les Cahiers Scientifiques du Transport
N° 64/2013 - Pages 91-118

Patrick Bonnel
Estimation d'un modèle agrégé de choix modal

JEL : C13, C15, C18

ESTIMATION D'UN MODÈLE AGRÉGÉ DE CHOIX MODAL

PATRICK BONNEL
LET
ENTPE, UNIVERSITÉ LUMIÈRE LYON 2, CNRS

Ces trente dernières années ont vu le développement de l'utilisation des modèles désagrégés de par le monde depuis les travaux fondateurs notamment de MCFADDEN et BEN-AKIVA (DOMENCICH, MCFADDEN, 1975 ; MCFADDEN, 2000 ; BEN-AKIVA, LERMAN, 1985) jusqu'aux avancées plus récentes des modèles mixtes multinomiaux (BHAT, 1997 ; BHAT, 2000) qui sont rendus plus accessibles grâce à certains logiciels disponibles sur le marché (ORTUZAR, WILLUMSEN, 2011). Toutefois, malgré ces avancées indéniables, l'approche désagrégée reste soumise à un certain nombre d'hypothèses qu'il n'est pas toujours aisé de tester (BONNEL, 2004). De ce fait, l'approche agrégée reste encore largement utilisée dans de nombreux pays et notamment en France où les développements de modèles désagrégés pour le choix du mode de transport reste encore peu fréquents. Jusqu'à une date récente, les travaux les plus significatifs en milieu urbain étaient peu nombreux : ABRAHAM et al., 1961 ; CETUR, 1985 ; DALY, 1985 ; CETE DE LYON et al., 1986 ; BOUYAUX, 1988 ; HIVERT et al., 1988 ; RATP, CAMBRIDGE SYSTEMATICS, 1982 ; ROUSSEAU, SAUT, 1997 ; CERTU, 1998b.

Si le développement des modèles agrégés de choix du mode de transport

reste assez simple sur le plan de la formalisation, ce n'est plus tout à fait le cas sur celui de l'estimation. Les enquêtes déplacements utilisées pour ces estimations disposent généralement d'effectifs trop faibles par rapport aux besoins des modèles agrégés. Il est de ce fait généralement nécessaire de procéder à des agrégations zonales conduisant à des incertitudes importantes sur la qualité des estimations. C'est pour faire face à cette situation que nous proposons une méthode d'estimation qui permet de conserver un zonage très fin. Nous débutons notre article par la présentation de la problématique de l'estimation des modèles agrégés (1.) pour exposer ensuite la méthode que nous avons développée (2.). Nous proposons alors une application de cette méthode pour l'estimation d'un modèle agrégé de choix du mode sur l'agglomération lyonnaise (3.). Cette analyse empirique, nous permet d'analyser l'apport de la méthode développée par rapport à des approches plus classiques (4.).

1. Problématique de l'estimation des modèles agrégés

Nous présentons tout d'abord la forme fonctionnelle logit (Ortuzar, Willumsen, 2011) qui est généralement retenue pour les modèles agrégés (1.1.) pour aborder ensuite la problématique du calage (1.2.).

1.1. Modèles logit ou régression logistique

Les fondements théoriques des modèles logit découlent de la théorie micro-économique néoclassique à partir d'une approche probabiliste de l'utilité définie dans sa seule composante déterministe. Cette forme fonctionnelle est toutefois le plus souvent justifiée par des considérations empiriques. En effet, l'analyse des données de répartition modale en fonction de la différence des temps généralisés en situation de choix bi-modal montre qu'une courbe en « S » représente très correctement les données d'enquête (Graphique 1).

Pour simplifier, selon la théorie micro-économique néoclassique, l'individu est supposé rationnel et cette rationalité s'exprime à travers le choix de l'alternative présentant la plus grande utilité (Henderson, Quandt, 1980). Dans l'approche déterministe (et agrégée) qui est la notre, l'utilité s'exprime par :

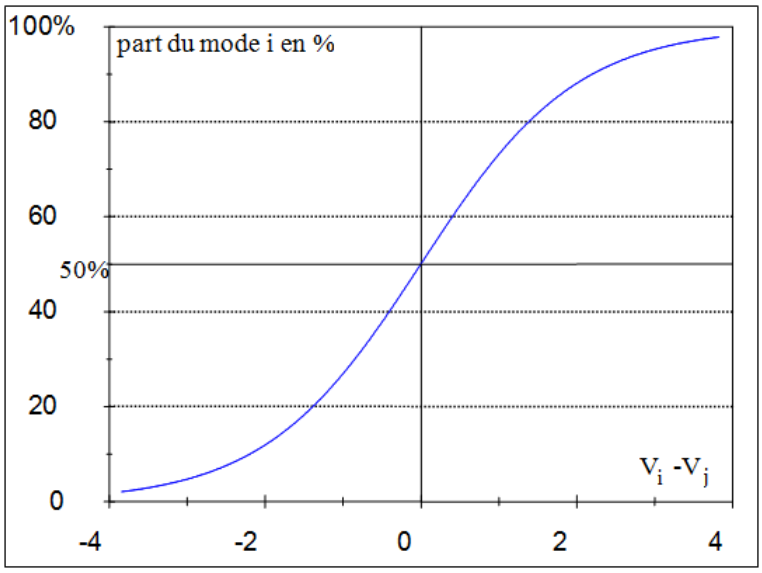
$$V_i = \sum_k \beta_{ki} X_{ki} \qquad (1)$$

où V_i est l'utilité du bien ;
X_{ki} correspond aux différentes variables explicatives permettant d'estimer l'utilité ;
β_{ki} est le coefficient de la variable X_{ki}.

Une utilisation déterministe du modèle à partir de l'utilité moyenne de la voiture et des transports collectifs conduit à un choix de la forme tout-ou-

rien (0 ou 100 %) pour chacune des origines-destinations. Ce résultat ne correspond bien évidemment pas aux données empiriques disponibles (Graphique 1). Pour y faire face, il est nécessaire de considérer une approche probabiliste. L'approche agrégée nous conduit à considérer un individu moyen se déplaçant entre une zone i et une zone j qui est confronté à un choix de mode de transport. Nous calculons ensuite pour cet individu moyen une utilité moyenne pour chacun des modes compte tenu des caractéristiques socio-économiques moyennes de cet individu moyen et de l'offre de transport de chacun des modes. Nous savons très bien que cet individu moyen est une fiction et qu'en fait les individus se distribuent autour de cet individu moyen. Nous pouvons alors utiliser l'approche probabiliste de l'utilité constante développée par LUCE et SUPPES (1965). Ces auteurs sont à l'origine de la justification théorique de l'utilisation de la forme fonctionnelle logistique dans les processus de choix discret.

Graphique 1 : Illustration de la forme générique des courbes de partage modal en S



Nous reprenons ici de manière simplifiée la présentation qu'en font BEN-AKIVA et LERMAN (1985) soit : $P\langle i|C_n\rangle$, la probabilité que l'individu n choisisse l'alternative i appartenant à l'ensemble C_n des alternatives disponibles pour l'individu n. Le modèle le plus simple s'appuyant sur une approche probabiliste de l'utilité constante a été construit, selon BEN-AKIVA et LERMAN (1985), par LUCE (1959). Il découle de l'hypothèse appelée « *choice axiom* » que pour toute alternative du sous-ensemble $\tilde{C}_n$ de l'ensemble C_n telle que $i \in \tilde{C}_n \subseteq C_n$, on a :

$$P\langle i|\tilde{C}_n \subseteq C_n\rangle = P\langle i|\tilde{C}_n\rangle \qquad (2)$$

LUCE (1959) a montré que dans le cas où cet axiome est respecté, la probabilité de choix de l'alternative i s'écrit très simplement. Dans un premier temps, nous nous limitons à un ensemble C_n ne comportant que deux alternatives i et j :

$$P\langle i|C_n\rangle = \frac{1}{1+e^{-(V_{in}-V_{jn})}} = \frac{e^{V_{in}}}{e^{V_{in}}+e^{V_{jn}}} \quad (3)$$

