine route pour une certaine période, s'écrit:

$$U_a^i = \beta^T x + \alpha p_a + \xi_a^i + \nu_a^i \tag{3.14}$$

où x est le vecteur des caractéristiques observées du produit a, p_a est le prix du produit a, ξ_a^i est le vecteur des caractéristiques inobservées par l'économètre du

produit a et v_a^i est le terme d'erreur idiosyncrasique. Pour permettre aux hausses dans tous les tarifs de réduire la demande totale agrégée, les biens externes (*outside goods*) sont introduits et ils représentent la décision de ne pas voler, les vols inter-lignes sur deux transporteurs domestiques différents et les vols à bord de transporteurs étrangers. Inversement, les biens internes (*internal goods*) sont les vols directs ou deux vols avec le même transporteur (*online flights*) parmi un groupe de 12 transporteurs domestiques.

Le cadre logit imbriqué utilisé regroupe tous les biens internes au sein d'un même nid et les biens externes dans un autre. Ceci admet la corrélation dans v_a^i à travers les biens internes, leur permettant d'être des substituts plus rapprochés entre-eux qu'avec les biens externes. Formellement,

$$v_a^i = \zeta_a^i(m) + (1-\sigma)\varepsilon_a^i \tag{3.15}$$

où $\zeta_a^i(m)$ est l'effet aléatoire commun à tous les produits internes et a une fonction de distribution qui dépend de σ, avec $0 \leq \sigma \leq 1$, et ε_a^i est la préférence idiosyncrasique du produit j distribuée de façon indépendante et identique. Lorsque σ se rapproche de 1, la corrélation des niveaux d'utilité au sein du groupe tend vers 1 et lorsque σ se rapproche de 0, la corrélation des niveaux d'utilité au sein du groupe tend vers 0. En faisant l'hypothèse de distribution GEV pour ε_a^i et notant C_{mi} l'ensemble de produits au sein du groupe m pour le passager i, la part de marché pour le produit a est:

$$\Pr_{C_i}^i(a) = e^{\frac{v_a^i}{(1-\sigma)}} \left(D_m^{\sigma} \left[\sum_m D_m^{(1-\sigma)} \right] \right)^{-1} \tag{3.16}$$

où, pour un produit du groupe m,

$$D_m = \sum_{b \in C_{mi}} e^{\frac{v_b^i}{(1-\sigma)}}. \tag{3.17}$$

Le système d'expressions de part de marché dans un modèle logit imbriqué pour être résolu analytiquement pour les niveaux d'utilité moyens, V_a^i. En fixant l'utilité moyenne du bien externe à 0, $V_0^i = 0$, et en substituant $V_a^i = \beta^T x + \alpha p_a + \xi_a^i$, l'équation d'estimation suivante pour le logit imbriqué peut être construite:

$$\ln\left(\Pr_{C_i}^i(a)\right) - \ln\left(\Pr_{C_i}^i(0)\right) = \beta^T x + \alpha p_a + \sigma \ln\left(\Pr_{C_i}^i(C_{mi})\right) + \xi_a^i \qquad (3.18)$$

où $\Pr_{C_i}^i(C_{mi})$ est la part du produit *a* au sein du groupe *m*, i.e., la part du produit *a* de tous les passagers à bord d'un des 12 transporteurs domestiques. ξ_a^i (le niveau de service inobservé du produit *a*) est le terme d'erreur.

3.2 Contexte de part d'itinéraire (*itinerary share*)

La plupart des études sur la demande de transport aérien entre deux villes se sont penchées sur la part du transporteur au niveau du réseau complet, de la paire de villes ou du vol-direct (*point-to-point*). Peu d'études ont utilisé des données au niveau de l'itinéraire et se sont penchées sur le choix de celui-ci. Pourtant, le choix d'itinéraire du passager est fondamental puisqu'il regroupe simultanément toutes les décisions au niveau du choix: du transporteur, du trajet, du créneau d'horaire et du type d'avion. Or, les modèles axés sur la part de l'itinéraire procurent aux transporteurs une compréhension de l'importance relative qu'ont différents facteurs de service sur la part du marché de l'itinéraire entre deux villes et de la compétitivité sous-jacente qui existe parmi ceux-ci. De plus, ils expliquent comment les changements de politiques à l'égard des caractéristiques de service peuvent augmenter la part du marché sur l'itinéraire.

Coldren et Koppelman (2003) obtiennent des résultats concluants à partir de leurs modèles agrégés de parts d'itinéraires estimés au niveau des paires de villes pour 500 des plus grands marchés est-ouest nord-américains. Un itinéraire est défini comme étant un segment (numéro de vol) ou une séquence de segments reliant une paire de villes. Les modèles prédisent le nombre de passagers transportés par itinéraire et

aident les transporteurs dans leurs décisions puisqu'ils distribuent la prévision totale de passagers transportés entre deux villes parmi les itinéraires offerts. Une fois que la part d'itinéraire est prédite pour toutes les paires de villes, les prévisions peuvent être agrégées (*rolled-up*) pour prédire la part du transporteur à différents niveaux d'agrégation. Plus de détails sur cette étude sont présentés plus loin.

Coldren *et al.* (2003) modélisent la part d'itinéraire selon une fonction agrégée logit multinomiale des caractéristiques de l'itinéraire. L'étude se penche sur l'influence des différentes caractéristiques de service telles le niveau de service, le transporteur, le tarif, le créneau horaire, le type d'avion, etc. Malheureusement, vu la propriété d'indépendance des alternatives non-pertinentes du modèle multinomial logit (Koppelman *et al.*, 2003; Ben-Akiva et Lerman, 1985), ces modèles tracent un portrait incomplet de la substitution parmi les itinéraires. Cette propriété est invraisemblable dans le contexte de la modélisation du transport aérien selon la part d'itinéraire puisqu'il est fort probable que la compétition entre les itinéraires (telle que mesurée par les élasticités croisées) est différenciée par la proximité de l'heure du départ, du niveau de service[11], du transporteur ou toute combinaison de ces dimensions. Des modèles plus avancés de part d'itinéraire tiennent compte de cette structure compétitive sous-jacente et examinent l'impact différentiel de changements dans un itinéraire sur chaque autre itinéraire basé sur la similitude des principales caractéristiques de l'itinéraire, soit: le créneau horaire, le transporteur et le niveau de service.

Dans cette optique, Coldren et Koppelman (2003) développent plusieurs modèles GEV pour tenir compte des différentes tendances (*patterns*) de substitution parmi les alternatives différenciées selon la proximité des dimensions du créneau horaire, du transporteur et du niveau de service. Ceci permet de tenir compte de la possibilité de corrélation entre les termes d'erreur pour des paires d'alternatives (McFadden, 1978; Ben-Akiva et Lerman, 1985). Leurs modèles incorporent une, deux ou trois de ces

[11]Types de vols, soit: direct sans escale, direct avec escale, à une connexion et à deux connexions.

dimensions simultanément et leurs spécifications incluent le logit multinomial, le logit imbriqué à un et à deux niveaux et le logit imbriqué pondéré (*weighted*) à un et à deux niveaux avec l'emphase mise sur les paramètres logs-sommes (*logsum*) qui représentent le niveau de compétition entre itinéraires au sein des nids.

Les auteurs utilisent la valeur d'un itinéraire pour représenter l'attrait relatif de chaque itinéraire reliant une paire de villes. La part de marché attribuée à chaque itinéraire est modélisée comme une fonction de la valeur de celui-ci et des valeurs des autres itinéraires desservant la même paire de villes pour un jour donné de la semaine. Les variables qui décrivent chaque itinéraire et leurs paramètres estimés correspondants déterminent la valeur de celui-ci. La valeur est formulée comme une fonction linéaire pondérée des variables indépendantes, notamment: le niveau de service, la qualité de la connexion, le transporteur, le tarif, le type d'avion et le créneau horaire.

Les modèles utilisent des données de réservations et d'horaires officiels et complets. Les itinéraires sont construits avec le *itinerary building engine* de United Airlines à partir de données d'horaires basées sur les segments. Les itinéraires sont générés pour chaque jour de la semaine en retenant les jours de la semaine où chaque segment opère. La variable dépendante résultante est le nombre de passagers qui ont des réservations sur chacun des itinéraires. Les ensembles d'alternatives (*choice sets*) comprennent tous les itinéraires entre chaque paire de villes pour chaque jour de la semaine. L'analyse statistique est basée sur les choix de passagers individuels, par contre, aucune donnée de nature personnelle (socio-économique, démographique, etc.) n'est disponible pour identifier les différences entre passagers voyageant entre deux villes. Se référer à leur étude antérieure (Coldren *et al.*, 2003) pour plus de détails concernant la structure des données du modèle.

Au point de vue des résultats, leur modèle logit imbriqué à un niveau démontre clairement que les itinéraires ayant le même créneau horaire et ceux opérés par le

même transporteur ont des caractéristiques communes que le passager considère dans son choix d'itinéraire. Par contre, le niveau de service n'a pas fourni de résultats acceptables au point de vue statistique. Leur modèle logit imbriqué à deux niveaux montrent qu'il y a compétition moyenne entre les itinéraires partageant le même créneau horaire mais qu'il y a forte compétition lorsque ceux-ci partagent, à la fois, le même créneau horaire ou encore, le même transporteur et le même niveau de service.

Leur modèle logit imbriqué pondéré à un niveau estime simultanément un modèle avec deux structures de nids parallèles et un paramètre de poids qui indique l'importance relative de chaque nid. Chaque itinéraire de l'ensemble des alternatives apparaît deux fois dans le modèle, i.e., une fois dans chaque nid parallèle, de sorte que chaque nid est équivalent à un modèle logit imbriqué à un niveau. Les paramètres logs-sommes pour les deux nids (créneau horaire et transporteur) sont significativement différents de 1, indiquant une compétition accrue parmi les itinéraires partageant le même créneau horaire ou le même transporteur. Les paramètres de poids sont proches de ½ et significativement différents de 0 ou 1, indiquant que chaque portion de la structure est importante. Cependant, le modèle performe moins bien que le modèle logit imbriqué à deux niveaux (créneau horaire, transporteur). Donc, la substitution entre itinéraires opérés par un transporteur au sein d'un créneau horaire est plus importante qu'entre transporteurs opérant dans différents créneaux.

Finalement, leur logit imbriqué pondéré à deux niveaux est une extension directe de celui à un niveau et est le seul à incorporer les trois dimensions. Ce qui montre le potentiel d'obtenir des améliorations encore plus intéressantes et significatives au niveau de la qualité et de la structure du modèle en l'imbriquant de plus en plus.

CHAPITRE 4: DONNÉES

Avant d'entamer toute tentative de modélisation, un procédé de traitement des données brutes doit être mis en place et exécuté. Ceci permet non seulement de créer de nouveaux champs calculés qui serviront dans l'étape de modélisation mais aussi, de réduire la taille du jeu de données en écartant les observations jugées inutiles. Cette étape demande un effort considérable en termes de ressources et constitue un aspect important de l'obtention des résultats de modélisation. Les sections 4.1 et 4.1.1 décrivent les données et le procédé utilisé pour construire des itinéraires ainsi que des attributs reliés aux alternatives et des pseudo-attributs reliés aux individus.

4.1 Données brutes

Les données brutes proviennent d'un grand transporteur nord-américain et sont masquées pour protéger les intérêts commerciaux de celui-ci. Le jeu de données ne contient pas de données des compétiteurs, ni de données démographiques ou socio-économiques sur les passagers.

Tableau 4.1: Description des données brutes.

CHAMP	exemple	NOTES
NUMÉRO DU PNR	A57916	- Deux PNRs avec le même numéro ne peuvent être valides en même temps dans le système de rés.
DATE DE LA RÉSERVATION	7/1/05	
DATE DE DÉPART	9/16/05	
NOMBRE DE PASSAGERS	1	
NUMÉRO DU SEGMENT	1	
SIGLE DU TRANSPORTEUR	ZZ	
NUMÉRO DU VOL	1125	
ORIGINE	WIT	
HEURE DE DÉPART	20:15	- Heure locale au décollage.
DESTINATION	NMY	
HEURE D'ARRIVÉE	21:37	- Heure locale à l'atterrissage.
CLASSE TARIFAIRE	L	- Classe dans laquelle se fait la réservation.
TARIF PAYÉ	312.43	- En dollars, avant taxes et frais de service.
RÉSERVATION FAITE VIA LE WEB[12]	1	- Booléen. Prend la valeur 1 si la réservation se fait via le Web; 0 sinon.

[12] Ce champ est calculé *ex ante* à partir d'un pourcentage global de réservations faites via le web qui est ensuite redistribué au niveau des segments. Il s'agit de remanier cette valeur pour certains segments au sein des PNRs pour restaurer la vraisemblance du processus de réservation. L'impact de ce remaniement (avec la règle: 0 si <50%; 1 sinon) sur le pourcentage global initial est jugé minime vu la taille du jeu de données.

Ainsi, tous les passagers sont identiques[13] et l'ensemble d'alternatives correspond aux combinaisons de vols opérés par le transporteur reliant deux villes.

Les champs disponibles dans les données brutes ainsi qu'un exemple et quelques notes explicatives se retrouvent au Tableau 4.1. Les données brutes couvrent tous les vols commerciaux dont les départs s'effectuent entre le 25 août et le 19 octobre 2005. À partir de ces données, basées sur le vol, il faut reconstruire les itinéraires de chaque passager et leur rattacher leurs attributs respectifs tels le temps en transit[14], le niveau de service, la classe de service dominante, le tarif payé, etc. De plus, bien que tous les passagers soient identiques, il est tout de même possible de segmenter cette population homogène en examinant les caractéristiques de la classe tarifaire choisie et celles du voyage en tant que tel. Le but de ce pas additionnel est de vérifier si des différences de comportement existent parmi les différents groupes de passagers et si c'est le cas, d'utiliser cette information additionnelle dans la gestion du revenu.

4.1.1 itinéraires

Les itinéraires sont créés à partir des données brutes et représentent un déplacement uni-directionnel entre deux villes. Ils se composent d'un vol ou une combinaison de vols opérés par le transporteur unique. Il est possible de segmenter un itinéraire en plusieurs itinéraires en considérant, par exemple, à quel moment se fait le départ. Coldren et Koppelman (2003) optent pour générer des itinéraires par jour de la semaine. Autrement dit, tous les vols du lundi dans leur échantillon qui relient une paire de villes dans un sens constituent un seul itinéraire. Cet itinéraire est différent des vols du mardi entre ces mêmes villes et des itinéraires des autres jours de la semaine. Il est possible de diviser d'avantage en considérant le jour du départ ou le moment de la journée où s'effectue le départ. Bref, la segmentation doit se faire de

[13] Puisqu'ils ont les mêmes caractéristiques démographiques et socio-économiques. Il peut être utile de considérer que tous les voyages sont entrepris par un passager typique, i.e., un passager dont les caractéristiques socio-économiques correspondent aux moyennes observées dans la population.

[14] Si l'itinéraire est composé d'un vol simple, le temps en transit est tout simplement égal à la durée du vol. Si l'itinéraire est composé d'une combinaison de deux ou plusieurs vols, le temps de transit est la somme des durées de vols et des temps d'attentes aux points intermédiaires.

façon à respecter le contexte économique tout en évitant les problèmes numériques qu'un échantillon trop petit peut provoquer.

Dans cette recherche, la paire de villes qui enregistre les flots de passagers les plus importants sur la période de temps à l'étude est WNZWIT. Le vol retour, WITWNZ, vient au second rang (se référer au Tableau 4.2). Il est donc intéressant de considérer les quatre itinéraires suivants: WNZWIT, WITWNZ, WNZWITR et WITWNZR où le suffixe "R" désigne que l'itinéraire est en fait un aller-retour de l'origine à l'origine, via la destination. La raison principale est que le tarif payé dans les données brutes s'applique au voyage complet et non au segment. Puisqu'un voyage peut contenir plusieurs segments qui forment plusieurs itinéraires, chaque itinéraire du voyage hérite du même tarif payé. Or, un itinéraire court au sein d'un long voyage risque d'afficher une valeur de tarif payé plus élevé que le même itinéraire entrepris seul. Il semble donc souhaitable de les séparer.

Tableau 4.2: Les dix (10) itinéraires les plus importants.

ORIGINE	DESTINATION	PASSAGERS	PNRS
WNZ	WIT	42227	30504
WIT	WNZ	40651	29311
WIT	AKT	30859	26456
AKT	WIT	28268	24537
WIT	OQK	26806	18801
OQK	WIT	26042	18492
WIT	OUE	16058	11731
WNZ	OQK	15818	12107
WIT	WPT	15570	11397
OUE	WIT	15436	11532

Note: échantillon partiel.

Un autre aspect important qui dépend du contexte économique est la qualité des connexions dans les combinaisons de vols. Une connexion sera de grande qualité si l'intervalle de temps qui sépare l'arrivée d'un vol et le départ du vol suivant sur l'itinéraire d'un passager est petit. Puisque le but d'un voyage est d'accomplir quelque chose à destination, les chances de réaliser cet objectif augmentent en maximisant la durée du séjour ou, inversement, en minimisant le temps en transit. Même si le voyage est un aller simple, le but demeure d'arriver le plus vite possible à destination. Si le passager s'attarde trop longtemps à un point intermédiaire sur son

itinéraire, l'itinéraire initial est remis en question. Au fait, au-delà d'un certain seuil, la qualité de la connexion devient si mauvaise que ce sont deux itinéraires: (1) de l'origine initiale au point intermédiaire et (2) du point intermédiaire à la destination initiale.

Dans cette recherche, une règle bien simple[15] est utilisée pour délimiter ce seuil: le passager qui arrive à un point intermédiaire doit avoir une réservation sur un des deux prochains vols reliant la destination suivante sur son itinéraire. Le cas échéant, la qualité de la connexion est jugée trop mauvaise et l'itinéraire est coupé en deux au point intermédiaire. Voici un exemple d'un PNR où un passager voyage entre WNZ et CZT en passant par WIT:

Tableau 4.3: Règle de construction d'itinéraire (exemple 4.1).

DATE DE DÉPART	NO. SEG.	NUMÉRO DU VOL	ORIG.	HEURE DE DÉPART	DEST.	HEURE D'ARRIVÉE	TARIF PAYÉ
11/3/05	1	0513	WNZ	10:00:00 AM	WIT	5:24:00 PM	390.76
11/3/05	2	1505	WIT	8:30:00 PM	CZT	10:45:00 PM	390.76
11/9/05	3	1500	CZT	10:30:00 AM	WIT	12:50:00 PM	390.76
11/9/05	4	0558	WIT	4:00:00 PM	WNZ	5:57:00 PM	390.76

- À l'aller, lorsque le passager arrive à WIT, il doit prendre soit sur le vol 1505 ou 1497 (le lendemain matin) pour que son itinéraire demeure WNZCZT; comme en témoigne le tableau suivant. Il prend effectivement le premier vol possible, la qualité de cette connexion est maximale et l'itinéraire initial est conservé.

WITCZT	DATE DE DÉPART	NUMÉRO DU VOL	HEURE DE DÉPART
1	11/3/05	1505	8:30:00 PM
2	11/4/05	1497	7:20:00 AM

- Au retour, par contre, lorsque le passager arrive à WIT, il doit connecter soit sur le vol 0566 ou 0506 pour que son itinéraire demeure CZTWNZ. Il prend le

[15] D'autres règles sont plus simples à implanter dans l'algorithme de construction d'itinéraire. Par exemple, un itinéraire est coupé en deux si la connexion entre le vol entrant et le vol sortant ne se fait pas en dedans de x heures.

quatrième vol possible. La qualité de cette connexion est mauvaise et l'itinéraire initial est coupé en deux itinéraires, CZTWIT et WITWNZ.

WITWNZ	DATE DE DÉPART	NUMÉRO DU VOL	HEURE DE DÉPART
1	11/9/05	0566	1:10:00 PM
2	11/9/05	0506	2:00:00 PM
3	11/9/05	0544	3:00:00 PM
4	11/9/05	0558	4:00:00 AM

Le constructeur d'itinéraire parcourt les PNRs et y relie les segments pour en faire des itinéraires. Il place les segments dans un ordre chronologique et géographique et tente de les relier. Dans la plupart des cas cependant, un segment représente un itinéraire. Ce procédé est lent et coûteux car il requiert beaucoup de ressources informatiques et de temps de calcul. De plus, il doit être refait à chaque fois que les règles par rapport aux connexions changent.

Il est utile d'illustrer les itinéraires sous forme de nœuds et d'arêtes sur un graphe espace-temps. L'espace (géographique) est représenté sur l'axe des ordonnées alors que le temps est représenté sur l'axe des abscisses. Les arêtes représentent des tâches, i.e., soit un vol entre deux villes ou une période d'attente à une ville pour y effectuer une escale ou une connexion. Les nœuds marquent le début ou la fin de la tâche. Le graphe se lit de gauche à droite[16] et pour chaque arête, le nœud de gauche marque le moment où cette tâche débute et celui de droite marque le moment où elle se termine. Sur les figures qui suivent, les itinéraires sont représentés par les chaînes d'arêtes plus foncées. Les deux premières figures illustrent deux voyages types WNZCZTR. Le voyage d'affaires type est entrepris durant la semaine et le séjour à destination est court. Les vols directs sont privilégiés puisqu'ils réduisent le temps en transit Le voyage de plaisance type est généralement plus long et au moins un samedi est passé à destination. Or, le passager est plus réceptif à faire des escales ou des connexions surtout s'il visite une destination éloignée ou peut économiser sur le prix des billets.

[16] Ceci permet d'utiliser des arêtes pour représenter les déplacements au lieu d'arcs et d'alléger le graphe.

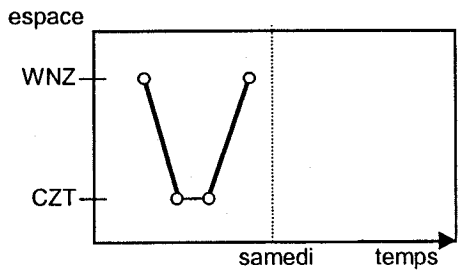


Figure 4.1: Voyage d'affaires type.

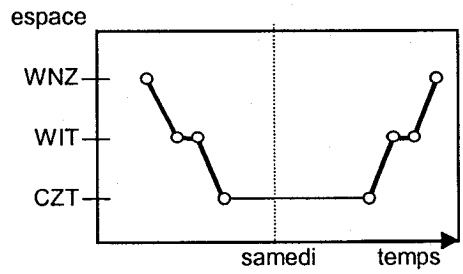


Figure 4.2: Voyage de plaisance type.

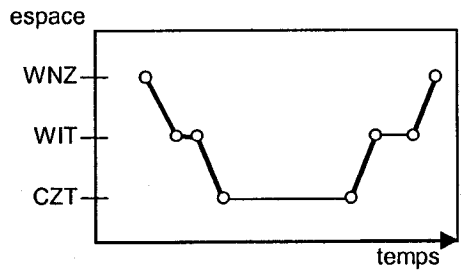


Figure 4.3: Voyage de l'exemple 4.1.

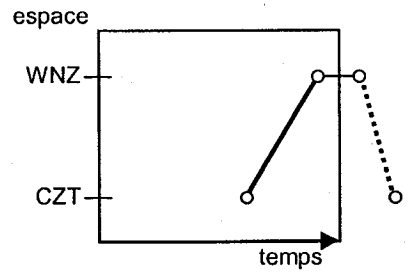


Figure 4.4: Voyage coupé par l'échantillon.

La Figure 4.3 illustre l'exemple 4.1 avec ces trois itinéraires. Finalement, la Figure 4.4 montre une autre difficulté que comporte l'échantillon utilisé qui couvre une période de temps bien délimitée. L'existence de l'arête en pointillé n'est pas garantie car le départ se fait à l'extérieur de la période à l'étude.

4.1.2 tarifs

Le constructeur d'itinéraires construit l'itinéraire et y attribue le tarif payé pour le voyage en entier. Certains itinéraires sont entrepris en échange de points de programmes de fidélité et leurs tarifs payés correspondants sont nuls. Si le voyage est un aller retour qui contient deux itinéraires, le tarif payé pour celui-ci risque d'être plus élevé que pour un aller-simple. Cependant, lorsque le voyage est entrepris près de la limite échantillonale, un aller-simple peut en fait être un aller-retour déguisé et afficher un tarif payé plus élevé. Il n'y pas de moyen de savoir et il est probable que les itinéraires allers-simples (*eg.* WNZWIT) soient plus nombreux près des limites de l'échantillon et affichent de fortes variations de tarifs au sein d'une même classe de service.

Pour y remédier, le filtre suivant est appliqué uniformément sur tout l'échantillon: un tarif en-dessous de $100 en classe affaires ou $50 en classe économie est considéré non-payant. Il n'est pas possible d'appliquer un tel filtre pour le tarif maximum puisque certaines classes tarifaires ne sont pas offertes sur les vols inter-continentaux et les caractéristiques du voyage influencent beaucoup le tarif payé final. La Figure 4.5 montre l'étalement de trois classes tarifaires en représentant les distributions de fréquences des réservations. Beaucoup d'observations des classes tarifaires *M* et *V* seraient perdues si une limite uniforme supérieure de $1300 était imposée. Une étude plus détaillée (pour chaque paire de villes) est requise pour écarter les observations extrêmes sans éliminer de données valides. Néanmoins, dans les trois cas précédents, il est possible d'exclure les observations questionnables en réduisant la durée de la période à l'étude ou en examinant uniquement les classes tarifaires appartenant aux cinq classes de service.

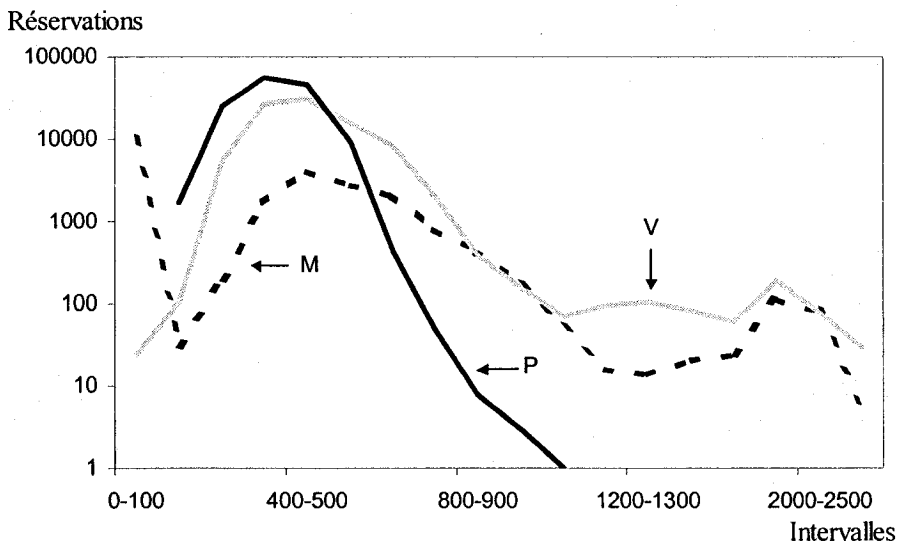


Figure 4.5: Étalement des tarifs (échelle logarithmique).

4.1.3 classes de service

Les caractéristiques des classes de service sont booléennes, en général, et dénotent par 1 les privilèges auxquels le passager à droit. Plus de détails, un exemple et quelques notes explicatives se retrouvent au Tableau 4.4.

Tableau 4.4: Description des caractéristiques des classes de service.

CHAMP	exemple	NOTES
NOM DE LA CLASSE	A	- Lettre désignant la classe tarifaire.
COMPARTIMENT	ECO	- Affaires (BUS) ou économie (ECO).
RESTREINT	1	- Prend la valeur 1 s'il est vendu en quantité limitée ou sur une période de temps limitée.
ACCES AU SALON DE REPOS	0	
ENREGISTREMENT PRIORIT.	0	
EMBARQUEMENT PRIORITAIRE	0	
POURCENT. MILLES AERIENS	100	- Pourcentage de milles aériens obtenus.
FRAIS POUR MODIFICATION	50	- Montant en dollars à payer pour modif.
REMBOURSABLE	0	- Prend la valeur 1 si le billet est remb.
SELECTION DU SIEGE	1	- Prend la valeur 1 si un siège peut être selectionné à l'enregistrement.
POSSIBILITE D'UPGRADE	0	- Prend la valeur 1 s'il permet de voyager en BUS en échange de points accumulés.

Note: la lettre de l'alphabet désignant une classe tarifaire n'est qu'une étiquette donnée par le transporteur à une classe de service. Chaque transporteur a son propre système de codage de classes tarifaires.

Puisque toutes les classes tarifaires ne sont pas disponibles sur chaque vol et que plusieurs classes tarifaires peuvent représenter la même classe de service, il est utile de les regrouper selon leurs caractéristiques et de rattacher les attributs du produit qu'elles définissent aux itinéraires. Ainsi, un itinéraire ne se fera pas en classe tarifaire *A* mais plutôt en classe de service *1*. Les classes de service utilisées dans cette recherche apparaissent dans le Tableau 4.5. Dans le cas d'itinéraires à plusieurs segments où les classes tarifaires diffèrent d'un segment à l'autre, la classe tarifaire associée au plus long vol est dite dominante et la classe de service qu'elle définit est attribuée à l'itinéraire complet. Le raisonnement est que les vols moins longs sont contraints par le vol plus long et que l'itinéraire du passager est construit autour du vol le plus long.

Tableau 4.5: Classification des classes de service.

CLASSE DE SERVICE	1	2	3	4	5
COMPARTIMENT	BUS	ECO	ECO	ECO	ECO
RESTREINT	0	0	1	1	1
ACCES AU SALON DE REPOS	1	1	0	0	0
ENREGISTREMENT PRIORIT.	1	1	0	0	0
EMBARQUEMENT PRIORITAIRE	1	1	0	0	0
POURCENT._MILLES AERIENS	150	125	100	100	50
FRAIS POUR MODIFICATION	0	0	0	50	150
REMBOURSABLE	1	1	1	0	0
SELECTION DU SIEGE	0	1	1	1	0
POSSIBILITE D'UPGRADE	0	1	1	0	0

La classe de service désirée sur ce vol principal dictera vraisemblablement le coût du billet et les autres vols sont secondaires puisque le passager est prêt à accepter un produit différent sur ceux-ci pour vu qu'il obtienne le produit désiré sur le vol principal. En se référant au Tableau 4.5, il y a essentiellement cinq produits offerts par le transporteur sur chaque vol et le passager en choisit un parmi ceux-ci. La numérotation est arbitraire mais elle représente, néanmoins, un rang. À prix égal, un passager choisirait le produit *4* plutôt que *5*, *1* plutôt que *2*, etc. Le produit *1* serait liquidé en premier parce que sa valeur, en termes de la somme des poids accordés à ses attributs, est plus grande que celle de tout autre produit. De plus, l'écart de valeur n'est pas le même entre *1* et *2* qu'entre *2* et *3*. Ceci doit être pris en compte lorsque la classe de service est une variable explicative dans un modèle.