On retrouve l'expression de l'équation de la distribution logistique. Cette justification théorique est le plus souvent ignorée au profit de l'observation empirique des données d'enquête combinée à une commodité analytique. L'observation empirique conduit à une courbe en S proche de la distribution logistique (Cf. Graphique 1). De plus, la forme fonctionnelle logistique du fait de sa propriété IIA (indépendance vis-à-vis des autres alternatives ; BEN-AKIVA, LERMAN, 1985) conduit à une équation qui reste très simple, même lorsque le nombre d'alternatives devient très grand. Il s'agit tout simplement d'une généralisation de l'équation 3 :

$$P\langle i|C_n\rangle = \frac{e^{V_{in}}}{\sum_{j\in C_n} e^{V_{jn}}} \quad (4)$$

La propriété IIA conduit à une forme analytique très simple, ce qui constitue la force du logit par rapport à d'autres formulations. Mais cette propriété constitue aussi la faiblesse de ce modèle, car elle débouche sur le paradoxe bien connu du bus bleu et bus rouge (BEN-AKIVA, LERMAN, 1985). C'est pour faire face à ces limites que les logit emboîtés ou d'autres formes fonctionnelles plus complexes ont été développées (BHAT, 1997 ; BHAT, 2000 ; BONNEL, 2004).

Si la forme analytique est très simple, l'estimation l'est beaucoup moins en raison de la faiblesse des effectifs généralement disponibles pour estimer les matrices origines-destinations.

1.2. CALAGE DES MODÈLES LOGIT

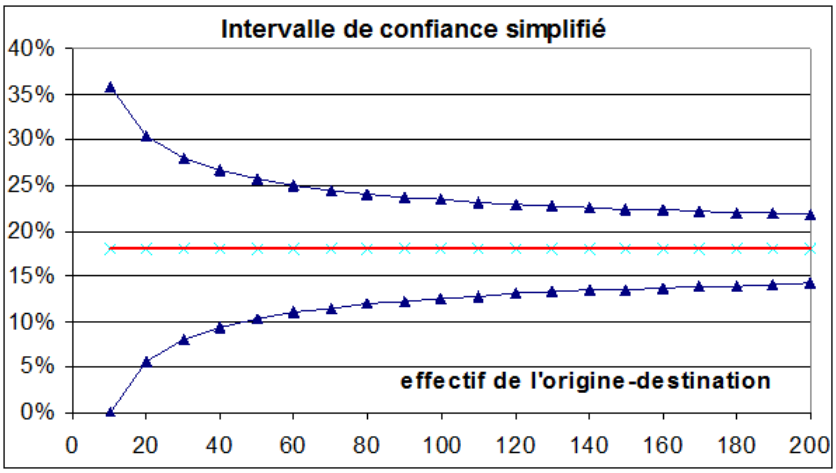
Le calage du modèle logit consiste à estimer les coefficients inconnus, à savoir les coefficients de la fonction d'utilité de chacun des modes de transport (les coefficients β_{ki} si l'on reprend les notations de l'équation 1). Cette estimation nécessite de disposer de données permettant de construire les matrices origines-destinations par mode. L'estimation vise à obtenir les coefficients β_{ki} qui fournissent la « meilleure » reproduction de ces matrices de demande par mode, en fonction des critères de convergence retenus.

La construction de ces matrices est souvent effectuée à partir des données d'enquêtes déplacements auprès des ménages qui permettent de disposer à la fois des données de mobilité et des données socio-économiques sur les

individus qui peuvent être prise en compte dans les fonctions d'utilité. La taille des échantillons de ces enquêtes est toutefois généralement trop faible pour disposer de matrices dont les données relatives à chacune des origines-destinations soient suffisamment significatives sur le plan statistique.

Pour illustrer simplement cette problématique, nous retenons l'exemple de l'agglomération lyonnaise. La transposition à d'autres agglomérations dans le monde ne posera aucune difficulté, puisque les conclusions dépendent du nombre de déplacements enquêtés pouvant être utilisés pour estimer les matrices de demande par mode. L'enquête ménages déplacements de Lyon a été réalisée en 1995 auprès de 6 000 ménages ce qui correspond à 14 000 personnes enquêtées qui ont effectué au total 53 000 déplacements. Par ailleurs, les modèles développés sur l'agglomération s'appuient le plus souvent sur un zonage comprenant entre 100 et 500 zones. Même en retenant le zonage le moins fin, la matrice contient déjà 10 000 cases. Ces 10 000 cases ne seront alimentées que par 53 000 déplacements. De plus, il faut construire autant de matrices qu'il y a de modes de déplacements pris en compte. Les matrices contiennent donc forcément de très nombreuses cellules vides et très peu de cellules dont le nombre de déplacements enquêtés sera suffisant pour que l'on dispose d'une précision statistique acceptable. Le Graphique 2 fournit l'ordre de grandeur de l'intervalle de confiance, en faisant l'hypothèse d'un tirage aléatoire simple sans refus pour une part de marché du mode de 18 %. Ces hypothèses sont évidemment plus favorables que celles rencontrées dans la réalité des enquêtes qui conduirait donc à des intervalles encore plus larges (RICHARDSON et al., 1995).

Graphique 2 : Estimation simplifiée de l'intervalle de confiance sur un pourcentage (pour une valeur de 18 %) dans le cas d'un tirage aléatoire simple



La combinaison des données d'enquêtes ménages déplacements avec d'autres sources de données, comme les comptages routiers ou les données

d'exploitant de réseaux, permet d'améliorer la qualité de l'estimation des matrices de déplacements. Toutefois, le problème de fiabilité statistique subsiste au niveau d'un grand nombre d'origines-destinations. Si la matrice de référence que l'on cherche à estimer ne présente pas une fiabilité suffisante, ce problème se reporte automatiquement sur l'estimation des coefficients des fonctions d'utilité de chacun des modes. Deux solutions sont généralement retenues pour faire face à cette difficulté.

La première consiste à ne sélectionner que les origines-destinations dont la précision statistique est jugée suffisante. L'estimation des coefficients des fonctions d'utilité est alors conduite sur une partie seulement de la matrice. Le principal problème de cette méthode réside dans l'élimination d'une partie de l'information disponible. De plus, il n'est pas certain que les origines-destinations sélectionnées soient représentatives de l'ensemble des déplacements. On peut même penser le contraire. Les origines-destinations présentant les effectifs les plus importants correspondent le plus souvent à des flux centraux ou à des flux radiaux. En revanche, les flux périphériques sont généralement beaucoup plus faibles. Le calage est donc effectué sur les flux pour lesquels la part des transports collectifs et des modes doux est la plus élevée. Il est donc probable que cette sélection conduit à réduire la variance du jeu de données initiales, ce qui ne peut que réduire la qualité de l'estimation.

La seconde conduit à une agrégation zonale de sorte à augmenter les effectifs des flux origines-destinations. Cette agrégation est généralement effectuée sur la base de la proximité géographique en essayant de regrouper des zones dont les caractéristiques ne sont pas trop différentes. Pour disposer d'effectifs suffisants, il faut réduire fortement le nombre de zones. Pour reprendre l'exemple lyonnais, même en réduisant à 25 zones (soit 625 origines-destinations), il reste un grand nombre d'origines-destinations dont les flux possèdent une précision extrêmement limitée. Le principal problème de cette agrégation concerne le calcul des données de temps et de coût pour chacune des origines-destinations. Quand la taille de la zone augmente, les temps d'accès à l'origine ou la destination deviennent difficiles à estimer pour les transports collectifs. Il y a plusieurs itinéraires pertinents pour aller de la zone origine à la zone de destination selon la localisation exacte de l'origine et de la destination au sein de chacune de ces deux zones… Il devient donc difficile de calculer une valeur pertinente pour chacune des variables explicatives. La solution généralement retenue conduit à produire une moyenne pondérée de l'ensemble des données en fonction du poids de chacune des zones les plus fines composant la macro-zone. Toutefois, compte tenu de l'étendue des zones, la pertinence de cette mesure n'est pas certaine car il y a une grande variabilité dans la durée des déplacements pour une origine-destination donnée selon la localisation exacte dans la zone origine et la zone de destination. Ainsi, si l'on a gagné en précision dans l'estimation des flux,

on a perdu en précision dans l'estimation de certaines variables explicatives entrant dans la définition de l'utilité de chacun des modes.