4.2 Choix d'alternatives

La section 4.1 a permis d'identifier quelques difficultés qui surgissaient lors de l'utilisation des données brutes pour définir les itinéraires des passagers. De plus, le dernier paragraphe a introduit l'idée de choix de produit auquel pourrait face un passager type. Il reste maintenant à définir quelques règles pour réduire l'ensemble d'alternatives à une taille raisonnable. Un ensemble trop petit et le modèle est biaisé vers certaines alternatives. Un ensemble trop grand et l'hypothèse d'indépendance des alternatives non-pertinentes est violée, rendant ainsi le modèle multinomial logit inutilisable.

Tout d'abord, l'usage d'aéroports fictifs ne permet pas de représentation géographique du réseau desservi. Or, en réalité, l'ensemble d'alternatives dépend beaucoup de l'emplacement géographique de deux villes. De plus, dans le cas d'une ville à aéroports multiples, l'emplacement de ceux-ci peut déterminer s'ils appartiennent à l'ensemble d'alternatives ou pas. S'il existe plusieurs alternatives pour atteindre une ville à partir d'une autre ville, il s'agit de limiter son choix à celles qui nous semblent les meilleures.

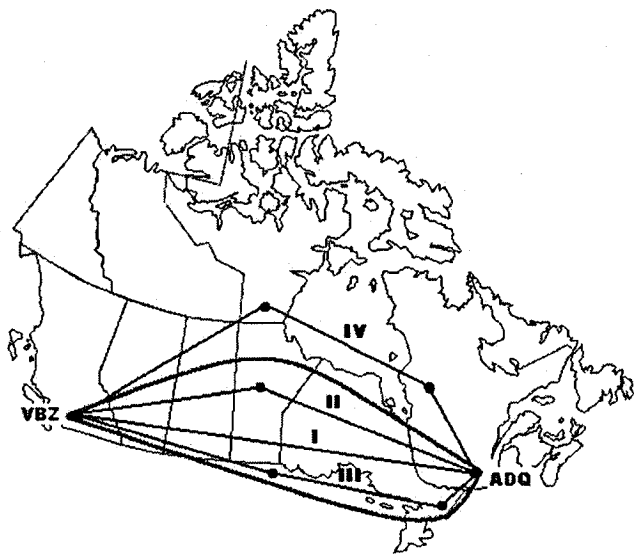


Figure 4.6: Règles de circuité et ensemble d'alternatives[17].

Dans la Figure 4.6, quatre combinaisons de vols, *I* à *IV*, sont offertes entre ADQ et VBZ. Les autres nœuds représentent les aéroports où se font les escales ou les connexions. Les arêtes représentent les vols et les longueurs des arcs correspondent aux distances parcourues lors des vols. La forme ellipsoïdale représente l'ensemble d'alternatives tel que défini par les règles de circuité (*circuity*). Cette notion vient du principe que le plus court chemin entre deux points est la ligne droite qui les relient. Dans ce cas-ci, le vol direct *I*. En optant pour ce vol, le passager minimise la distance parcourue et maximise l'utilité qu'il retire du voyage. C'est l'alternative qui a la plus grande valeur à ses yeux mais, il y en a d'autres qui s'offrent à lui. Il est cependant prêt à faire certains compromis (jusqu'à un certain seuil) en choisissant des alternatives dont les chemins sont plus longs tant que la réduction d'utilité ainsi encourue soit compensée par un autre facteur, généralement monétaire. Le seuil est représenté par le périmètre de la forme ellipsoïdale et le passager considérera tous les chemins qui se trouvent à l'intérieur. En termes de règle de circuité, par exemple, il considère toutes les alternatives dont les chemins ont une ou deux escales (ou connexions), situées à l'intérieur de la forme ellipsoïdale, dont la somme des

[17] Source: http://www.transcanadahighway.com/Images/CanadianProvinces.jpg.

distances ne dépasse pas 1.5 fois la distance du vol direct *I*. C'est le cas des alternatives *II* et *III* dont les valeurs sont moindres mais néanmoins positives[18], puisque le passager en retire encore une certaine utilité en les choisissant. Ce n'est pas le cas de la combinaison de vols *IV* qui n'est pas considérée comme une alternative puisqu'elle repose à l'extérieur du périmètre. Elle a une valeur de zéro et ne procure aucune utilité au passager.

Le principe des règles de circuité peut être utilisé avec d'autres paramètres tels le temps de transit ou le prix du billet. En supposant que les longueurs des arêtes sur la Figure 4.6 représentent maintenant les temps de vols (incluant les escales), les vols à l'intérieur de la forme ellipsoïdale font partie de l'ensemble d'alternatives du passager. En termes de règles de circuité, par exemple, seuls les combinaisons de vols dont le temps de transit total est au-dessous de 1.5 fois le temps de transit minimum observé entre ADQ et VBZ font partie de l'ensemble d'alternatives. Les alternatives *II* et *III* demandent plus de temps en transit et sont moins attirantes aux yeux du passager que l'alternative *I* à moins d'offrir un écart de prix suffisant pour compenser cette perte d'utilité. L'alternative *IV* n'est pas considérée comme une alternative car son temps de transit est jugé trop long. Dans ce mémoire, les itinéraires qui sont deux fois plus longs (en termes de temps de transit) que le vol direct le plus long entre deux villes sont exclus de l'ensemble d'alternatives. Les détails reliés au calcul du temps de transit sont présentés aux paragraphes suivants.

4.2.1 temps en transit

La difficulté principale reliée au calcul du temps de transit provient de l'utilisation d'heures locales de décollage et d'atterrissage dans le calcul des durées de vol. Lorsque l'information sur les fuseaux-horaires est fournie, la durée du vol est calculée en prenant la différence entre l'heure d'arrivée et de départ du vol où toutes les deux sont exprimées en temps moyen de Greenwich (*GMT*). Or, sans information sur les fuseaux-horaires à l'origine et à la destination du vol, comme c'est le cas ici, les

[18] Leurs probabilités de sélection sont positives puisqu'elles reposent à l'intérieur du périmètre.

calculs de durée de vol sont délicats et imprécis. Elles peuvent être négatives sur certains vols ou être plus longues que ce qu'un avion commercial moderne est capable d'accomplir. Puisque la durée du vol a sûrement un impact important sur le choix du passager et sur la taille de l'ensemble d'alternatives, il faut du moins y remédier en partie.

Pour ce faire, il s'agit d'examiner les paires de villes où il existe des vols les reliant dans les deux directions. Il s'agit ensuite de faire l'hypothèse que presque la même trajectoire de vol est utilisée dans les deux directions et que celle-ci respecte le principe du plus court chemin entre ces deux villes. La résultante est donc deux vols de même durée en termes de temps que l'avion passe dans les airs. Il suffit ensuite d'examiner les durées de vol entre deux villes dans chacune des directions et d'appliquer le raisonnement suivant:

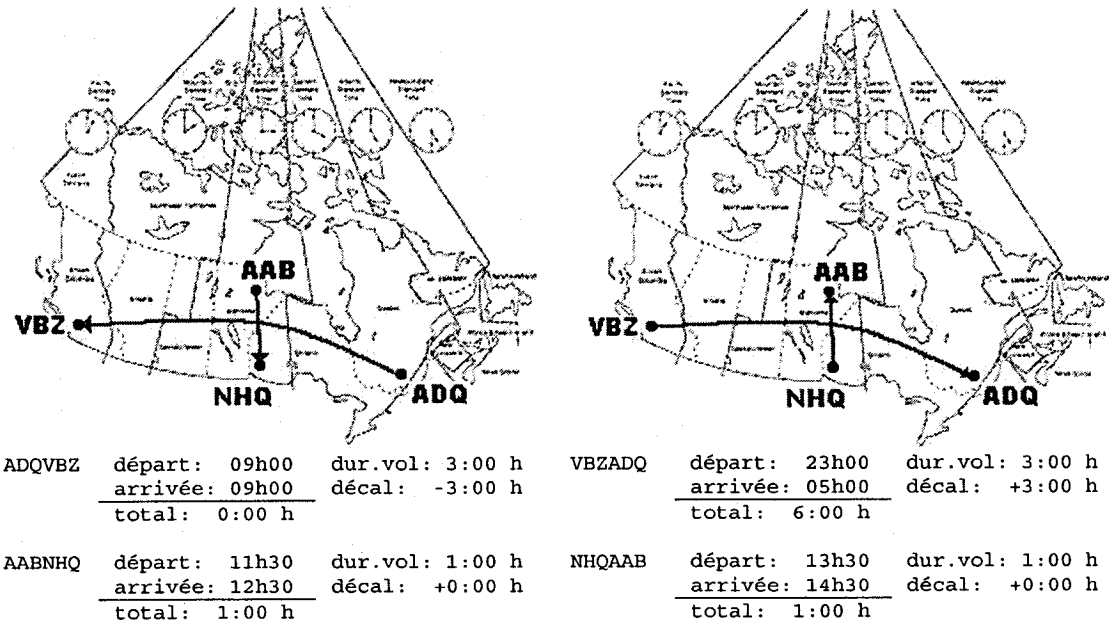
- Deux villes dont les durées de vol sont presque identiques dans les deux directions se trouvent dans le même fuseau horaire (axe nord-sud).
- Deux villes dont la différence entre les durées de vol est un nombre entier pair d'heures se trouvent dans deux fuseaux horaires différents (axe est-ouest).

Deux exemples sont illustrés à la Figure 4.7. En supposant que les villes ADQ et AAB sont des origines et que VBZ et NHQ sont des destinations, les informations relatives aux vols-allers (ADQVBZ et AABNHQ) apparaissent à gauche alors que celles des vols-retours (VBZADQ et NHQAAB) apparaissent à droite. Les résultats sont obtenus en résolvant le système d'équations (4.1) pour les deux paires de villes:

$$\begin{aligned} dur.vol + décal &= \text{départ}_{all} - \text{arrivée}_{all} \\ dur.vol - décal &= \text{départ}_{ret} - \text{arrivée}_{ret} \end{aligned} \tag{4.1}$$

où *dur.vol* et *décal* sont les deux inconnues et représentent, respectivement, le temps que l'avion passe les airs et le décalage-horaire entre la ville d'origine et celle de destination. Les valeurs de départ_{all}, arrivée_{all}, départ_{ret} et arrivée_{ret} sont connues et exprimées en heures locales. AAB et NHQ sont dans le même fuseau-horaire car les

durées de vol dans les deux directions sont les mêmes (*dur.vol* = 1 et *décal* = 0). ADQ et VBZ ne sont pas dans le même fuseau-horaire (*dur.vol* = 3 et *décal*= -3).



ADQVBZ	départ: 09h00	dur.vol: 3:00 h		VBZADQ	départ: 23h00	dur.vol: 3:00 h
	arrivée: 09h00	décal: -3:00 h			arrivée: 05h00	décal: +3:00 h
	total: 0:00 h				total: 6:00 h	
AABNHQ	départ: 11h30	dur.vol: 1:00 h		NHQAAB	départ: 13h30	dur.vol: 1:00 h
	arrivée: 12h30	décal: +0:00 h			arrivée: 14h30	décal: +0:00 h
	total: 1:00 h				total: 1:00 h	

Figure 4.7: Durée de vol et décalage-horaire[19].

4.2.2 heures de départ

Les heures de départ des itinéraires sont regroupées selon les périodes de pointe pour permettre un certain degré de compétition entre les itinéraires. Puisque les heures de décollage sont exprimées en heures et en minutes, il est possible de les regrouper ainsi. De plus, ce concept de vagues, i.e., période de décollages et de d'atterrissages intenses suivie d'une période où ces fréquences sont plus basses, est présent dans les aéroports achalandés. Il y a deux raisons pour ce phénomène. La première est la rotation des avions que les transporteurs utilisent où le même avion est utilisé pour opérer un ou plusieurs vols allers-retours au cours de la journée. Par exemple, un avion peut faire un vol sortant le matin, revenir à l'aéroport d'origine en fin d'avant-midi, faire un autre vol sortant en début d'après-midi puis revenir ensuite, etc. et l'horaire se répète le lendemain. L'autre raison est l'importance de l'heure d'arrivée à destination. Ainsi, l'heure du départ n'est pas si importante pour vu que l'arrivée se

[19] Source: http://www.cbsa-asfc.gc.ca/E/pub/cp/rc4032/map-e.jpg.

fasse à l'heure souhaitée. Les heures de départ sont regroupées selon les intervalles présentés à la Figure 4.8 et ces intervalles seront utilisés comme variables explicatives (attributs) dans les modèles pour déceler les préférences des passagers pour itinéraires débutants durant les pointes du matin (*MA*), du midi (*MI*), de l'après-midi (*PM*) ou du soir (SO). Les périodes hors-pointes sont représentées par les tirets pointillés.

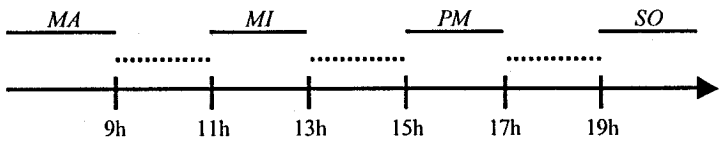


Figure 4.8: Intervalles d'heures de départ.

4.3 Attributs

Les sections précédentes de ce chapitre ont traité les données brutes pour produire des itinéraires et des ensembles de choix cohérents. Avant de procéder à la modélisation, il reste à définir et regrouper les attributs qui serviront à caractériser les passagers, les itinéraires et les alternatives. Le Tableau 4.6 contient la liste des attributs, leurs acronymes et s'ils se rattachent au passager, à l'itinéraire ou à la classe de service.

Tableau 4.6: Description des attributs.

ATTRIBUTS	ACR.	CAT.	VARIABLE	NOTES
NB. DE PASSAGERS	NP	PASSAG.	ENTIÈRE NON	- nombre de passagers par PNR.
NUIT DU SAMEDI	SA	PASSAG.	BOOLÉENNE	- 1 si la nuit du samedi est passée à destination; 0 sinon.
ORIGINE	OR	PASSAG.	BOOLÉENNE	- 1 si l'itin. commence à un certain aéroport; 0 sinon.
HEURES DE DÉPART	MA	ITIN	BOOLÉENNE	- 1 si le départ s'effectue avant 09:00; 0.5 si entre 09:00 et 10:59; 0 sinon.
	MI	ITIN	BOOLÉENNE	- 1 si le départ s'effectue entre 11:00 et 12:59; 0.5 si entre 09:00 et 10:59 ou entre 13:01 et 14:59; 0 sinon.
	PM	ITIN.	BOOLÉENNE	- 1 si le départ s'effectue entre 15:00 et 16:59; 0.5 si entre 13:00 et 14:59 ou entre 17:01 et 18:59; 0 sinon.
	SO	ITIN.	BOOLÉENNE	- 1 si le départ s'effectue après 19:00; 0 sinon.
TEMPS EN TRANSIT	TT	ITIN.	CONTINUE	- temps en transit (durée de vol et d'escale) en heures.
CONNEXION	MU	ITIN.	BOOLÉENNE	- 1 si l'itinéraire comprend au moins une connexion; 0 sinon..
LIMITE ECHANTILL.	LE	ITIN.	BOOLÉENNE	- 1 si le départ se fait dans les 2 premières ou dernières semaines de l'échan.; 0 sinon.
TARIF PAYÉ	TA	CLASSE	CONTINUE	- tarif payé en dollars.
CLASSE DE SERVICE	SE	CLASSE	CATÉGORIQUE	- indique la classe de service. - sert d'alternative.
COMPARTIMENT	CO	CLASSE	BOOLÉENNE	- 1 pour compartiment affaires.
RESTREINT	RE	CLASSE	BOOLÉENNE	- Prend la valeur 1 s'il est vendu en quantité restreinte.
ACCES AU SALON	LO	CLASSE	BOOLÉENNE	- 1 si accès au salon; 0 sinon.
ENREGIS. PRIORIT.	CI	CLASSE	BOOLÉENNE	- 1 s'il permet l'enreg. prior.
EMBARQU. PRIORIT.	BO	CLASSE	BOOLÉENNE	- 1 s'il permet l'embarq. prior.
POURCENT. MILLES	AM	CLASSE	CATÉGORIQUE	- % de milles aériens récoltés.
FRAIS POUR MODIF.	CF	CLASSE	CATÉGORIQUE	- Montant en dollars.
REMBOURSABLE	RF	CLASSE	BOOLÉENNE	- 1 si le billet est rembours.
SELECT. DU SIEGE	SS	CLASSE	BOOLÉENNE	- 1 s'il permet sélec. de siège.
POSS. D'UPGRADE	PU	CLASSE	BOOLÉENNE	- 1 s'il permet de voyager en BUS en échange de points.

Note: les attributs liés au passager sont en fait des pseudo-attributs socio-économiques.

CHAPITRE 5: MODÈLE PROPOSÉ

Ce chapitre examine le comportement de différents modèles de choix discret appliqués sur le jeu de données décrit au chapitre 4. Comme dans la plupart des articles présentés au chapitre 3, la démarche suivie sera de commencer par un modèle logit multinomial relativement simple et d'en examiner les résultats. Ensuite, dans le but d'améliorer ces résultats, une formulation logit imbriquée, puis logit imbriquée croisée, sera présentée.

5.1 Algorithmes et logiciel d'estimation

Le gratuiciel (*freeware*) BIOGEME (version 1.4) est utilisé pour estimer les paramètres dans tous les cas. L'acronyme est pour BIerlaire Optmization toolbox for GEv Model Estimation. BIOGEME est conçu pour estimer les modèles logit binaire, logit multinomial, logit imbriqué ainsi que les modèles plus complexes de la famille GEV et les combinaisons de ces modèles (*eg.* logit mixte).

Les entrants sont constitués de deux fichiers: un qui contient les données et le second qui contient les caractéristiques du modèle à estimer et les paramètres que doit estimer BIOGEME. Plusieurs fichiers sortants sont produits mais les deux plus importants comportent les résultats de l'estimation et une analyse sommaire des données utilisées. BIOGEME peut utiliser cinq algorithmes d'optimisation différents: BIO, BIOMC, CFSQP, DONLP2 et SOLVOPT. Il est possible que chacun d'entre eux produise des solutions différentes. D'ailleurs, il est toujours souhaitable de résoudre le même problème avec différents algorithmes. En général, les variations sont petites et dues aux différents critères d'arrêt. De plus, aucun de ces algorithmes ne prétend trouver un maximum global de la fonction de vraisemblance. Donc, il peut arriver qu'un d'eux soit pris dans un maximum local différent des maxima locaux trouvés par les autres algorithmes.

La sélection de l'algorithme à utiliser n'est pas toujours évidente. Ils ont tous leurs avantages et leurs inconvénients. BIO (pour BIerlaire's Optimization, comme dans BIOGEME) a été spécialement adapté pour ce paquetage logiciel mais, pour l'instant, il ne peut accommoder les contraintes non-triviales. BIOMC est une version qui estime le maximum de vraisemblance simulé. CFSQP n'est pas gratuit et donc, n'est pas inclus dans la distribution générale de BIOGEME (Version 1.4). DONLP2 est plus lent que CFSQP mais plus rapide que SOLVOPT qui est parfois très lent.

Tableau 5.1: Algorithmes d'optimisation que peut utiliser BIOGEME.

BIO (Conn et al., 2000)

- Conçu pour les problèmes avec contraintes de bornes simples.
- Algorithme basé sur la méthode de région de confiance.
- Utilise la méthode du gradient conjugué tronqué pour résoudre le sous-problème de région de confiance.
- Utilise, par défault, la matrice BHHH (Berndt *et al.*, 1974) comme approximation de la matrice des dérivées secondes. Les erreurs suivent la distribution normale.

BIOMC: version de BIO

- Conçu pour l'estimation du maximum de vraisemblance simulé.
- L'idée est d'utiliser BIO avec quelques tirages au début pour obtenir une valeur assez fiable des paramètres et ensuite d'augmenter le nombre de tirages jusqu'au seuil désiré par l'usager.
- En général, BIOMC n'est pas plus rapide que BIO mais permet d'obtenir de bonnes approximations des paramètres assez rapidement.

CFSQP (http://www.aemdesign.com/FSQPwhatis.htm)

- Utilise la méthode de programmation quadratique séquentielle réalisable.
- Implémentation en langage C de l'algorithme d'optimisation FSQP développé par E.R. Panier, A.L. Tits, J.L. Zhou, et C.T. Lawrence (Lawrence *et al.*, 1997).

DONLP2 (Spellucci, 1993)

- Utilise la méthode de programmation quadratique contrainte.
- L'auteur décrit l'algorithme dans deux ouvrages (Spellucci, 1998a et 1998b).

SOLVOPT (Kuntsevich et Kappel, 1997)

- Acronyme pour SOLVer for local OPTimization problems.
- Résoud les problèmes containts où la fonction objectif est non-linéaire (ou même non-lisse) par la méthode de pénalisation exacte.

Source: Bierlaire (2005), pp.40-43.

Cependant, ce dernier n'est pas à blâmer car il n'a pas été conçu pour solutionner le type de problèmes d'optimisation résolus avec BIOGEME. Il est robuste dans le cas de modèles mal spécifiés et peut trouver une solution là où les autres algorithmes ne convergent pas. Le Tableau 5.1 fournit quelques informations supplémentaires sur ces algorithmes.

La suggestion de Bierlaire est d'utiliser BIO si le modèle ne contient pas de contraintes non-triviales sur les paramètres. S'il en contient ou qu'il est très lent, utiliser DONLP2. Si ceci échoue, tenter de solutionner le modèle avec SOLVOPT. Finalement, si SOLVOPT échoue, redéfinir le modèle. Dans ce mémoire, tous les modèles logit multinomial et logit imbriqué sont estimés à l'aide de DONLP2 puisqu'il atteint les valeurs plus hautes de log-vraisemblance. DONLP2 est inévitablement utilisé avec les modèles logit imbriqué croisé puisqu'ils contiennent des contraintes non-triviales sur les paramètres α_{am}[20].

Les modèles suivants s'intéressent à la part d'itinéraire. Le modèle 5.2 (équation (5.2)) offre aux passagers cinq alternatives de classes de services sur leurs itinéraires. Le modèle 5.3 examine une structure imbriquée du modèle 5.2 où les nids sont basés sur le volume de réservations. Finalement, le modèle 5.4 présente la formulation logit imbriqué croisé du modèle 5.3 et permet à une alternative d'appartenir à plusieurs nids simultanément en plus d'estimer son degré d'appartenance à ce nid.

5.2 Logit multinomial

Itinéraires: WITWNZ/R, WNZWIT/R, avec règle de circuité.
Période: 8/19/05-11/18/05, répartis selon 1, 234, 5, 67.
Passagers: payants seulement avec limites (5000, 2500, 2500, 2500, 2500).

[20] Ces paramètres représentent le degré d'appartenance de l'alternative b au nid m et respectent les contraintes suivantes: $\sum_{b \in C_{mi}} \alpha_{bm} = 1, \forall m$ et $0 \leq \alpha_{bm} \leq 1, \forall b, m$.

Ce modèle tente de capter les facteurs qui caractérisent le choix des passagers sur 16 sous-échantillons (*j*) de l'échantillon principal (Tableau 5.2). Il est justifié de procéder à la segmentation; comme l'indique le test du ratio de vraisemblance suivant:

$$-2\left(L^*(\hat{\beta}) - \sum_{j=1}^{16} L^*(\hat{\beta})_j\right) = -2(-36665{,}2 + 31884{,}5) = 4821{,}7 > 131{,}8 = \chi^2_{(0.95,160)} \quad (5.1)$$

où l'hypothèse nulle voulant que $\hat{\beta}_j = 0, \forall j$, est rejetée. Puisqu'un des buts est d'étudier l'impact du tarif sur son choix, seuls les passagers payants sont considérés. De plus, puisque les deux villes à l'étude sont reliées par plusieurs vols directs, une règle de circuité est appliquée selon laquelle: Tout itinéraire dont le temps en transit total dépasse deux fois la durée du plus long vol direct entre les deux villes ce jour-là sont exclus de l'échantillon. Les itinéraires sont ainsi réduits à de simples vols directs.

Tableau 5.2: Définitions des modèles.

ITINÉRAIRES	ORIGINE	RETOUR	JOURS D'OPÉRATION
WITWNZ-EX-52-WIT-1	WIT	NON	LUNDI
WITWNZ-EX-52-WIT-2	WIT	NON	MARDI, MARCREDI, JEUDI
WITWNZ-EX-52-WIT-3	WIT	NON	VENDREDI
WITWNZ-EX-52-WIT-4	WIT	NON	SAMEDI, DIMANCHE
WITWNZ-EX-52-WNZ-1	WNZ	NON	LUNDI
WITWNZ-EX-52-WNZ-2	WNZ	NON	MARDI, MARCREDI, JEUDI
WITWNZ-EX-52-WNZ-3	WNZ	NON	VENDREDI
WITWNZ-EX-52-WNZ-4	WNZ	NON	SAMEDI, DIMANCHE
WITWNZ-EX-52-WITR-1	WIT	OUI	LUNDI
WITWNZ-EX-52-WITR-2	WIT	OUI	MARDI, MARCREDI, JEUDI
WITWNZ-EX-52-WITR-3	WIT	OUI	VENDREDI
WITWNZ-EX-52-WITR-4	WIT	OUI	SAMEDI, DIMANCHE
WITWNZ-EX-52-WNZR-1	WNZ	OUI	LUNDI
WITWNZ-EX-52-WNZR-2	WNZ	OUI	MARDI, MARCREDI, JEUDI
WITWNZ-EX-52-WNZR-3	WNZ	OUI	VENDREDI
WITWNZ-EX-52-WNZR-4	WNZ	OUI	SAMEDI, DIMANCHE

Le passager fait face à un choix multiple: utiliser une classe de service (produit) pour son itinéraire parmi l'ensemble qui est offert. La variable dépendante (*choix*) prend la valeur 1 si le passager choisit la classe de service *1*, 2 si le passager la classe de service *2* et ainsi de suite jusqu'à 5 pour la classe de service *5*. Lors de l'estimation, l'alternative 1, classe de service *1*, est normalisée et les coefficients estimés sont interprétés par rapport à celle-ci. Chacune des alternatives n'est pas disponible à tout

passager à tout moment. De plus, le tarif payé peu inclure un vol retour. Il faut donc regrouper les itinéraires selon les caractéristiques suivantes:

- *RETOUR*: si l'itinéraire fait partie d'un aller-retour ou pas,
- *CONNEXION*: si l'itinéraire contient une connexion ou pas,
- *DATE DE DÉPART*: jour où se fait du premier segment (vol) de l'itinéraire.

Par conséquent, les alternatives offertes au passager ont toutes les mêmes caractéristiques que l'alternative choisie. Il n'est peut-être pas possible de choisir certaines alternatives en période de pointe *MA* un certain jour si aucune autre réservation n'a été faite à ce moment-là. Similairement, le passager qui a payé un certain montant pour son billet aller-retour ne considérait pas les tarifs aller-simples des autres alternatives au moment de son choix. En réalité, c'est de moins en moins vrai mais avec le jeu de données sous-jacent, il est nécessaire d'imposer ces restrictions puisque sinon, l'ensemble d'alternatives comporterait des alternatives fictives ou encore, des alternatives dont le coût est nul et qui ne seraient pas choisies par le passager.

Les composantes déterministes des utilités des cinq alternatives sont définies comme suit:

$$\begin{aligned}
V_1 &= \beta_{CF} CF + \beta_{MA} MA + \beta_{SS} SS + \beta_{TA} TA + \beta_{TT} TT \\
V_2 &= ASC_2 + \beta_{CF} CF + \beta_{MA} MA + \beta_{SS} SS + \beta_{NP} NP + \beta_{TA} TA + \beta_{TT} TT \\
V_3 &= ASC_3 + \beta_{CF} CF + \beta_{MA} MA + \beta_{SS} SS + \beta_{NP} NP + \beta_{TA} TA + \beta_{TT} TT \\
V_4 &= ASC_4 + \beta_{CF} CF + \beta_{MA} MA + \beta_{SS} SS + \beta_{NP} NP + \beta_{TA} TA + \beta_{TT} TT \\
V_5 &= ASC_5 + \beta_{CF} CF + \beta_{MA} MA + \beta_{SS} SS + \beta_{NP} NP + \beta_{TA} TA + \beta_{TT} TT
\end{aligned} \tag{5.2}$$

Il n'est pas possible d'inclure tous les attributs puisqu'il y aurait multicollinéarité (*multicollinearity*), i.e., corrélation entre les variables. Pour que les variables explicatives soient indépendantes et que l'effet de chacune d'elle soit reflété dans le modèle, il ne doit pas être possible d'exprimer une variable comme combinaison linéaire des autres variables. Même deux variables qui prennent étroitement les mêmes valeurs peuvent être problématiques. Or ici, les relations suivantes existent:

- $LO = CI = BO$ donc, CI et BO sont retirées;
- $RE + LO = 1$ donc, LO est retirée;
- $CO = RF - PU$ donc, RF et PU sont retirées;
- $MA + MI + PM + SO = 1$ donc, SO est retirée.

De plus, la variable AM peut être écrite comme une fonction non linéaire de plusieurs variables et est donc retirée. Finalement, CO et RE, ainsi que MI et PM, varient très étroitement ensemble et somment presque toujours à 1. Elles sont donc retirées, au même titre que SA qui est corrélée avec toutes les autres variables. Les résultats de l'estimation sur l'itinéraire WITWNZ_EX_52_WNZ_4 sont affichés au Tableau 5.3. Dans la dernière colonne, lorsque l'étoile apparaît à côté d'une valeur, ceci indique que le coefficient estimé associé n'est pas significativement différent de zéro au sens statistique. Par abus de langage, la tournure "différent de zéro" sera utilisée pour désigner le cas contraire.

Tableau 5.3: Modèle 5.2.

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	ASC_2	-0,80	13,88	-0,06*
2	ASC_3	-0,27	13,85	-0,02*
3	ASC_4	0,78	22,96	0,03*
4	ASC_5	-0,26	27,44	-0,01*
5	β_{CF}	0,02	0,18	0,10*
6	β_{MA}	0,03	0,12	0,22*
7	β_{NP}	0,03	0,08	0,36*
8	β_{SS}	-0,29	--	-0,02*
9	β_{TA}	0,00	0,00	-3,34
10	β_{TT}	-0,28	0,46	-0,62*

Statistiques générales
nombre d'observations = 1826
log-vraisemblance(0) = -2236,97
log-vraisemblance($\hat{\beta}$) = -1109,87
$\bar{\rho}^2$ = 0,50

Avec cette spécification du modèle, aucun coefficient relié aux caractéristiques de la classe n'est différent de zéro. Le coefficient β_{TA} est positif et différent de zéro indiquant que plus le tarif est élevé, moins l'alternative est attirante pour le passager.