Le processus d'agrégation zonale préalable au calage doit donc répondre à deux objectifs contradictoires :
- le nombre de zones doit être le plus faible possible afin d'accroître le nombre de déplacements enquêtés pris en compte pour l'estimation et donc la précision des matrices origines-destinations de déplacements par mode ;
- la taille des zones doit être la plus petite possible afin de produire des données pertinentes pour les variables explicatives prises en compte dans les fonctions d'utilité.

C'est pour faire face à cette contradiction que nous proposons une autre méthode d'agrégation des données permettant de conserver la totalité de l'information et le zonage le plus fin.

2. UNE NOUVELLE MÉTHODE DE CALAGE

Cette méthode consiste à proposer un autre processus d'agrégation des origines-destinations permettant de conserver le découpage le plus fin. Les effectifs étant trop faibles pour de nombreuses origines-destinations, il faut proposer une procédure d'agrégation la plus pertinente possible. Cette méthode étant beaucoup plus simple à mettre en place pour un choix entre deux modes, nous limitons notre exposé à cette situation. Dans ce cas, la probabilité de choix du mode i peut s'écrire sous la forme suivante :

$$P\langle i|C_n\rangle = \frac{1}{1+e^{-(V_{in}-V_{jn})}} \qquad (5)$$

Le choix du mode s'effectue donc en fonction de la valeur de la différence d'utilité entre les deux modes. Il est logique d'utiliser cette quantité pour la procédure d'agrégation, compte tenu de la formulation de l'équation du logit. Le principe est simple. La méthode usuelle consiste à agréger les zones en fonction de leur proximité spatiale. La nouvelle méthode proposée ne s'appuie plus sur une agrégation zonale, mais sur une agrégation des origines-destinations. L'agrégation ne se fait plus sur la base de la proximité spatiale des origines-destinations mais en fonction de leur proximité en termes de différence d'utilité entre les deux modes. Cette procédure se justifie par le fait qu'en général la mesure de la différence d'utilité possède une précision statistique nettement supérieure à celle de la part de marché observée sur les données d'enquêtes ménages déplacements. En effet, les variables entrant dans la mesure de la différence d'utilité sont généralement soit des variables zonales (l'effectif enquêté est forcément supérieur au niveau de la zone qu'au niveau de l'origine-destination), soit des variables de niveaux de service dont la précision dépend d'une part de la codification des réseaux, et d'autre part de la finesse du zonage. Dans la pratique, il faut donc trouver le

bon équilibre entre la finesse du zonage pour les données de niveaux de service et des zones de taille suffisante pour les données zonales.

L'application de cette méthode suppose de connaître la valeur des coefficients des fonctions d'utilité. Or ces valeurs sont obtenues lors de la phase de calage du modèle. Pour lever cette contradiction, nous proposons un processus itératif dont les principales étapes sont décrites ci-dessous :

- choix de valeurs d'initialisation pour les coefficients des fonctions d'utilité des deux modes de transport afin d'initialiser le processus itératif. Ces valeurs peuvent résulter par exemple d'une précédente étude ou d'un calage effectué selon la méthode classique exposée en 1.2.. Le choix de ces valeurs n'a en fait pas beaucoup d'importance, car l'application de la méthode montre que les valeurs d'initialisation n'influent pas sur les résultats finaux dans la mesure où le nombre d'itérations est suffisamment important (Cf. 4.3.) ;
- à partir de ces premiers coefficients, on calcule l'utilité de chacun des modes, puis la différence d'utilité entre les deux modes pour chacune des origines-destinations du système zonal le plus fin pour lequel les variables explicatives sont disponibles avec un niveau de précision suffisant ;
- classement des origines-destinations selon la valeur croissante de la différence d'utilité ;
- processus d'agrégation des origines-destinations. Il s'agit d'agréger les origines-destinations en fonction de la proximité de la différence d'utilité. La méthode retenue consiste à considérer l'origine-destination avec la différence d'utilité la plus faible, puis à la regrouper avec l'origine-destination suivante (dans l'ordre croissant des différences d'utilité). Ce processus est poursuivi jusqu'à ce que le nombre de déplacements enquêtés au sein de ce regroupement soit supérieur à un seuil fixé en fonction de la précision attendue sur l'estimation des flux de déplacements. Lorsque le seuil est atteint, on procède à la construction du second regroupement et ainsi de suite jusqu'à ce que chaque origine-destination appartienne à un groupe ;
- pour chacune des nouvelles classes d'origines-destinations, il est nécessaire de calculer la valeur des variables explicatives composant les fonctions d'utilité de chacun des modes. Cette valeur est obtenue par moyenne pondérée (par le poids de l'origine-destination en nombre de déplacements) des valeurs observées au sein de chacune des origines-destinations composant la classe. Il est de même nécessaire de calculer la part de marché de chacun des modes pour cette classe. Elle est obtenue très simplement par sommation des déplacements de chacun des modes sur l'ensemble des origines-destinations composant la classe ;
- nouvelle estimation des coefficients de la fonction d'utilité.

Ce processus doit bien évidemment être renouvelé un grand nombre de fois afin d'obtenir une convergence suffisante dans l'estimation des coefficients de la fonction d'utilité. Cette méthode a été testée sur l'agglomération lyonnaise afin de vérifier d'une part sa convergence et d'autre sa fiabilité face à des données empiriques.

3. TEST EMPIRIQUE : APPLICATION SUR UN MODELE DE CHOIX MODAL DANS L'AGGLOMÉRATION LYONNAISE

Avant de présenter les résultats de l'application de cette nouvelle méthode de calage (4.), nous décrivons le modèle retenu pour ce test (3.1), les données utilisées (3.2) et la méthode d'estimation (3.3).

3.1. MODÈLE DE CHOIX MODAL POUR L'AGGLOMÉRATION LYONNAISE

Nous avons retenu un modèle développé dans une précédente étude par le LET et la SEMALY (LICHÈRE, RAUX, 1997a ; LICHÈRE, RAUX, 1997b). Ce modèle vise à estimer la répartition du marché des modes motorisés entre la voiture et les transports collectifs. La forme fonctionnelle est de type logit, même si l'expression de l'utilité n'est pas complètement additive relativement aux variables explicatives retenues :

$$\%TC = \frac{1}{1+\exp\left[k_m + \tau tc_m \cdot \left[ttc_{ij} \cdot mot_i\right] - \tau vp_m \cdot \frac{tvp_{ij}}{mot_i} - \delta_m \cdot d_j\right]} \qquad (6)$$

avec ttc_{ij}, le temps généralisé TC entre les zones i et j ;
tvp_{ij}, le temps généralisé VP entre les zones i et j ;
mot_i, la motorisation de la zone i ;
d_j, la densité de la zone j, exprimée en population + emplois à l'hectare ;
k_m, τtc_m, τvp_m et δ_m, les paramètres de répartition modale pour le motif m.