Des résultats similaires sont obtenus pour les 15 autres modèles (Tableau 5.4). La plupart du temps, d'autres coefficients sont aussi significatifs. Par exemple, lorsque β_{NP} est positif est significatif, plus le passager voyage en gros groupe, moins il optera pour l'alternative. B_{CF} est positif et différents de zéro parfois ce qui est contre-intuitif puisqu'en général, un passager ne devrait pas préférer les alternatives où les frais de modification de réservation sont élevés. Cette variable peut aussi capter d'autres effets tel qu'être associé aux tarifs les plus bas ou à une action de vente. Lorsque β_{TT} est significatif, il prend généralement des valeurs très grandes par rapport aux autres coefficients. Ceci est signe de multicollinéarité. Cependant, l'effet sur le $\overline{\rho}^2$ n'est pas dramatique et l'impact sur la valeur des autres coefficients n'est pas vraiment visible. *TT* varie très peu si les tous les vols de l'échantillon sont directs, comme c'est le cas WITWNZ. Elle peut être retirée dans ces rares cas mais, en général, *TT* varie et contribue au pouvoir explicatif du modèle.

Tableau 5.4: Résultats des 16 modèles multinomial logit.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
ASC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_CF	-	0,02	-	-	-	-	-	-	0,02	-	0,02	-	-	-	-	-
beta_MA	-	-	-	-	-	-	0,38	-	-	0,18	-	-	-	0,73	0,53	-
beta_NP	-	-	-	0,74	-	-	-	-	-	1,04	-	0,72	-	-	0,88	-
beta_SS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_TA	0,00	-	0,00	0,00	0,00	0,00	-	0,00	-	-	-	0,00	-	0,27	-	0,00
beta_TT	-	29,67	20,69	37,53	-	-	14,79	-	-	-	-	-	-	-	53,54	-

En termes de statistiques générales, ces modèles arrivent à expliquer 47% (en moyenne) de la log-vraisemblance originale, ce qui est considérable, mais aucun des attributs liés aux classes de service n'y contribue vraiment.

Tableau 5.5: Statistiques générales des modèles multinomial logit.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
param.	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
obs.	1066	2986	975	2076	1042	3122	1056	1826	2344	6179	2262	4047	2255	6545	2554	4181
L*(0)	-1385,24	-3673,25	-1196,71	-2424,59	-1172,80	-3925,96	-1400,14	-2236,97	-3547,48	0,00	-3351,07	-5654,72	-3234,51	-9522,27	-3664,43	-5956,99
L*(beta)	-683,98	-1836,49	-596,24	-1199,09	-615,57	-2058,79	-758,67	-1109,87	-1847,21	-4854,86	-1671,08	-2802,41	-1712,29	-5125,42	-2005,45	-2966,17
rho-aj.	0,50	0,50	0,49	0,50	0,47	0,47	0,45	0,50	0,48	0,45	0,50	0,50	0,47	0,46	0,45	0,50

Les résultats du modèle 5.2 sont un peu décevants puisque le modèle comporte plusieurs attributs et a été protégé de plusieurs sources de multicollinéarité possibles. La propriété d'indépendance des alternatives non-pertinentes (*IIA*) n'est sans doute

pas respectée ici. Sans cette propriété, les modèles logit multinomial échouent. Puisque la plupart des itinéraires sont entrepris dans les deux dernières classes de service (*4* et *5*), il faut vérifier si elles ne sont, au fait, qu'une seule et même alternative. Pour se faire, Ben-Akiva et Lerman (1985, p.183) proposent un test pour révéler ceci. En voici un résumé:

- à partir des coefficients estimés et des données utilisées, calculer les utilités (V_b^i) et probabilités $\Pr_{C_i}^i(b)$ de chaque alternative pour chaque observation,
- déterminer le sous-ensemble d'alternatives susceptibles d'être corrélées, C={4,5},
- calculer la statistique: $Z_C = \frac{\sum_{b \in C_i} \Pr_{C_i}^i(b) V_b^i}{\sum_{b \in C_i} \Pr_{C_i}^i(b)} - V_b^i, \forall b \in C$,
- ajouter les nouvelles variables Z_C créées à l'étape précédente aux données initiales sous formes de variables explicatives. Ici, Z_{13}, Z_{23}, Z_{33}, Z_{45} et Z_{55},
- ajouter un coefficient β au modèle initial. Ici, $\beta_{\text{IIA\_13}}$ et $\beta_{\text{IIA\_45}}$,
- estimer le tout.

Le signe du nouveau paramètre IIA estimé n'est pas important. Par contre, la valeur de la statistique-*t* doit être telle que le β associé soit différent de zéro pour que la propriété IIA ne soit pas respectée pour les alternatives du sous-ensemble *C*. Le Tableau 5.6 contient les résultats de l'estimation pour le modèle 5.2 sur l'itinéraire WITWNZ_EX_52_WNZ_4 avec les paramètres IIA.

Tableau 5.6: Modèle 5.2 (avec paramètres IIA).

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	ASC_2	-0,86	1,65	-0.52*
2	ASC_3	-0,27	--	--
3	ASC_4	0,81	--	1.12*
4	ASC_5	-0,27	--	-0.07*
5	β_{CF}	0,02	0,03	0.74*
6	$\beta_{\text{IIA\_13}}$	0,00	--	0.09*
7	$\beta_{\text{IIA\_45}}$	0,00	0,00	-2.06
8	β_{MA}	0,03	--	0.26*
9	β_{NP}	-0,01	0,10	-0.14*
10	β_{SS}	-0,32	1,66	-0.19*
11	β_{TA}	0,00	0,00	-3.13
12	β_{TT}	-0,07	0,26	-0.26*

β_{IIA_45} n'est pas significativement différent de zéro au seuil de 95%. Ceci indique que la propriété IIA est respectée pour les alternatives *4* et *5*. Ceci indique que la propriété IIA est respectée pour les alternatives *4* et *5* mais pas pour les alternatives *1* à *3*. Les résultats des 15 autres modèles sont présentés au Tableau 5.7. L'impact sur la valeur des coefficients est négligeable dans certains cas mais, en général, plus de coefficients sont significatifs. En termes de statistiques générales, les valeurs des $\bar{\rho}^2$ sont un peu plus grandes pour deux raisons.

Tableau 5.7: Résultats des 16 modèles 5.2 (avec paramètres IIA).

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	#VALUE!
ASC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC2	-	-	-	-	-	-	-	-	-	-	-	-	-	-3,40	-	-
ASC3	-	-	-	-	-	-	-	-	-	-	-	-	-	-2,25	-	-
ASC4	-	-	-	-	-	-	-	-	-	-	-	-	-	5,42	-	-
ASC5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_CF	-	-	-	-	-	-	0,02	-	-	-	0,02	-	-	-	-	-
beta_IIA_Z13	-	-	-	-	-	-	-	-	-	0,70	-	-	2,57	1,46	0,44	-
beta_IIA_Z45	-	-0,46	-	-	-	0,00	-0,69	0,00	-	-	-	-	-1,24	-6,48	-1,17	-
beta_MA	-	-	-	-	0,61	-	-	-	-	-	-	-	0,50	-0,43	-	-
beta_NP	-	-	-	0,97	-	-	-	-	-	1,44	-	0,72	1,87	1,31	1,17	-
beta_SS	-	-	-	-	-	-	-	-	-	-	-1,13	-	-	-	-	-
beta_TA	0,00	-	0,00	0,00	-	0,00	-	0,00	-	-	-	0,00	-	-	-	0,00
beta_TT	-	19,73	20,31	39,26	29,88	-	12,97	-	142,80	23,85	-	-	52,71	-	33,83	-

Tout d'abord, le nombre d'observations a diminué puisque le calcul des Z_C n'est pas toujours possible. Moins d'observations impliquent des valeurs de la log-vraisemblance initiale plus grandes. Ensuite, puisque les modèles possèdent deux variables explicatives supplémentaires, les valeurs de la log-vraisemblance finale seront aussi plus grandes. L'impact sur les valeurs de $\bar{\rho}^2$ est positif et le pouvoir explicatif de ces modèles avec β_{IIA_13} et β_{IIA_45} est légèrement supérieur que les modèles originaux.

Tableau 5.8: Statistiques générales des modèles 5.2 (avec paramètres IIA).

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	#VALUE!
param.	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
obs.	988	2596	912	1812	819	2856	906	1574	2340	6155	2261	4036	2244	6514	2544	4164
L*(0)	-1327,31	-3375,76	-1136,86	-2127,54	-992,63	-3612,74	-1233,49	-1935,15	-3545,40	-8854,84	-3351,07	-5645,01	-3233,13	-9519,50	-3664,43	-5938,49
L*(beta)	-649,10	-1664,73	-560,11	-1054,20	-511,41	-1908,05	-666,88	-964,19	-1723,46	-4771,20	-1670,00	-2798,13	-1571,90	-5095,68	-1958,45	-2954,37
rho-aj.	0,50	0,50	0,50	0,50	0,47	0,47	0,45	0,50	0,51	0,46	0,50	0,50	0,51	0,46	0,46	0,50

5.3 Logit imbriqué

Le modèle suivant tenter de corriger cette corrélation qui existe entre les alternatives en introduisant des nids basés sur le volume de réservations. La structure imbriquée est:

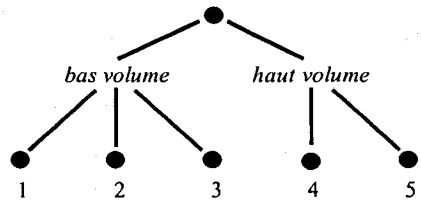


Figure 5.1: Structure imbriquée de nids.

où le nid *bas volume* contient les trois premières alternatives alors que nid *haut volume* contient les deux dernières. Les résultats de l'estimation sur l'itinéraire WITWNZ_EX_52_WNZ_4 sont affichés au Tableau 5.9. Comme précédemment, β_{TA} est négatif et différent de zéro. C'est d'ailleurs un résultat souhaité dans tous les modèles. Ainsi, plus une alternative est dispendieuse, moins elle est attirante. Si le passager n'est pas sensible aux différences de prix entre les alternatives, le transporteur pourrait changer le prix d'une alternative sans voir de différence significative en termes de part de marché.

Tableau 5.9: Formulation logit imbriquée du modèle 5.2.

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	ASC_2	-0,94	--	0,00*
2	ASC_3	-0,53	--	0,00*
3	ASC_4	1,74	--	0,00*
4	ASC_5	-0,58	--	0,00*
5	B_{CF}	0,03	--	--
6	B_{MA}	0,01	--	0,35*
7	B_{NP}	0,03	0,09	0,33*
8	B_{SS}	0,27	--	0,00*
9	B_{TA}	0,00	0,00	-3,23
10	B_{TT}	0,01	--	0,23*
11	μ_{BAS}	1,00	85,50	85,50
12	μ_{HAU}	10,00	15091,94	15091,94

Statistiques générales
nombre d'observations = 1826
log-vraisemblance(0) = -1099,33
log-vraisemblance($\hat{\beta}$) = -217,93
$\bar{\rho}^2$ = 0,50

Le paramètre μ/μ_m reflète le degré de corrélation parmi les portions inobservables de l'utilité pour les alternatives du nid C_m. Pour le modèle au Tableau 5.9, μ_{HAU} a été estimé et vaut 10 (μ_{BAS} vaut 0). Puisque μ est normalisé à 1, le ratio μ/μ_{HAU} vaut 1/10 = 0,1. En théorie, $0 < \mu/\mu_m < 1$de sorte que le résultat obtenu est en accord avec la théorie. Ayant estimé μ_{HAU}, BIOGEME fait les deux tests suivants:

1. Hypothèse nulle H_0: $\mu_m = 0$. Rejette H_0 si $\left| \frac{(\mu_m - 0)}{\text{écart - type de } \mu_m} \right| > t_{\text{valeur critique}}$;

2. Hypothèse nulle H_0: $\mu_m > 1$. Rejette H_0 si $\left| \frac{(\mu_m - 1)}{\text{écart - type de } \mu_m} \right| > t_{\text{valeur critique}}$.

Dans les deux tests, H_0 est rejetée et l'hypothèse suivante est validée: il y a corrélation des facteurs inobservables parmi les alternatives du nid *haut volume*. Le modèle logit imbriqué est donc préféré au modèle logit multinomial.

Tableau 5.10: Formulation logit imbriquée du modèle 5.2.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
ASC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_CF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_MA	-	-	-	-	-	0,05	0,05	-	-	0,18	-	-	0,24	0,02	0,20	-
beta_NP	-	-	-	-	-	-	-	-	-	1,04	-	0,73	2,21	0,73	0,94	-
beta_SS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_TA	-	-	0,00	0,00	0,00	0,00	0,00	0,00	-	-	-	0,00	-	-	0,00	-
beta_TT	-	-	8,75	21,83	-	3,22	3,37	-	-	-	-	-	10,74	0,38	26,91	-
bas	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hau	-	-	10,00	-	-	10,00	10,00	10,00	-	-	-	10,00	1,01	-	3,64	-

Le Tableau 5.10 présente les résultats des 15 autres modèles logit imbriqué. La plupart du temps, les quatre coefficients reliés à la pointe du matin (*MA*) sont significatifs et parfois autrement, β_{NP} et β_{TT} sont positifs et différents de zéro. Les paramètres μ_{BAS} ne sont pas différents de zéro indiquant que les trois alternatives du nid *bas volume* sont distinctes et pertinentes. C'est moins vrai pour les alternatives du nid *haut volume*.

Les statistiques générales au Tableau 5.11 révèlent des $\bar{\rho}^2$ de 0,48. Ceci peut être dû au fait que la structure de nids imposée est rigide. La démarche a été d'assigner une alternative à un nid et si cette assignation initiale s'avère correcte, les $\bar{\rho}^2$ en sont marginalement haussé. Si, par contre, la formulation logit imbriqué change la valeur des coefficients estimés de façon marquante, une chute importante dans la valeur du $\bar{\rho}^2$ est aussi observée. Donc, la formulation logit imbriquée, de par sa structure rigide, n'est pas toujours supérieure à la formulation multinomiale logit sous-jacente. Le modèle imbriqué croisé présenté à la section suivante est une relaxation de celui-ci et devrait produire de meilleurs résultats dans les 16 cas.

Tableau 5.11: Statistiques générales des modèles logit imbriqués.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
param.	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
obs.	1066	2986	975	2076	1042	3122	1056	1826	2344	6179	2262	4047	2255	6545	2554	4181
L*(0)	-1385,24	-3673,25	-1196,71	-2424,59	-1172,80	-3925,96	-1400,14	-2236,97	-3547,48	-8856,23	-3351,07	-5654,72	-3234,51	-9522,27	-3664,43	-5956,99
L*(beta)	-683,95	-1828,48	-582,04	-1196,52	-612,79	-2013,87	-744,97	-1099,33	-1847,21	-4854,86	-1671,08	-2794,98	-1721,87	-5117,66	-1959,39	-2956,98
rho-aj.	0,50	0,50	0,50	0,50	0,47	0,48	0,46	0,50	0,48	0,45	0,50	0,50	0,46	0,46	0,46	0,50

5.4 Logit imbriqué croisé

Le modèle logit imbriqué croisé présenté à cette section s'inspire du modèle précédent et tente d'améliorer les résultats obtenus jusqu'ici en permettant à une alternative *i* d'appartenir à plusieurs nids, avec un certain degré d'appartenance α_{am}. Le concept de nids basés sur le volume de réservations est conservé. La structure imbriquée croisée est:

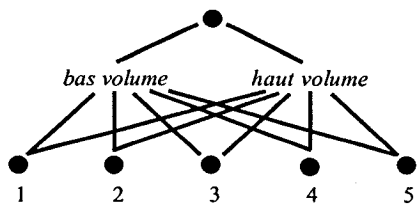


Figure 5.2: Structure imbriquée croisée de nids.

où le nid *bas volume* contient toutes les alternatives et le nid *haut volume* aussi. Il faut noter que la condition $\sum_m \alpha_{im} = 1$ a été imposée. Une telle condition n'est pas nécessaire pour la validité du modèle mais semble souhaitable (Abbé *et al.*, 2005). Les résultats au Tableau 5.12 montrent que l'alternative *1* appartient aux deux nids avec une proportion de 0,88 pour le nid *bas volume* et 0,12 pour le nid *haut volume*. Auparavant, l'alternative *1* appartenait uniquement au nid *bas volume*. L'alternative *4* appartient aux deux nids avec une proportion de 0,77 pour le nid *bas volume* et 0,23 pour le nid *haut volume*. Auparavant, l'alternative *4* appartenait uniquement au nid *haut volume*. β_{TA} est négatif et différent de zéro indiquant que plus une alternative est dispendieuse, moins elle est attirante. Le Tableau 5.13 présente les détails des 15 autres modèles imbriqués croisés. Les coefficients α_{im} sont significatifs parfois et varient d'un modèle à l'autre. β_{NP} et β_{MA} sont généralement positifs et différents de zéro pour les itinéraires à partir de WNZ.

Tableau 5.12: Formulation logit imbriquée croisée du modèle 5.2.

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	ASC_2	-1,29	--	0,00*
2	ASC_3	-0,41	--	--
3	ASC_4	0,80	--	0,00*
4	ASC_5	-0,27	342,67	0,00*
5	B_{CP}	0,01	--	--
6	B_{MA}	0,01	--	0,43*
7	B_{NP}	0,00	0,03	-0,15*
8	B_{SS}	-0,90	--	0,00*
9	B_{TA}	0,00	0,00	-4,78
10	B_{TT}	0,02	0,04	0,58*
11	μ_{BAS}	10,00	--	0,00
12	μ_{HAU}	1,00	--	--
13	α_{1BAS}	0,88	0,04	24,61
14	α_{2BAS}	0,52	--	0,00
15	α_{3BAS}	0,68	--	0,00
16	α_{4BAS}	0,77	0,07	11,61
17	α_{5BAS}	0,38	0,08	4,80
18	α_{1HAU}	0,12	0,04	3,21
19	α_{2HAU}	0,48	--	0,00
20	α_{3HAU}	0,32	--	0,00
21	α_{4HAU}	0,23	0,07	3,50
22	α_{5HAU}	0,62	0,08	7,76

Statistiques générales
nombre d'observations = 1826
log-vraisemblance(0) = -2236,97
log-vraisemblance($\hat{\beta}$) = -1080,54
$\bar{\rho}^2$ = 0,51

Dans le cas β_{NP}, l'interprétation est la suivante: plus le passager doit voyager en gros groupe, plus il optera pour une des autres classes de service que la classe de service *1*.

Tableau 5.13: Résultats des 16 modèles logit imbriqués croisés.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
ASC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ASC5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_CF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_MA	-	-	-	-	-	0,08	0,07	-	-	-	-	-	0,08	-	0,07	0,05
beta_NP	-	-	-	-	-	-	-	-	-	0,12	-	-	-	0,73	0,91	0,37
beta_SS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
beta_TA	-	-	0,00	0,00	0,00	0,00	0,00	0,00	-	-	-	0,00	0,00	-	0,00	0,00
beta_TT	-	-	16,77	24,94	-	5,57	5,60	-	-	-	-	-	7,13	0,27	10,17	5,89
bas	-	-	-	-	-	-	10,00	-	-	-	-	-	-	-	10,00	-
hau	-	-	-	10,00	-	-	-	-	-	10,00	-	-	-	-	-	-
bas_Alt1	-	-	-	0,15	-	-	-	0,88	-	0,00	-	-	0,05	-	0,91	-
bas_Alt2	-	-	0,90	-	-	-	-	-	-	-	-	-	-	-	-	-
bas_Alt3	-	-	0,93	-	-	-	-	-	-	-	-	-	-	-	0,94	-
bas_Alt4	-	-	-	0,39	-	-	0,89	0,77	-	-	-	-	-	-	-	-
bas_Alt5	-	-	0,46	-	-	-	-	0,38	-	-	-	-	-	-	-	-
hau_Alt1	-	-	-	0,85	-	-	-	0,12	-	-	-	-	0,95	-	0,09	-
hau_Alt2	-	-	0,10	-	-	-	-	-	-	-	-	-	-	-	-	-
hau_Alt3	-	-	0,07	-	-	-	-	-	-	-	-	-	-	-	0,06	-
hau_Alt4	-	-	-	0,61	-	-	0,11	0,23	-	-	-	-	-	-	-	-
hau_Alt5	-	-	0,54	-	-	-	-	0,62	-	-	-	-	-	-	0,00	0,00

Les statistiques générales au Tableau 5.14 révèlent des $\bar{\rho}^2$ un peu plus bas qu'avec la formulation précédente; ce qui ne concorde pas tout à fait avec les attentes initiales.

Tableau 5.14: Statistiques générales des modèles logit imbriqués croisés.

	WIT_1	WIT_2	WIT_3	WIT_4	WNZ_1	WNZ_2	WNZ_3	WNZ_4	WITR_1	WITR_2	WITR_3	WITR_4	WNZR_1	WNZR_2	WNZR_3	WNZR_4
param.	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
obs.	1066	2986	975	2076	1042	3122	1056	1826	2344	6179	2262	4047	2255	6545	2554	4181
L*(0)	-1385,24	-3673,25	-1196,71	-2424,59	-1172,80	-3925,96	-1400,14	-2236,97	-3547,48	-8856,23	-3351,07	-5654,72	-3234,51	-9522,27	-3664,43	-5956,99
L*(beta)	-683,98	-1828,22	-577,89	-1190,59	-611,53	-1983,04	-739,41	-1080,54	-1847,21	-4850,00	-1671,08	-2793,06	-1603,08	-5125,42	-1946,16	-2830,40
rho-aj.	0,49	0,50	0,50	0,50	0,46	0,49	0,46	0,51	0,47	0,45	0,49	0,50	0,50	0,46	0,46	0,52

En conclusion, le bilan des valeurs des $\overline{\rho}^2$ est présenté à la Figure 5.3. Les trois modèles produisent des valeurs des $\overline{\rho}^2$ semblables sauf que le modèle multinomial logit a moins de variables explicatives. La propriété d'indépendance des alternatives non-pertinentes n'est pas toujours respectée et dans ces cas, la formulation logit imbriquée devrait être utilisée. De plus, lorsque les nids sont basés sur les volumes de réservations, plusieurs alternatives appartiennent simultanément à deux nids comment en témoignent les modèles logit imbriqués croisés. Finalement, sur les vols allers-retours originaires de WIT, les trois spécifications sont équivalentes alors que la formulation multinomial logit est inférieure dans plusieurs autres cas.

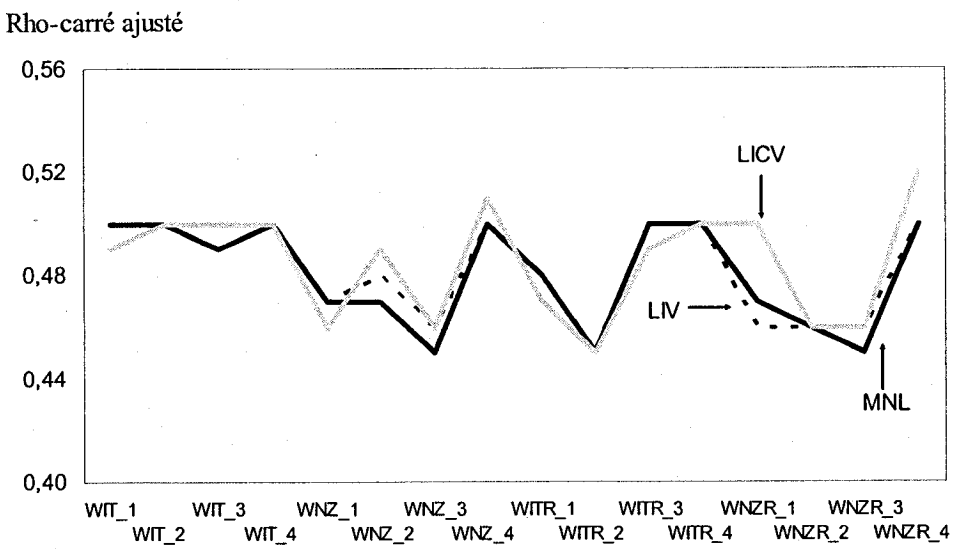


Figure 5.3: Comparaison des rho-carrés ajustés.

CHAPITRE 6: VALIDATION

Ce chapitre est composé de trois sous-sections. À la sous-section 6.1, les trois modèles présentés au chapitre précédent sont validés selon une méthode proposée par Ben-Akiva et Lerman (1985). La sous-section 6.2 présente deux modèles logit binaire qui tentent de valider deux hypothèses.

6.1 Validation

La validation des résultats est une étape importante de toute analyse statistique. Les articles sur les modèles de choix discret appliqués à la demande en transport font souvent référence à une mesure de qualité du modèle nommée "%-correct." Cette statistique s'écrit:

$$\frac{100}{N}\sum_{i=1}^{N}\hat{g}_b^i \tag{6.1}$$

où $\hat{g}_b^i$ prend la valeur 1 si la probabilité prédite la plus haute correspond à l'alternative choisie b et 0 sinon. N représente le nombre d'observations. Les auteurs ne recommandent pas l'utilisation de cette statistique pour l'évaluation des modèles car elle peut masquer le fait que la qualité d'un modèle soit faible.

Par exemple, dans un échantillon de 100 observations, 90 d'entre elles révèlent que l'alternative a est choisie et 10 autres que l'alternative b est choisie, $C_i = \{a,b\}, \forall i$. Le modèle suivant est adopté pour générer les probabilités: $\Pr_{C_i}^i(a) = 0{,}9, \forall i$. La statistique "%-correct" sera de 90% pour ce modèle en dépit du fait que le modèle classifie incorrectement chaque observation où l'alternative b est choisie.

L'approche différente pour calculer "%-correct" est d'utiliser les probabilités calculées:

$$\frac{100}{N}\sum_{i=1}^{N}\sum_{b\in C_i}\Pr_{C_i}^{i}(b)g_b^i \tag{6.2}$$

où g_b^i prend la valeur 1 si l'alternative b est choisie à l'observation i et 0 sinon. Pour le même modèle, cette mesure donnerait: 90(0,9) + 10(0,1) = 82%, ce qui est plus petit que le 90% obtenu précédemment. Cependant, dans cette dernière méthode de prévision, la caractéristique désirable de reproduire les parts des alternatives est maintenue. Ainsi,

$$\frac{1}{N}\sum_{i=1}^{N}\Pr_{C_i}^{i}(b)=\frac{1}{N}\sum_{i=1}^{N}g_b^i\,. \tag{6.3}$$

Or, cette dernière mesure du "%-correct" est mieux représentée par la valeur de la log-vraisemblance qui est la fonction objectif maximisée dans les modèles et est plus sensible aux basses valeur des probabilités prédites pour l'alternative choisie.

En conclusion, la procédure de prédilection pour déterminer la qualité d'un modèle réside dans l'utilisation de la valeur de la log-vraisemblance, $L^*(\hat{\beta})$, ou des transformées de celle-ci telles que ρ^2 et $\bar{\rho}^2$. Le manque de sensibilité de la statistique "%-correct" et son potentiel pour les fausses interprétations sont deux arguments contre son utilisation.

BioSim est un paquetage (*package*) fourni avec BIOGEME qui peut être utilisé pour calculer les probabilités estimées. Le procédé est le suivant:

- à partir des coefficients estimés et des données utilisées, calculer la probabilité pour l'alternative choisie, $\Pr_{C_i}^{i}(a)$, ainsi que la probabilité de chacune des autres alternatives dans l'ensemble de choix, $\Pr_{C_i}^{i}(b), b\neq a, \forall b\in C_i$, pour chaque individu t,
- comparer les $\Pr_{C_i}^{i}(a)$ avec les données utilisées pour déterminer les g^i,

- calculer la valeur de la statistique "%-correct" selon: $\frac{100}{N}\sum_{i=1}^{N}\sum_{b\in C_i}\Pr_{C_i}^{i}(b)g_b^i$.

Cette démarche n'est pas entreprise dans ce mémoire pour les raisons de sensibilité mentionnées plus haut. La qualité des modèles repose sur les $\overline{\rho}^2$. Les trois modèles affichent des $\overline{\rho}^2$ de l'ordre de 0,48 indiquant que les modèles finaux, $L^*(\hat{\beta})$, n'arrivent seulement qu'à expliquer la moitié de la log-vraisemblance originale, $L^*(0)$. Cette statistique dépend du type de modèle estimé et est plus utile pour comparer différentes spécifications de modèles entre elles (utilisant le même échantillon de données) qu'en soi-même.

L'approche utilisée dans cette recherche est différente. Il s'agit de comparer les prévisions obtenues à partir de l'échantillon initial avec les données observées d'un nouvel échantillon de taille égale pour chacun des seize modèles. Plus précisément, il s'agit de:

- calculer la probabilité de chacune des alternatives dans l'ensemble de choix, $\Pr_{C_i}^{i}(b), \forall b \in C_i$, pour chaque individu *i*, à partir des coefficients estimés et des données utilisées,
- calculer la moyenne des probabilités de chacune des alternatives, $\Pr(\bar{b}), b \in C$, pour obtenir la part de marché de l'alternative *b*,
- calculer la part de chaque alternative dans le nouvel échantillon en divisant le nombre de réservations que compte chaque alternative par le nombre total de réservations,
- Prendre la différence entre ces deux parts.

Les résultats apparaissent à la Figure 6.2. Les modèles produisent des prévisions dont le biais est de 4% en moyenne. Le modèle logit imbriqué croisée est moins précis et plus volatile que les deux autres modèles. Le multinomial logit ordinaire produit des résultats plus stables et précis. Les prévisions pour les itinéraires allers-simples se comportent différemment de celles des allers-retours.

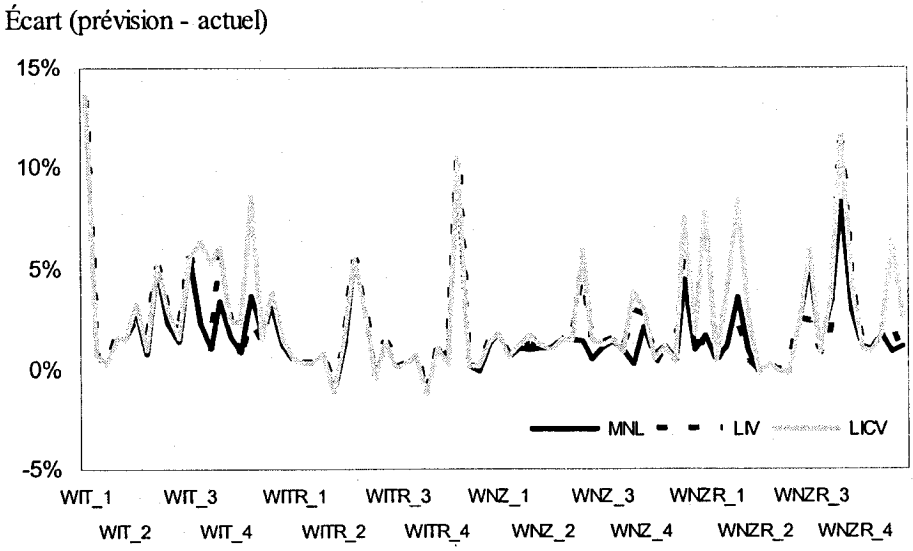


Figure 6.1: Écarts entre les prévisions et données observées.