Cette formulation mérite quelques commentaires pour expliciter les composantes des fonctions d'utilité voiture et transports collectifs :

- $\tau tc_m \cdot \left[ttc_{ij} \cdot mot_i\right]$ correspond au temps généralisé transports collectifs entre les zones i et j. Il est pondéré par un facteur de perception de ce temps qui dépend de la motorisation de la zone origine. Plus la zone sera motorisée, plus ce temps sera perçu de manière défavorable (c'est-à-dire sera majoré). τtc_m est le coefficient de cette variable qui doit être déterminé lors du calage du logit ;

- $\tau vp_m \cdot \frac{tvp_{ij}}{mot_i}$ correspond au temps généralisé voiture particulière entre les zones i et j. Il est pondéré par un facteur de perception de ce temps qui dépend de la motorisation de la zone origine. Plus la zone sera motorisée,

moins ce temps sera perçu de manière défavorable (c'est-à-dire sera minoré). τ_{vp_m} est le coefficient de cette variable qui doit être déterminé lors du calage du logit ;

- $\delta_m \cdot d_j$ correspond à la densité de la zone j, exprimée en population+ emplois à l'hectare. Ce terme traduit la pression sur le stationnement. Il a été introduit faute de données pertinentes sur le stationnement sur l'agglomération lyonnaise. Plus la densité (en termes de population et d'emplois) de la zone est élevée, plus la pression sur le stationnement sera forte. L'expérience montre que cette densité est souvent une variable pertinente pour traduire la pression sur le stationnement. Dans une autre étude sur Lyon où la contrainte de stationnement pour les zones centrales a été introduite sous la forme d'une variable muette, il a été possible de retrouver des rapports de coefficients identiques au rapport des densités (Bonnel, Cabanne, 2000). En revanche, si l'introduction de cette variable est importante pour caler le modèle et ne pas biaiser la valeur des coefficients des autres variables, cette définition pose problème en simulation. Cette variable ne permet pas de simuler directement une modification de l'offre de stationnement dans une zone. Elle permettra tout au plus de simuler des variations en termes de pourcentage. δ_m est le coefficient de cette variable qui doit être déterminé lors du calage du logit.

Six motifs ont été retenus : travail, études niveau primaire, études niveau collège/lycée, études niveau universitaire, achats/services et autres motifs. Pour cette application, nous nous limitons toutefois à l'analyse du motif travail (une analyse de l'ensemble des motifs a été conduite dans Ferey, 2002).

3.2. Présentation des données

Le calage a été conduit à partir des matrices origines-destinations de déplacements pour les deux modes considérés dans cette analyse : voiture particulière et transports collectifs. Ces matrices ont été construites à partir des données de l'enquête ménages déplacements réalisée en 1995 sur l'agglomération lyonnaise (CETE de Lyon et al., 1995). Cette enquête a permis d'interroger tous les individus de 5 ans et plus de 6 000 ménages sur les déplacements de la veille du jour d'enquête (Tableau 1). La méthodologie suit un cahier des charges commun à toutes les enquêtes ménages déplacements réalisées dans les agglomérations françaises (CERTU, 1998a). L'échantillon est obtenu à partir d'un tirage aléatoire après stratification géographique sur 87 secteurs.

Les données de temps généralisés pour les transports collectifs ont été produites par le logiciel d'affectation transports collectifs de la SEMALY (SEMALY, 2000). Il s'agit d'un modèle d'affectation au plus court chemin. Cette modélisation a été réalisée à partir de la description du réseau de

transports collectifs disponibles à la date de réalisation de l'enquête ménages déplacements (FEREY, 2002).

Tableau 1 : Principales données de l'enquête ménages déplacements de Lyon 1995

	Nombre de ménages	Nombre d'individus	Nombre de déplacements	Nombre de déplacements pour le motif travail
Effectifs bruts	6 001	13 997	53 213	8 123
Effectifs redressés	536 317	1 195 189	4 659 777	728 818

source : LET d'après enquête ménages déplacements, Lyon 1995

Enfin, les données de temps généralisés pour la voiture particulière ont été produites par le CETE de Lyon à l'aide du logiciel DAVIS d'affectation pour la voiture particulière (PTV-ISIS, 2001). L'affectation a été réalisée à l'aide du modèle d'affectation à l'équilibre de Wardrop (FEREY, 2002).

L'utilisation de ces deux logiciels est nécessaire du fait de l'absence de modèle multimodal pour l'agglomération lyonnaise. L'utilisation de deux logiciels différents ne permet pas de s'assurer complètement de la cohérence des données de temps généralisés pour les deux modes de transport. Cela serait préjudiciable si notre objectif consistait à caler un modèle de prévision pour l'agglomération lyonnaise. Pour notre étude, nous nous limitons au test de la méthode de calage. L'incidence est donc marginale. Il est toutefois évident que les valeurs de calage que nous produisons ne peuvent être réutilisées pour faire de la prévision. Cette utilisation de deux logiciels pose un dernier problème dans la mesure où les découpages utilisés ne sont pas tout à fait identiques. Nous avons donc dû construire une matrice de passage entre ces deux découpages. De nouveau, cette situation est préjudiciable pour l'estimation des coefficients de calage, mais elle ne l'est pas vraiment pour le test que nous conduisons. De ce fait, nous travaillons avec le plus petit découpage commun aux découpages utilisés par la SEMALY et le CETE de Lyon, puis avec des découpages plus agrégés. Les découpages disponibles sont synthétisés ci-dessous :

- D196 : découpage en 196 zones utilisé pour le logiciel TERESE de la SEMALY ;
- D87 : découpage en 87 zones utilisé pour la stratification géographique lors du tirage de l'échantillon de l'enquête ménages déplacements ;
- D25 : découpage en 25 zones combinant une logique de couronnes et de bassins versants (LICHÈRE, RAUX, 1997a ; LICHÈRE, RAUX, 1997b) ;
- D7 : découpage en 7 zones combinant une logique de couronnes et une séparation entre l'Est et l'Ouest du Rhône (fleuve traversant l'agglomération qui exerce à la fois une coupure physique, mais aussi une différenciation sociologique).

L'analyse du calage à partir de chacun de ces découpages permettra d'identifier l'incidence de l'agrégation zonale sur l'estimation des coefficients.

3.3. Application des méthodes d'estimation

Ayant seulement deux modes de transport, l'équation du logit (équation 6) peut s'écrire d'une autre manière :

$$\log\left[\frac{1}{\%TC}-1\right]=k_m+\tau tc_m\cdot[ttc_{ij}\cdot mot_i]-\tau vp_m\cdot\frac{tvp_{ij}}{mot_i}-\delta_m\cdot d_j \qquad (7)$$

qui nous ramène à une forme linéaire, nous permettant l'utilisation de la régression linéaire pour estimer le modèle au lieu de méthodes reposant sur la maximisation de la vraisemblance (Ortuzar, Willumsen, 2011). Nous avons donc utilisé la régression linéaire pour les différentes estimations du modèle de répartition modale. Il est donc possible de caler le modèle à l'aide de n'importe quel tableur (cela nous a conduit à intégrer les deux méthodes d'estimation (présentées en 1.2 et 2.) dans un didacticiel de formation à la modélisation des transports développé par LET, IMTRANS et MVA ; Bonnel et al., 2002).

4. Résultats

Nous débutons notre présentation par l'analyse de la méthode « classique » de calage présentée en 1.2. afin d'illustrer les limites déjà évoquées (4.1.). Nous exposons ensuite les résultats de la nouvelle méthode de calage proposée (4.2.).

4.1. Méthode « classique » de calage

La première étape du calage consiste à sélectionner le zonage pertinent pour cette analyse. Pour cela, nous devons concilier deux objectifs contradictoires (Cf. 1.2.) :

- le nombre de zones doit être le plus faible possible afin d'accroître le nombre de déplacements enquêtés pris en compte pour l'estimation et donc la précision des matrices origines-destinations de déplacements par mode ;
- la taille des zones doit être la plus petite possible afin de produire des données pertinentes pour les variables explicatives prises en compte dans les fonctions d'utilité.

Nous débutons donc notre analyse avec le zonage le plus fin, soit 196 zones. Seuls 8 123 déplacements (effectifs non redressés) ont été enquêtés pour le motif travail (Cf. Tableau 1). Avec 38 416 origines-destinations, il est évident que la plupart des flux seront nuls et très peu auront un effectif suffisant pour que la précision de l'estimation de la part de marché des transports collectifs soit acceptable. Il faut descendre à un seuil de 10 déplacements enquêtés pour avoir un nombre d'origines-destinations qui ne soit pas trop faibles pour effectuer la régression (Tableau 2). Toutefois dans ce cas, d'une part seuls 9 % des déplacements sont pris en compte dans ces origines-destinations et d'autre part le Graphique 2 nous indique que pour une part de

marché de 18 % (qui est celle observée en moyenne sur l'enquête ménages déplacements de Lyon pour le motif travail), la précision est extrêmement faible (l'intervalle de confiance du pourcentage au seuil de confiance de 95 % est compris entre –0,5 % et 35 %).

Tableau 2 : Nombre d'origines-destinations (OD) dont l'effectif est supérieur à un seuil donné, découpage D196

D196, Seuil	30	25	20	15	10
Nombre d'OD dont l'effectif est supérieur au seuil	3	4	8	17	48
Poids de ces OD parmi l'ensemble des déplacements enquêtés pour le motif travail (%)	1,3	1,6	2,6	4,6	9,1

Source : LET d'après enquête ménages déplacements, Lyon 1995

Une agrégation zonale est donc indispensable. Le diagnostic est cependant identique pour le découpage en 87 zones (Tableau 3). Nous proposons donc une nouvelle agrégation zonale avec le découpage en 25 zones (Tableau 4). Même si les résultats sont moins problématiques que pour les découpages précédents, le nombre d'origines-destinations dont les effectifs sont suffisants reste assez limité. Avec un seuil de 40 déplacements enquêtés, qui conduit tout de même à un intervalle de confiance assez large ([8,6 %-25,4 %] au seuil de 95 %), seules 52 origines-destinations disposent d'un effectif enquêté suffisant. De plus, la moitié des déplacements enquêtés n'est pas inclus dans ces origines-destinations. Du point de vue de ces critères, le dernier découpage en 7 zones conduit à des résultats plus satisfaisants (Tableau 5).

Tableau 3 : Nombre d'origines-destinations dont l'effectif est supérieur à un seuil donné, découpage D87

D87, Seuil	30	25	20	15	10
Nombre d'OD dont l'effectif est supérieur au seuil	13	23	38	59	164
Poids de ces OD parmi l'ensemble des déplacements enquêtés pour le motif travail (%)	7,5	10,8	14,7	18,9	33,6

Source : LET d'après enquête ménages déplacements, Lyon 1995

Tableau 4 : Nombre d'origines-destinations dont l'effectif est supérieur à un seuil donné, découpage D25

D25, Seuil	100	80	60	40	20
Nombre d'OD dont l'effectif est supérieur au seuil	12	19	29	52	132
Poids de ces OD parmi l'ensemble des déplacements enquêtés pour le motif travail (%)	21,2	28,8	37,6	51,1	77,3

Source : LET d'après enquête ménages déplacements, Lyon 1995

Tableau 2 : Nombre d'origines-destinations dont l'effectif est supérieur à un seuil donné, découpage D7

D7, Seuil	100	80	60	40	20
Nombre d'OD dont l'effectif est supérieur au seuil	23	23	30	34	40
Poids de ces OD parmi l'ensemble des déplacements enquêtés pour le motif travail (%)	88,6	88,6	94,4	96,8	99,1

Source : LET d'après enquête ménages déplacements, Lyon 1995

Le choix du niveau de zonage et du niveau de précision attendue sur l'estimation de la matrice de part de marché de chacun des modes doit donc résulter d'un compromis entre quatre dimensions en grande partie contradictoires :
- le choix du seuil en nombre de déplacements enquêtés doit être le plus élevé possible pour garantir un niveau de précision acceptable pour la matrice de part de marché des modes ;
- le nombre de classes de la régression doit être le plus grand possible afin d'accroître le nombre de degrés de liberté de la régression et très probablement la variance initiale du jeu de données ;
- le nombre de zones du découpage doit être le plus grand possible afin de s'assurer d'une pertinence suffisante pour la mesure des variables explicatives au niveau de l'origine-destination ;
- le nombre de déplacements inclus dans les origines-destinations sélectionnées doit être le plus élevé possible afin d'utiliser au mieux l'information disponible.

L'analyse des Tableaux 2 à 5 montre qu'il est extrêmement difficile de concilier ces critères. De plus, une analyse des origines-destinations sélectionnées montre qu'il s'agit le plus souvent de flux centraux ou de flux radiaux. Même avec le découpage le plus grossier en 7 zones, il y a peu de flux périphériques inclus dans les origines-destinations sélectionnées. La variance du jeu de données ne peut que s'en trouver réduite et donc également la qualité de l'estimation.

Pour le calage, nous sélectionnons donc les configurations qui nous semblent les plus pertinentes. Cela nous conduit à éliminer les découpages en 87 et 196 zones pour ne retenir que les découpages les plus agrégés. Pour ces deux découpages, les Tableaux 6 et 7 présentent les résultats de l'estimation des coefficients. L'analyse sur le découpage en 25 zones met en évidence plusieurs problèmes :
- au-delà d'un seuil de 70 déplacements enquêtés, certains coefficients ne sont plus significatifs au seuil de 5 % (t de STUDENT < 2,1). Pour ces mêmes seuils, le nombre de classes devient également très faible conduisant à un faible nombre de degrés de liberté. Enfin le poids des déplacements représentés dans les origines-destinations conservées pour la régression est trop faible (Cf. Tableau 4) ;
- pour les seuils compris entre 40 et 60 déplacements enquêtés, la part estimée est assez proche de la part observée qui est de 17,9 %. Toutefois, on constate une instabilité assez forte, selon le seuil retenu, des coefficients estimés, à l'exception du coefficient de la variable tvp/mot. De plus, le poids des déplacements appartenant aux origines-destinations sélectionnées reste encore assez faible (entre un tiers et la moitié des déplacements sont pris en compte pour la régression).

Tableau 6 : Résultats du calage pour différents seuils en nombre de déplacements enquêtés sur le D25

Seuil	Nombre d'origines-destinations	R^2	Coefficient (t de Student entre parenthèses) Densité δ_m	tvp/mot τvp_m	ttc*mot τtc_m	Constante	Part TC estimée
40	52	75 %	-0,0044 (-4,96)	-0,038 (-3,44)	0,052 (6,64)	2,70 (8,24)	16,71
50	37	70 %	-0,0050 (-3,76)	-0,038 (-2,20)	0,045 (4,48)	2,93 (6,43)	16,89
60	29	72 %	-0,0063 (-3,99)	-0,037 (-2,14)	0,030 (2,33)	3,47 (6,25)	17,36
70	25	73 %	-0,0072 (-4,46)	-0,054 (-2,57)	0,017 (1,20)	4,29 (5,96)	22,25
80	19	84 %	-0,0066 (-4,82)	-0,042 (-2,09)	0,024 (2,09)	3,69 (5,93)	19,40
90	15	70 %	-0,0046 (-2,06)	-0,068 (-1,85)	0,027 (2,19)	4,16 (4,14)	23,14
100	12	65 %	-0,0045 (-1,40)	-0,080 (-1,60)	0,023 (1,38)	4,52 (3,24)	26,24

Source : LET d'après enquête ménages déplacements, Lyon 1995

Les résultats sont nettement plus satisfaisants dans le cas du découpage le plus agrégé en 7 zones (Tableau 7). La part transports collectifs estimée est toujours proche de celle observée dans l'enquête (17,9 %). Le R^2 est toujours élevé avec des t de STUDENT suffisants pour les coefficients estimés. Enfin, les coefficients estimés restent globalement stables quel que soit le seuil retenu. Cette stabilité découle toutefois en partie de celle relative au nombre de zones retenu pour effectuer la régression, indépendamment du seuil choisi, ce qui n'est pas le cas avec le découpage en 25 zones. Ces résultats nous conduisent à privilégier le découpage le plus synthétique. Toutefois, on peut s'interroger sur la pertinence des mesures de temps généralisé tant voiture particulière que transports collectifs sur un découpage aussi agrégé. Les valeurs moyennes calculées cachent des disparités extrêmement fortes.