6.2 Modèles logit binaire

Les deux modèles suivants ne s'intéressent pas à la part d'itinéraire. Ils tentent plutôt de valider deux hypothèses soulevées antérieurement. Tout d'abord, les itinéraires simples entrepris près des limites de l'échantillon risquent de faire partie de voyage allers-retours et d'afficher des tarifs payés plus hauts. Ensuite, certains itinéraires sont entrepris en échange de points de programmes de fidélité et leurs tarifs payés correspondants sont nuls. Or, ces programmes ont souvent des périodes d'invalidité (*blackout*) et n'offrent qu'un nombre limité de sièges aux passagers non-payants. Ces politiques pourraient empêcher les passagers non-payants d'emprunter les vols directs autant qu'ils ne le voudraient.

6.2.1 logit binaire sur les tarifs avec attributs génériques

Itinéraires: WITWNZ, WNZWIT, sans règle de circuité

Période: 8/19/05-11/18/05

Passagers: Payants seulement

Ce modèle tente de capter les facteurs qui caractérisent le choix des passagers ayant payé un tarif jugé trop élevé pour un itinéraire simple. Les données brutes sont telles

que certains itinéraires simples sont plus dispendieux que d'autres puisqu'ils font partie de voyages allers-retours. Or, par analogie, le passager fait face à un choix binaire: payer plus ou moins cher pour un itinéraire simple par rapport à la moyenne observée dans une classe de service donnée. La variable dépendante (*choix*) prend la valeur 1 si le passager a payé plus que la moyenne pour son billet et 0 sinon. Lors de l'estimation, l'alternative de payer moins cher que la moyenne est normalisée et les coefficients estimés sont interprétés par rapport à celle-ci. Les équations suivantes définissent les composantes déterministes des utilités des deux alternatives:

$$\begin{aligned} V_{\text{moins}} &= 0 \\ V_{\text{plus}} &= ASC_{\text{plus}} + \beta_{\text{LE}} LE + \beta_{\text{NP}} NP + \beta_{\text{TT}} TT. \end{aligned} \tag{6.4}$$

Les résultats de l'estimation sont affichés au Tableau 6.1.Étant donné la spécification initiale du modèle, le signe négatif de ASC_{plus} pourrait s'interpréter comme la préférence du passager de payer le tarif moins cher. Le coefficient β_{LE} est positif et différent de zéro indiquant que plus le passager part près des limites de l'échantillon, plus il paye cher pour son billet. Donc, couper un échantillon selon les dates de départ biaise le comportement des passagers près des limites de l'échantillon. Le coefficient β_{LE} peut aussi capter d'autres effets tel le retour des vacances estivales où la demande et les tarifs sont généralement les plus élevés.

Tableau 6.1: Modèle (6.4).

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	ASC_{plus}	-1,03	0,13	-8,00
2	β_{LE}	0,59	0,04	16,90
3	β_{NP}	-0,01	0,03	-3,75
4	β_{TT}	0,13	0,03	5,29

Statistiques générales
nombre d'observations = 14443
log-vraisemblance(0) = -10011,1
log-vraisemblance($\hat{\beta}$) = -9673,25
$\bar{\rho}^2$ = 0,033

Le coefficient β_{NP} est négatif et différent de zéro indiquant que plus le passager voyage en gros groupe, moins il est prêt à payer cher pour son billet. Effectivement, les groupes bénéficient souvent d'un tarif privilégié. Finalement, le coefficient β_{TT} est

positif et différent de zéro indiquant que plus l'itinéraire est long, plus le passager est prêt à payer cher pour son billet. L'interprétation de ce résultat est basée sur les représentations graphiques des voyages types d'affaires et de plaisance (section 4.1.1). Puisque les vols directs affichent les temps en transit les plus courts, ils seront sûrement plus populaires auprès des passagers d'affaires. Les séjours des voyages d'affaires sont courts et ceci réduit les chances qu'un tel voyage soit coupé par la limite de l'échantillon. Par contre, les voyages de plaisance ont plus de chances de chevaucher la limite de l'échantillon puisque les vacanciers se déplacent souvent plus loin, font quelques escales en cours de route et restent à destination plus longtemps. Ces derniers apparaissent donc comme des itinéraires où le temps en transit et le tarif payé sont élevés lorsque comparés à d'autres itinéraires simples.

Il est possible d'analyser l'effet de la variable dichotomique *LE* sur la distribution au complet en calculant Pr(*Payer plus cher*) = Pr(*Choix* = 1) sur l'intervalle de $\hat{\beta}^{\mathrm{T}} x$ (utilisant les estimés de l'échantillon) et avec deux valeurs de LE. Les probabilités suivantes sont fonctions de *TT*, à la valeur moyenne de *NP*:

$$\begin{aligned} LE = 0 &: \Pr(Choix = 1) = \Lambda(-1{,}03 - 0{,}01(1{,}35) + 0{,}14TT) \\ LE = 1 &: \Pr(Choix = 1) = \Lambda(-1{,}03 + 0{,}59 - 0{,}01(1{,}35) + 0{,}14TT) \end{aligned} \tag{6.5}$$

où Λ représente la distribution cumulative de la fonction logistique. La Figure 6.2 montre ces deux fonctions tracées sur un intervalle décalé de *TT* pour faire ressortir la courbure des fonctions. *TT* varie de 1,08 à 48,91 dans l'échantillon alors que sur le graphe, il varie de –15 à 30. L'effet marginal de *LE* est la différence entre les deux fonctions. Ceci montre que la probabilité de payer plus cher pour des itinéraires décollant près des limites de l'échantillon est plus grande que ceux en milieu d'échantillon. À la moyenne de *TT*, soit 4,71, l'effet de *LE* sur la probabilité est 0,145. La dérivée simple à ce point vaut 0,148 mais encore, ceci ne montre la gamme de différences illustrée à la Figure 6.2.

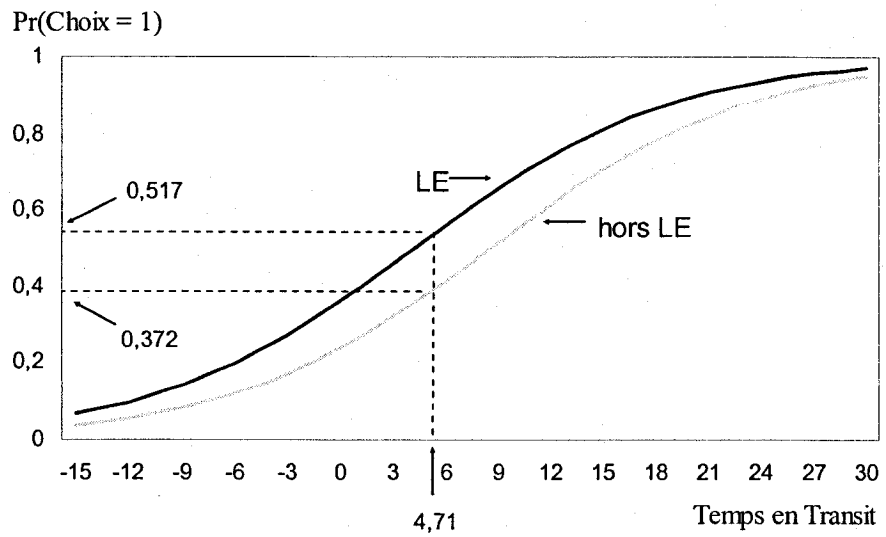


Figure 6.2: Effet de LE sur les probabilités estimées (*TT* en abscisses).

Comme le démontre la Figure 6.3, le même résultat est obtenu en fixant *TT* à sa valeur moyenne et en traçant les fonctions sur un échantillon décalé de *NP*. Les pentes des courbes sont maintenant négatives et à la moyenne de *NP*, soit 1,35, l'effet de *LE* sur la probabilité est toujours 0,145.

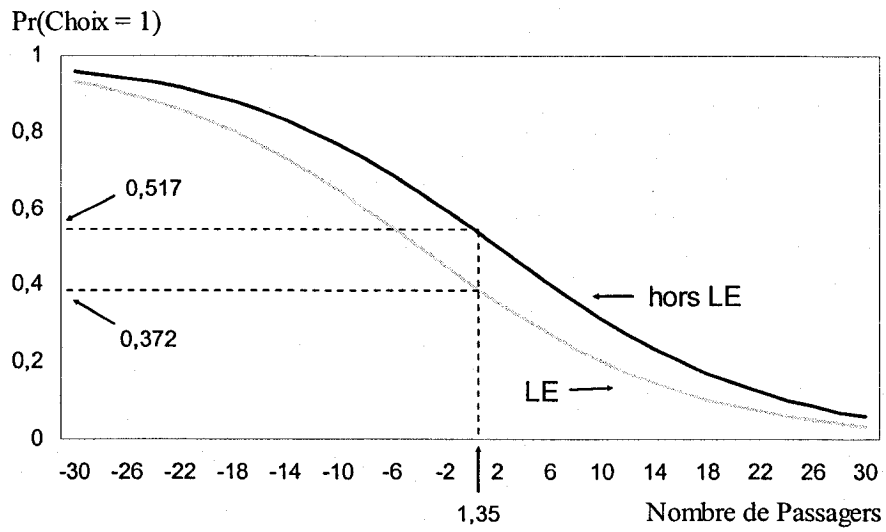


Figure 6.3: Effet de LE sur les probabilités estimées (*NP* en abscisses).

Les statistiques générales montrent que tous les paramètres du modèle, pris ensemble, sont différents de zéro. Ceci s'obtient par le test de ratio de vraisemblance où la

statistique $-2\left(L^*(0)-L^*(\hat{\beta})\right)$ est asymptotiquement distribuée χ^2 avec K degrés de liberté. Ici, K = 4 et la valeur de la statistique est 675,75, indiquant que l'hypothèse nulle: tous les paramètres du modèle prennent la valeur 0, peut être rejetée au niveau de confiance 99%. Cependant, $\overline{\rho}^2$ indique que le modèle final n'arrive seulement qu'à expliquer 3,3% de la log-vraisemblance originale mais encore, cette statistique dépend du type de modèle estimé. Elle est plus utile pour comparer différentes spécifications de modèles entre elles (utilisant le même échantillon de données) qu'en soi-même.

6.2.2 logit binaire sur les récompenses avec attributs génériques

Itinéraires: WITWNZ/R, WNZWIT/R, sans règle de circuité

Période: 8/19/05-11/18/05

Passagers: Tous

Ce modèle tente de capter les facteurs qui caractérisent le choix des passagers ayant payé pour leur itinéraire en utilisant des points des programmes de fidélité. Les données brutes sont telles que certains itinéraires affichent un tarif payé très bas ou même 0. Un tarif en- dessous de $100 en classe affaires ou $50 en classe économie est jugé très bas. Ceci s'applique sur tous les itinéraires simples qu'ils fassent partie de voyages allers-retours ou pas. Le passager fait face à un choix binaire: utiliser un vol direct ou choisir une combinaison de vols sur son itinéraire. La variable dépendante (*connect*) prend la valeur 1 si le passager connecte à un ou plusieurs aéroports et 0 sinon. Lors de l'estimation, l'alternative du vol direct est normalisée et les coefficients estimés sont interprétés par rapport à celle-ci. Les équations suivantes définissent les composantes déterministes des utilités des deux alternatives:

$$\begin{aligned} V_{\text{direct}} &= 0 \\ V_{\text{connect}} &= ASC_{\text{connect}} + \beta_{\text{MU}} MU + \beta_{\text{NP}} NP + \beta_{\text{TT}} TT. \end{aligned} \tag{6.6}$$

Les résultats de l'estimation sont affichés au Tableau 6.2. Avec cette spécification initiale du modèle, le signe négatif de ASC_{connect} indique que toutes choses étant

égales par ailleurs, la préférence du passager est de prendre le vol direct. Le coefficient β_{UM} est négatif et différent de zéro indiquant que si un passager doit payer pour son billet, il optera davantage pour le vol direct. β_{NP} est négatif mais pas différent de zéro indiquant que la taille du groupe voyageant avec le passager a un effet négligeable sur sa décision. Finalement, le coefficient β_{TT} est positif et différent de zéro indiquant que plus l'itinéraire est long, plus le passager est prêt à opter pour un itinéraire avec une connexion ou plus. Ce résultat, bien que significatif, n'est pas pertinent puisque les deux villes à l'étude sont reliées par plusieurs vols directs. Le fait que les règles de circuité ne soient pas utilisées pour réduire la taille l'échantillon n'est pas la cause. En éliminant de l'échantillon les itinéraires plus courts que la durée du vol direct le plus court et ceux plus longs que deux fois la durée du vol direct le plus long, la valeur de β_{TT} ne fait qu'augmenter. En enlevant la variable *TT* du modèle et en estimant de nouveau, le $\overline{\rho}^2$ tombe de 10% et le test de ratio de log-vraisemblance donne 8286. Par conséquent, *TT* améliore le pouvoir explicatif du modèle lorsqu'elle y est incluse.

Tableau 6.2: Modèle (6.6).

NUMERO VARIABLE	NOM VARIABLE	COEFFICIENT ESTIME	ECART-TYPE ROBUSTE	STATISTIQUE-T ROBUSTE
1	$ASC_{connect}$	-7,65	9,34	-8,19
2	β_{UM}	-44,72	0,05	-923,1
3	β_{NP}	-0,12	0,08	-1,54*
4	β_{TT}	13,23	1,71	7,75

Statistiques générales
nombre d'observations = 61284
log-vraisemblance(0) = -42478,8
log-vraisemblance($\hat{\beta}$) = -1182,53
$\overline{\rho}^2$ = 0,972

Les statistiques générales montrent que tous les paramètres du modèle, pris ensemble, sont différents de zéro. $\overline{\rho}^2$ indique que le modèle final arrive à expliquer 97% de la log-vraisemblance originale mais encore, cette statistique dépend du type de modèle estimé. Elle est plus utile pour comparer différentes spécifications de modèles entre elles (sur les mêmes données) comme ce fût le cas avec *TT* omise du modèle.

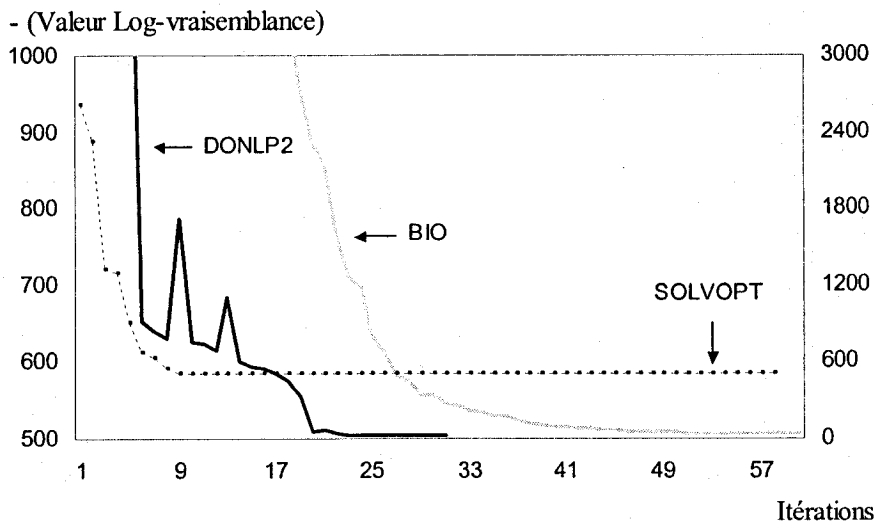


Figure 6.4: Comparaison des méthodes (classes de service régulières).

L'échantillon initial contient des classes tarifaires qui sont réservées aux agences de voyages, aux groupes, etc. Ces classes spéciales ne font pas partie des cinq classes de service régulières et les seuils de $50 et $100 ne leurs s'appliquent pas (*UM* prend toujours la valeur 0). En les enlevant, le même modèle est utilisé sur les observations appartenant à l'une ou l'autre des cinq classes de service régulières. Les résultats sont similaires mais ce qui est intéressant à noter ici est que la méthode BIO est très lente. Les résultats sont obtenus avec DONLP2.