4.2. MÉTHODE DE CALAGE CONSERVANT TOUTES LES DONNÉES ET UN ZONAGE FIN

Comme pour la première méthode, nous avons testé l'incidence du choix du seuil en nombre de déplacements. Toutefois, l'incidence est radicalement différente par rapport à la méthode précédente puisqu'il ne s'agit plus d'éliminer une partie de l'information disponible, mais au contraire de la conserver en regroupant les origines-destinations de telle sorte que les effectifs en nombre de déplacements enquêtés soient suffisants. De même, nous avons voulu tester l'incidence du choix du découpage. Il s'agit de savoir si l'incertitude plus forte sur la mesure des variables explicatives entrant dans les fonctions d'utilité peut avoir une incidence sur l'estimation des coefficients. Nous consignons les résultats dans le Tableau 8.

Tableau 7 : Résultats du calage pour différents seuils en nombre de déplacements enquêtés sur le D7

Seuil	Nombre d'origines-destinations	R^2	Coefficient (t de Student entre parenthèses) Densité δ_m	tvp/mot τvp_m	ttc*mot τtc_m	constante	Part TC estimée
40	34	85 %	-0,0041 (-4,79)	-0,035 (-4,64)	0,064 (9,33)	2,05 (6,62)	17,67
50	31	84 %	-0,0039 (-4,22)	-0,038 (-4,44)	0,065 (8,47)	2,16 (6,15)	17,50
60	30	84 %	-0,0041 (-4,38)	-0,038 (-4,40)	0,063 (8,02)	2,19 (6,28)	17,63
70	25	87 %	-0,0039 (-4,58)	-0,037 (-3,99)	0,065 (8,30)	2,04 (6,13)	18,51
80	23	88 %	-0,0045 (-4,73)	-0,031 (-3,04)	0,062 (7,15)	2,00 (5,86)	18,14
90	23	87 %	-0,0045 (-4,56)	-0,031 (-2,95)	0,063 (7,04)	1,97 (5,60)	18,19
100	23	87 %	-0,0044 (-4,41)	-0,031 (-2,88)	0,064 (6,93)	1,95 (5,38)	18,24

Source : LET d'après enquête ménages déplacements, Lyon 1995

Tableau 8 : Estimation des coefficients avec la méthode de calage conservant toutes les données et un zonage fin

	seuil 70		seuil 80		seuil 90		seuil 120		seuil 150		seuil 200	
D196	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type
Densité δ_m	**-0,0053**	0,0018	**-0,0054**	0,0020	**-0,0054**	0,0020	**-0,0055**	0,0024	**-0,0059**	0,0029	**-0,0060**	0,0028
Coeff. VP τvp_m	**-0,040**	0,012	**-0,039**	0,013	**-0,040**	0,014	**-0,039**	0,016	**-0,038**	0,022	**-0,030**	0,030
Coeff. TC τtc_m	**0,013**	0,009	**0,016**	0,014	**0,017**	0,015	**0,020**	0,020	**0,022**	0,023	**0,022**	0,030
constante	**3,896**	0,510	**3,793**	0,587	**3,783**	0,675	**3,659**	0,774	**3,599**	0,993	**3,256**	1,189
R^2	62,2%	4,7%	65,7%	4,8%	67,8%	5,2%	71,4%	5,7%	75,1%	6,3%	78,0%	6,8%
nombre de classes	112		99		89		66		53		40	
part TC estimée	17,0%		17,3%		17,5%		18,0%		18,5%		18,5%	
D87	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type
Densité δ_m	**-0,0049**	0,0008	**-0,0048**	0,0009	**-0,0048**	0,0012	**-0,0048**	0,0015	**-0,0051**	0,0020	**-0,0052**	0,0024
Coeff. VP τvp_m	**-0,037**	0,006	**-0,038**	0,008	**-0,040**	0,010	**-0,040**	0,013	**-0,041**	0,017	**-0,040**	0,020
Coeff. TC τtc_m	**0,015**	0,011	**0,015**	0,010	**0,017**	0,014	**0,019**	0,016	**0,020**	0,018	**0,021**	0,022
constante	**3,698**	0,540	**3,702**	0,553	**3,709**	0,712	**3,657**	0,795	**3,636**	0,944	**3,579**	1,004
R^2	59,2%	3,2%	61,2%	3,3%	63,0%	3,7%	67,1%	4,5%	71,9%	4,5%	75,1%	5,3%
nombre de classes	108		95		87		66		52		40	
part TC estimée	16,2%		16,3%		16,6%		17,1%		17,6%		18,0%	
D25	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type
Densité δ_m	**-0,0047**	0,0003	**-0,0044**	0,0003	**-0,0043**	0,0003	**-0,0042**	0,0003	**-0,0041**	0,0003	**-0,0039**	0,0004
Coeff. VP τvp_m	**-0,037**	0,004	**-0,040**	0,005	**-0,041**	0,004	**-0,045**	0,006	**-0,042**	0,006	**-0,044**	0,005
Coeff. TC τtc_m	**0,033**	0,005	**0,037**	0,005	**0,037**	0,005	**0,038**	0,005	**0,034**	0,005	**0,033**	0,004
constante	**3,125**	0,231	**3,097**	0,227	**3,128**	0,208	**3,236**	0,239	**3,185**	0,314	**3,225**	0,306
R^2	66,5%	4,3%	70,0%	3,5%	72,3%	3,9%	75,0%	3,6%	80,1%	3,5%	83,7%	2,6%
nombre de classes	83		76		68		53		44		36	
part TC estimée	16,5%		16,6%		16,8%		17,0%		17,0%		17,2%	
D7	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type	coefficient	écart-type
Densité δ_m	**-0,0041**	0,0002	**-0,0043**	0,0002	**-0,0045**	0,0002	**-0,0045**	0,0000	**-0,0042**	0,0002	**-0,0047**	0,0001
Coeff. VP τvp_m	**-0,040**	0,006	**-0,037**	0,005	**-0,033**	0,004	**-0,031**	0,000	**-0,041**	0,002	**-0,030**	0,004
Coeff. TC τtc_m	**0,059**	0,004	**0,057**	0,004	**0,056**	0,004	**0,057**	0,000	**0,061**	0,002	**0,052**	0,004
constante	**2,388**	0,299	**2,337**	0,239	**2,245**	0,294	**2,096**	0,000	**2,419**	0,035	**2,198**	0,162
R^2	85,5%	1,3%	86,1%	2,3%	86,6%	2,1%	84,9%	0,0%	82,2%	1,0%	91,8%	1,3%
nombre de classes	30		29		27		26		23		21	
part TC estimée	17,4%		17,4%		17,5%		17,9%		17,3%		17,8%	

Avant de commenter les résultats quelques précisions s'imposent sur ce tableau. L'estimation des coefficients résulte d'un processus itératif (Cf. 2.). Il est donc nécessaire d'analyser la convergence de ce processus. Les nombreux tests réalisés indiquent une forte convergence pour les premières itérations, mais après quelques centaines d'itérations, on débouche sur un résultat oscillatoire qui ne converge plus du tout (Cf. 4.3.). Afin d'éviter l'incidence

sur les résultats du nombre d'itérations, nous estimons chacun des coefficients en effectuant la moyenne sur les dernières itérations. Les données du Tableau 8 correspondent à 2 500 itérations avec un calcul du coefficient sur les 300 dernières itérations. De plus, afin d'analyser la convergence, nous indiquons dans le tableau à côté de la valeur du coefficient, le calcul de l'écart-type sur l'estimation du coefficient pour les 300 dernières itérations.

Contrairement à la méthode classique, il n'est plus possible de calculer un R^2 et des valeurs de t de STUDENT pour chacune des variables. En première approche, nous avons donc produit un R^2 moyen calculé sur les 300 dernières itérations qui sont celles qui ont été utilisées pour l'estimation du modèle. Une valeur moyenne pour les t de STUDENT n'aurait pas eu beaucoup d'intérêt. Nous avons donc vérifié qu'ils étaient suffisamment élevés pour chacune des 300 dernières itérations utilisées pour l'estimation du modèle.

Quels que soient le découpage et le seuil, la part de marché estimée des transports collectifs est proche de celle observée, même si elle est systématiquement un peu faible pour les seuils les plus faibles. Deuxième élément positif, la valeur des coefficients évolue peu en fonction du seuil pour un découpage donné. L'estimation des coefficients est donc peu sensible au choix du seuil dans la mesure où la taille de celui-ci est suffisante (pour des seuils plus faibles, la stabilité est moins grande, mais la précision de l'estimation de la part de marché également). En revanche, des différences apparaissent selon le choix du découpage. Si les résultats sont très proches pour les deux découpages les plus fins, ce n'est plus le cas avec les deux autres découpages, et tout particulièrement pour le découpage en 7 zones. On peut penser que cela est dû à l'incertitude générée par l'étendue des zones. Toutefois, les données ne nous permettent pas de conclure sur cette question. Il reste que l'objectif de cette méthode est de conserver un zonage le plus fin possible. L'utilisation de cette méthode avec un zonage grossier a donc beaucoup moins d'intérêt.

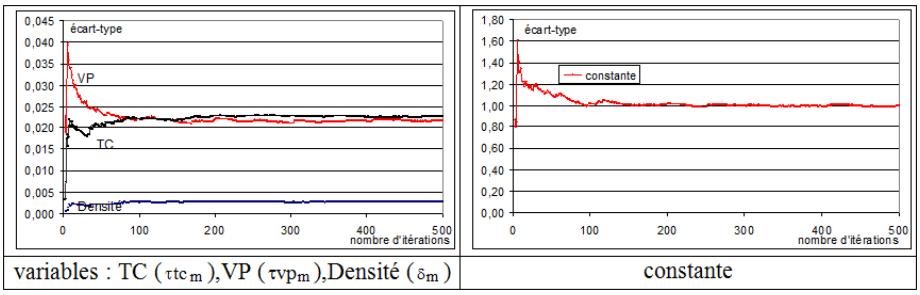
Pour le zonage en 7 zones, les résultats à l'aide de cette méthode sont très proches de ceux obtenus à l'aide de la méthode « classique » de calage, car les effectifs de nombreuses cases sont supérieurs au seuil fixé. Les classes de la régression sont donc assez proches selon les deux méthodes.

4.3. UNE CONVERGENCE RAPIDE, MAIS EN MOYENNE SEULEMENT

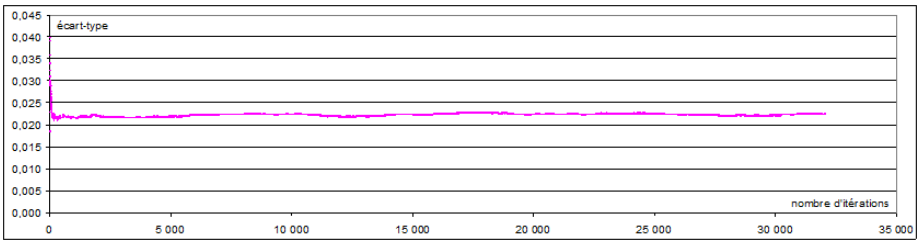
Dernier élément du diagnostic, les écarts-types sur les coefficients estimés lors du processus itératif. Ils renseignent sur la convergence de la méthode. Logiquement, ces écarts-types sont plus élevés lorsque le découpage est plus fin. Toutefois, on peut constater que les valeurs pour les découpages en 87 et 196 zones sont assez élevées au regard des valeurs des coefficients. Ce résultat est plutôt décevant dans la mesure où il désigne une convergence faible du processus itératif. Nous avons donc analysé cette convergence en effec-

tuant un grand nombre d'itérations. Il en ressort une convergence en moyenne rapide de l'estimation des coefficients. Cette convergence est d'autant plus rapide que le nombre de zones et le seuil sont plus faibles. Par ailleurs, la convergence est très peu dépendante des valeurs d'initialisation. Au-delà des premières centaines d'itérations, les résultats sont identiques en moyenne quels que soient les choix d'initialisation. Le Graphique 3 présente l'évolution de l'écart-type calculé sur l'ensemble des itérations en fonction du nombre d'itérations pour les trois variables explicatives et la constante. Ce calcul est effectué pour le découpage en 196 zones avec un seuil de 150 déplacements. Au-delà d'une centaine d'itérations, il n'y a plus que des évolutions marginales. Nous obtenons des résultats identiques pour les 3 variables et la constante. Nous avons ensuite vérifié cette stabilité pour un très grand nombre d'itérations (Graphique 4). Ce graphique fournit l'écart-type calculé sur les 5 000 dernières itérations en fonction du nombre d'itérations. Il illustre la stabilité totale de l'écart-type pour le coefficient de la variable VP (τvp_m). Des résultats identiques ont été obtenus pour les autres variables et la constante.

Graphique 3 : Evolution de l'écart-type calculé sur l'ensemble des itérations, découpage 196 zones, seuil 150 déplacements



Graphique 4 : Evolution de l'écart-type calculé sur les 5 000 dernières itérations, variable VP (τvp_m), découpage 196 zones, seuil 150 déplacements



Ces résultats nous ont conduits à rechercher une périodicité dans les résultats. Cette recherche a nécessité la réalisation d'un grand nombre d'itérations, surtout pour le découpage en 196 zones. Nous avons limité l'analyse à deux seuils de significativité de 120 et 150 déplacements pour

lesquels nous obtenons les meilleurs résultats en termes de reproduction de la part de marché observée sur l'ensemble du périmètre d'étude. Dans chacune des configurations, nous avons effectivement observé une périodicité. Les résultats sont consignés dans le Tableau 9 (le Tableau se lit de la manière suivante : pour le découpage en 196 zones et le seuil de 120 déplacements, à partir de l'itération 27 188, on observe des résultats cycliques, à savoir que l'itération 37 680 (27 188 + 10 492) fournit les mêmes résultats que l'itération 27 188 et ainsi de suite). L'identification de ce cycle est intéressante dans la mesure où elle confirme que l'on ne peut obtenir qu'une convergence en moyenne. Si l'on souhaitait une estimation très précise de cette moyenne, il serait possible d'utiliser cette périodicité pour faire le calcul de moyenne sur l'ensemble des itérations inclues dans un cycle. Toutefois, dans les cas que nous avons étudiés, la différence entre cette moyenne et celle obtenue dans le Tableau 8, c'est-à-dire avec les 300 dernières itérations d'une procédure d'estimation en comportant 2 500, les différences sont systématiquement inférieures à 1 %. Le « gain » de précision est ainsi totalement illusoire. Nous avons effectué la même analyse pour les autres motifs de déplacements. Elle débouche sur un constat tout à fait similaire à celui que nous avons pu faire pour le motif travail (FEREY, 2002).

Tableau 9 : Détermination de la périodicité des résultats de l'estimation des coefficients de la régression logistique

Découpage en 87 zones	Numéro de l'itération pour laquelle commence la périodicité des résultats	Nombre d'itérations inclues dans le cycle
Seuil 70	2 491	1 303
Seuil 80	2 219	5 448
Seuil 90	1 706	3 024
Seuil 120	8 863	5 297
Seuil 150	1 657	4 496
Seuil 200	330	5 515
Découpage en 196 zones		
Seuil 120	27 188	10 492
Seuil 150	22 983	17 187

Cette méthode répond donc aux objectifs que nous nous étions fixés à savoir une méthode permettant d'utiliser la totalité de l'information disponible et indépendante du choix du découpage ou du seuil de significativité des données. En l'état, cette méthode a toutefois une limite importante dans la mesure où elle ne permet pas de calculer d'intervalles de confiance sur les coefficients. Nous avons donc testé l'application des méthodes de bootstrap à notre contexte pour estimer des intervalles de confiance sur les coefficients et la part de marché estimée.