La Figure 6.4 montre le cheminement de la valeur négative de la log-vraisemblance pour les 60 premières itérations en utilisant chacune des trois méthodes disponibles. DONLP2 est la plus rapide à converger. BIO prend 5 fois plus de temps et 8 fois plus d'itérations pour converger que DONLP2. SOLVOPT prend 2 fois plus de temps et 2 fois plus d'itérations pour converger que DONLP2. SOLVOPT est tracée par rapport à l'axe des ordonnées secondaire (celui à droite de la figure) puisqu'elle se confondait avec DONLP2 à certains endroits.

En conclusion, les modèles vus à cette sous-section s'intéressaient aux facteurs influençant le passager face à un choix binaire. Dans le premier modèle, le fait de couper l'échantillon selon une date précise biaisait les résultats. Le modèle (mais

surtout les données) laissait croire que les passagers qui entreprenaient leurs itinéraires simples près des limites de l'échantillon étaient plus disposés à payer leur billet plus cher que ceux en milieu d'échantillon. Il n'y a pas de raison *a priori* de croire cet argument. Le deuxième modèle montrait que les passagers utilisant des points de programmes de fidélité pour payer leurs itinéraires préféraient les vols directs aux itinéraires à une connexion ou plus. Ceci rejoint les résultats de Lederman (2004) où les membres de ces programmes préfèrent utiliser les vols du transporteur dominant à leur aéroport d'origine. En supposant que le transporteur dominant à un certain aéroport est celui qui offre le plus de vols directs, ce qui est généralement le cas, les passagers membres de ces programmes préfèreront échanger leurs points contre des sièges sur les vols directs de celui-ci.

CHAPITRE 7: CONCLUSION

Trois modèles de choix discret pour analyser la demande des passagers ont été présentés dans ce mémoire. Dans le contexte de l'analyse de la demande de transport, les modèles de choix discret ont joué un rôle important au cours des 25 dernières années. L'échantillon de données provenait d'un grand transporteur aérien nord-américain et permettait de définir des itinéraires (déplacement unidirectionnel entre deux villes), puis de comparer les différentes classes de service offertes aux passagers sur ces itinéraires.

La démarche suivie a été de commencer par un modèle logit multinomial relativement simple et d'en examiner les résultats. La restriction de ce modèle venait du fait que la propriété d'indépendance des alternatives pertinentes était violée. Or, ceci impliquait que tous les classes de service au sein d'un même groupe (paire de villes, jour de la semaine) ne se concurrençaient pas de façon équivalente les unes avec les autres.

Ensuite, dans le but d'améliorer ces résultats, des formulations logit imbriquée et logit imbriquée croisée ont été présentées. L'hypothèse suivante à été démontrée dans plusieurs sous-échantillons: les classes de service partageant un volume similaire de réservations montrent un plus grand niveau de substitution entre elles qu'avec les classes de service ne partageant pas cette même caractéristique. Ainsi, les classes où le volume de réservations est élevé (ou bas) se concurrencent davantage entre elles.

L'approche innovatrice de ce mémoire est le nombre et la gamme d'attributs reliés aux alternatives et aux individus qui sont utilisés dans les modèles. Chacun des trois modèles arrive à expliquer 48% de la log-vraisemblance originale dans son échantillon respectif. Les résultats montrent que les attributs reliés à la classe de service ne sont pas statistiquement significatifs dans le choix de celles-ci pour le passager. Par contre, le fait de voyager en groupe réduit la probabilité de choisir la classe de service reliée au compartiment affaires (*business class*).

Les structures de nids utilisées dans ce mémoire sont simples et basées sur les volumes de réservations. Une amélioration possible serait d'introduire un nid basé sur le tarif payé (*eg.* tarif haut, tarif bas) et de tester le modèle logit imbriqué à deux niveaux ou encore, le logit imbriqué croisé avec quatre nids. Néanmoins, les modèles utilisés font ressortir la corrélation qui existe au sein des nids et il faut surtout éviter de faire des nids avec des attributs qui servent de variables explicatives dans un même modèle.

Les attributs reliés à la classe de service sont représentés sous forme de variables dichotomiques et le fait d'ajouter une dizaine de variables dichotomiques dans un modèle de choix ne fait pas ressortir les différences de perception des passagers entre elles. L'autre difficulté à travailler avec de telles variables est d'utiliser des moyennes pour représenter les niveaux de ces variables pour les alternatives non-choisies au sein de C_i. Par exemple, un billet peut être remboursable (ou non) mais pas 0,28 remboursable. Finalement, les valeurs {0,1} que prennent ces variables, lorsque prises ensemble, ont une certaine structure. Elles donnent au passager le droit à certains privilèges et ces privilèges s'accumulent d'une classe de service à l'autre. Ainsi, uniquement certains combinaisons de ces attributs sont possibles, pas toutes. Il y a donc risque de corrélation parmi les attributs. Une amélioration consisterait à tester les attributs individuellement (un à un) dans les modèles ou encore, en prendre quelques-uns parmi ceux qui jugés non-corrélées et de faire des nids avec les autres.

Finalement, le jeu de données brutes utilisé demande beaucoup de traitements et d'hypothèses avant l'étape de modélisation. Le fait que le tarif payé soit exprimé par passager, pour le voyage au complet, nécessite de séparer l'échantillon selon la direction du vol initial et de considérer si le retour est compris dans le tarif ou pas. Bien que les compagnies aériennes aient en place un système de *pro-rata* du millage

volé lorsqu'elles séparent les revenus des voyages inter-lignes[21], diviser le tarif payé par le nombre de segments contenus dans le PNR pour obtenir un tarif moyen par personne par segment n'est pas recommandé pour deux raisons. Tout d'abord, les coûts variables d'un transporteur augmentent avec la distance parcourue et en s'appuyant sur la théorie économique voulant qu'un transporteur fixe ses tarifs au seuil où il couvre ses coûts et sa marge de profit, traiter tous les segments comme s'ils étaient de distance égale est non seulement peu probable en réalité mais, contrevient à la théorie économique. Ensuite, si les aéroports n'étaient pas fictifs et que les distances entre eux étaient connues, diviser le tarif payé au *pro-rata* du millage volé à chaque segment par rapport au millage total du voyage serait une meilleure mesure de la contribution de chaque segment que la moyenne ne l'est. Cependant, le tarif moyen par segment (ou au *pro-rata* du millage volé) peut surestimer, dans le cas d'une guerre des prix sur une route, ou sous-estimer, dans le cas d'un monopole sur une route, les tarifs en vigueur.

Puisque aucune information n'est disponible sur la capacité des avions, ni sur la politique de gestion du revenu chez le transporteur, l'ensemble d'alternatives est réduit aux alternatives comparables. Une classe de service est considérée "non disponible" si aucune réservation n'a eu lieu dans cette classe pour le même niveau de segmentation de l'échantillon. En revanche, la valeur moyenne des attributs des alternatives non choisies est calculée à partir de ces mêmes réservations (si elles existent). Ainsi, les alternatives non choisies dont les valeurs des attributs prennent la valeur 0 sont évitées mais l'ensemble d'alternatives est plus limité qu'il ne le devrait.

[21] Un passager achète un billet aller-retour d'un transporteur qui opère le vol aller. Au retour, c'est un partenaire du transporteur initial qui opère le vol et celui-ci exige du premier une somme monétaire égale au pro-rata de la distance volée par lui par rapport à la distance totale parcourue.

RÉFÉRENCES

ABBÉ, E., BIERLAIRE, M. et TOLEDO, T. (2005). Normalisation and correlation of cross-nested logit models, Technical report, Report RO-050912.

ANDERSON, S. P., DE PALMA, A. et THISSE, J.-F. (1992). "Discrete Choice Theory of Product Differentiation", Cambridge, Mass.: MIT Press, 1992, 445 p.

BEN-AKIVA, M. E. (1973). Structure of Passenger Travel Demand Models, PhD thesis, Department of Civil Engineering, MIT, Cambridge, Mass. 1973.

BEN-AKIVA, M. E. (1974). Structure of passenger travel demand models, *Transportation Research Record*, No. 526.

BEN-AKIVA, M. E. et BIERLAIRE, M. (1999). Discrete Choice Methods and Their Applications to Short Term Travel Decisions, in Hall, R.W. (ed). *The Handbook of Transportation Science*, 1999, pp. 5-33.

BEN-AKIVA, M. E. et LERMAN, S. R. (1985). "Discrete Choice Analysis: Theory and Application to Travel Demand", MIT Press, Cambridge, Mass. 1985.

BERNDT, E. K., HALL, B. H., HALL, R. E. et HAUSMAN, J. A. (1974). Estimation and inference in nonlinear structural models, "Annals of Economic and Social Measurement", Vol.3, Issue 4, 1974, pp.653-665.

BIERLAIRE, M. (1997). "Discrete Choice Models", Intelligent Transportation Systems Prog., MIT,1997. http://roso.epfl.ch/mbi/papers/discretechoice/paper.html.

BIERLAIRE, M. (2003). BIOGEME: A free package for the estimation of discrete choice models, Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

BIERLAIRE, M., (2005). "An introduction to BIOGEME (Version 1.4)", décembre, 2005. 82 p. http://transp-or2.epfl.ch/biogeme/doc/tutorial.pdf.

BROOKE, A. S., CAVES R. E. et PITFIELD, D. E. (1994). Methodology for predicting European short-haul air transport demand from regional airports: An application to East Midlands International Airport, *Journal of Air Transport Management*, Vol. 1, Issue 1, mars 1994, pp. 37-46.

CAMBRIDGE SYSTEMATICS (1973). Series of unpublished technical memoranda on the automobile ownership project, Cambridge, Mass. 1973-74.

CHARLES RIVER ASSOCIATES (1972). A Disaggregate Behavioral Model of Urban Travel Demand, Federal Highway Administration, US Department of Transportation, Washington, DC, 1972.

COLDREN, G. et KOPPELMAN, F. S. (2005). Modeling the Competition among Air-travel Itinerary Shares: GEV Model Development, *Transportation Research A*, Vol. 39, Num. 4, pp.345-365.

COLDREN, G., KOPPELMAN, F. S., KASTURIRANGAN, K. et MUKHERJEE, A. (2003). Air Travel Itinerary Share Prediction: Logit Model Development at a Major U.S. Airline, *Journal of Air Transport Management*, Vol. 9, Num. 6, pp.361-369.

CONN, A. R., GOULD, N. I. M., et TOINT, P. (2000). "Trust region methods", MPS/SIAM Series on Optimization. SIAM.

CÔTÉ, J.-P., RISS, M., SCHOEB, A., et SAVARD G. (2006). An Application of Choice-Based Dynamic Inventory Management, AGIFORS Joint Revenue Management, Distribution and Cargo Study Group Meeting, Cancun, 2006.

CÔTÉ, J.-P. (2001). Un modèle bi-niveau pour la gestion du revenu dans le transport aérien, mémoire de maîtrise, département d'informatique et de recherche opérationnelle, Université de Montréal, avril, 2001, 94 p.

CÔTÉ, J.-P., MARCOTTE, P et SAVARD, G. (2003). A bilevel modeling approach to pricing and fare optimization in the airline industry, Journal of Revenue and Pricing Management, Vol. 2, No. 1, April, 2003, pp.23-36.

DE DONNEA, F. X. (1971). "The Determinants of Transport Mode Choice in Dutch Cities", Rotterdam: Rotterdam University Press, 1971.

GARROW, L. A. et KOPPELMAN, F. S. (2004a). Predicting air travelers show, no-show and standby behavior using passenger and directional itinerary information, *Journal of Air Transport Management*, Vol. 10, pp. 401-411.

GARROW, L. A. et KOPPELMAN, F. S. (2004b). Multinomial and nested logit models of airlines passengers no-show and standby behavior, *Journal of Revenue and Pricing Management*, Vol. 3, No. 3, pp. 237-253.

GUMBEL, E. J. (1958). "Statistics of Extremes", New York: Columbia University Press, 1958.

HESS, S. et POLAK, J. W. (2006). Cross-Nested Logit Modeling of Combined Choice of Airport, Airline, and Access Mode, Transportation Research Board Annual Meeting 2006, Paper #06-2428, 32 p.

KUNTSEVICH, A. V. et KAPPEL, F. (1997). The Solver for Local Nonlinear Optimization Problems, Institute for Mathematics, Karl-Franzens University of Graz, A-8010 Graz. http://www.kfunigraz.ac.at/imawww/kuntsevich/solvopt/content.html.

LAWRENCE, C., ZHOU, J. et TITS, A. (1997). User's guide for CFSQP version 2.5: A C code for solving (large scale). constrained nonlinear (minimax). optimization

problems, generating iterates satisfying all inequality constraints, Technical Report TR-94-16r1, Institute for Systems Research, University of Maryland, College Park, MD 20742, 1997.

LEDERMAN, M. (2004). Do Enhancements to Loyalty Programs Affect Demand? The Impact of International Frequent Flyer Partnerships on Domestic Airline Demand, Rotman School of Management, University of Toronto, June 2004, 54 p.

LUCE, R. D. (1959). "Individual Choice Behaviour", New York: John Wiley, 1959.

LUCE, R. D. et SUPPES, P. (1965). Preference, utility and subjective probability dans LUCE, R. D. BUSH, R. R. et GALANTER, E. (eds.). "Handbook of Mathematical Psychology", Vol. 3, New York: John Wiley, 1965.

MANSKI, C. (1973). Qualitative Choice Analysis, unpublished PhD thesis, Department of Economics, MIT, Cambridge, Mass., 1970.

MCFADDEN, D. (1968). The Revealed Preference of a Government Bureaucracy, Technical Report no. 17, Institute of Intl. Studies, University of California, Berkeley, 1968.

MCFADDEN, D. (1978). Modeling the choice of residential location, in Karlqvist, A. *et al.* (eds), "Spatial interaction theory and residential location", Amsterdam: North-Holland, 1978, pp. 75–96.

PEAT, MARWICK, MITCHELL et cie. (1972). Implementation of the N-dimensional Logit Model, Comprehensive Planning Organization, San Diego County, California, 1972.

PELS, E., P. NIJKAMP et P. RIETVELD (1997). Substitution and Complementarity in Aviation: Airports vs Airlines, *Transportation Research*, Vol. 33E, Issue 4, 1997, pp.275-286.

REICHMAN, S. et STOPHER, P. R. (1971). Disaggregate stochastic models of travel mode choice, *Highway Research Record*, No. 369, Highway Research Board, Washington, DC, 1971.

SPELLUCCI, P. (1993). "DONLP2 Users Guide", department of Mathematics, Technical University at Darmstadt, 64289 Darmstadt, Germany.

SPELLUCCI, P. (1998a). A new technique for inconsistent problems in the SQP method, *Mathematical Methods of Operations Research*, Vol. 47, pp.355-400.

SPELLUCCI, P. (1998b). An SQP method for general nonlinear programs using only equality constrained subproblems, *Mathematical Programming*, Vol. 82, pp.413-448.

SWAIT, J.D. et BEN-AKIVA, M. (1987). Incorporating Random Constraints in Discrete Models of Choice Set Generation, *Transportation Research B*, Vol. 21, No. 2, 1987.

THEIL, H. (1971). "Principles of Econometrics", New York: John Wiley, 1971.

WALKER, J. L. et PARKER, R. A. (2006). Estimating Utility of Time-of-Day Demand for Airline Schedules Using Mixed Logit Model, Transportation Research Board Annual Meeting 2006, Paper #06-2434, 25 p.

WATSON, P. (1974). "The Value of Time: Behavioral Models of Modal Choice", Lexington, Mass.: Heath Lexington Books, 1974.