4.4. Estimation d'intervalles de confiance par Bootstrap

Depuis les travaux fondateurs d'EFRON (1979), les techniques de bootstrap

(EFRON, TIBSHIRANI, 1993) se sont rapidement développées pour étudier la distribution d'estimateurs pour lesquels la statistique classique ne propose pas de solution. Le principe du bootstrap consiste à partir de l'échantillon initial (le jeu de données dont on dispose), puis à procéder de manière aléatoire à de multiples ré-échantillonnages avec remise sur cet échantillon afin de produire la distribution d'estimateurs. Dans la pratique, l'analyse de la littérature fournit un nombre de ré-échantillonnage compris le plus souvent entre 300 et 1 000.

Les estimateurs pour lesquels nous souhaitons produire une distribution sont d'une part les coefficients et la constante de la fonction d'utilité et d'autre part la part de marché estimée. L'échantillon que nous devons considérer est constitué des origines-destinations sur lesquelles nous mettons en œuvre notre méthode de calage. Nous avons donc constitué par tirages aléatoires avec remise 1 000 échantillons d'origines-destinations à partir de l'échantillon initial d'origines-destinations issu de l'exploitation des données de l'enquête ménages déplacements. La difficulté de l'application à notre contexte d'étude vient de ce que la méthode de calage que nous proposons repose déjà sur un processus itératif comportant un grand nombre d'itérations. Même en nous limitant à 1 000 itérations pour la méthode de calage, l'application du bootstrap avec 1 000 ré-échantillonnage conduit à un million d'estimation de la régression logistique. Même avec les moyens de calcul actuels, l'opération reste consommatrice de temps.

Nous avons donc cherché à optimiser le nombre d'itérations nécessaires pour utiliser notre méthode de calage. Pour cela, nous avons fait le choix d'un seul découpage en retenant le découpage le plus fin en 196 zones, (soit 9 515 origines-destinations). Nous avons également limité notre investigation au seuil de 150 déplacements. L'optimisation du nombre d'itérations consiste à déterminer le nombre minimum d'itérations nécessaires pour obtenir des résultats stables quel que soit l'échantillon bootstrapé que l'on utilise. L'analyse ne pouvant être conduite sur les 1 000 échantillons du bootstrap puisque l'objectif est de minimiser le temps de calcul, nous avons effectué cette analyse sur une vingtaine d'échantillons de bootstrap. Pour ces 20 échantillons, nous avons retenu 3 000 itérations pour la méthode de calage : nous avons observé une convergence rapide à partir de l'échantillon initial (Cf. 4.3), nous espérons donc pouvoir appliquer la méthode sur un nombre d'itérations nettement plus faible. Pour ces 20 échantillons, nous avons observé une convergence tout à fait similaire avec un écart-type qui se stabilise très rapidement. Toutefois, pour compléter l'analyse, nous avons étudié la variation de l'estimation des coefficients à chacune des itérations par rapport à leur estimation sur la moyenne des itérations (en excluant toutefois les 100 premières itérations qui sont trop variables). Cette analyse permet de s'assurer d'une part de la stabilité de l'estimation et d'autre part de la minimisation des variations autour de l'estimation moyenne pour chacune des ité-

rations entrant dans le calcul de l'estimation moyenne. L'analyse montre qu'à partir de 1 500 itérations, l'écart entre l'estimation de chacune des itérations suivantes (c'est-à-dire 1 500 itérations, puisque le test a été effectué avec un nombre de 3 000 itérations) et l'estimation obtenue sur la moyenne des itérations est inférieur à 5 % pour chacune des trois variables et la constante, pour chacun des 20 échantillons bootstrap qui ont servi à ce test. Nous avons donc retenu 1 800 itérations de manière à pouvoir estimer nos coefficients sur les 300 dernières itérations pour lesquelles les variations par rapport à la moyenne sont inférieures à 5 %.

La méthode de calage a ensuite été utilisée avec 1 800 itérations sur l'ensemble des échantillons bootstrap. Le nombre d'échantillons bootstrap nécessaires au calcul des intervalles de confiance ne peut être déterminé a priori. Seule l'analyse empirique nous permettra de déterminer le nombre d'échantillons réellement nécessaires. Nous avons préparé 1 000 échantillons, l'analyse montre toutefois que 600 échantillons sont largement suffisants pour estimer les intervalles de confiance.

Deux méthodes de calcul de l'intervalle de confiance sont disponibles :
- méthode dite t-bootstrap (EFRON, TIBSHIRANI, 1993). Elle se rapproche de la méthode classique de calcul de l'intervalle de confiance en statistique inférentielle :

$$\left(E_f(B) - t_{1-\alpha/2}\sqrt{V_f(B)}; E_f(B) + t_{1-\alpha/2}\sqrt{V_f(B)}\right) \quad (8)$$

où $E_f(B)$ représente l'estimation de l'espérance de l'estimateur calculée sur les B échantillons de bootstrap ($E_f(B) = \frac{1}{B}\sum_{b=1}^{B} \theta_b^*$, avec θ_b^* la valeur de l'estimateur pour l'échantillon b) et $V_f(B)$ représente l'estimation de la variance de l'estimateur ($V_f(B) = \sum_{b=1}^{B} \frac{(\theta_b^* - E_f(B))^2}{B-1}$ avec $t_{1-\alpha/2}$ la valeur de la loi normale centrée réduite pour $1-\alpha/2$, pour un risque α) ;
- méthode du percentile bootstrap (EFRON, 1979 ; HALL, 1992) : $(a_\alpha ; b_\alpha)$ telles que $\alpha/2$ % des observations aient une valeur inférieure à a_α et $\alpha/2$ % des observations aient une valeur supérieure à b_α .

L'application de la première méthode suppose que les distributions des estimateurs suivent une loi normale, ce qui est le cas des distributions empiriques obtenues à partir des échantillons bootstrap pour chacune des variables explicatives, pour la constante et la part de marché estimée.

Nous avons préparé 1 000 échantillons bootstrap. Toutefois, il est possible qu'un plus faible nombre d'échantillons soit suffisant. Nous avons donc effectué les simulations par paquet de 100 échantillons bootstrap afin de calculer les intervalles de confiance pour chacun des effectifs multiples de 100 échantillons. Les résultats sont consignés dans le Tableau 10 pour

chacune des deux méthodes.

Tableau 10 : Coefficient estimé par la méthode du bootstrap et variation du coefficient en fonction du nombre d'échantillons bootstrap*

	Nombre d'échantillons bootstrap					
	100	200	300	400	500	600
Densité δ_m						
Coefficient	-0,00570	-0,00566	-0,00569	-0,00573	-0,00573	-0,00573
Variation t-bootstrap	15,7%	15,0%	14,8%	15,2%	15,1%	15,0%
Variation percentile-bootstrap	16,3%	14,6%	15,4%	15,4%	14,9%	14,6%
Coefficient VP τvp_m						
Coefficient	-0,0380	-0,0384	-0,0384	-0,0383	-0,0383	-0,0382
Variation t-bootstrap	35,3%	31,7%	29,9%	31,2%	30,8%	31,2%
Variation percentile-bootstrap	38,3%	30,6%	32,1%	31,4%	31,5%	31,4%
Coefficient TC τvp_m						
Coefficient	0,0174	0,0176	0,0174	0,0173	0,0172	0,0173
Variation t-bootstrap	61,3%	55,6%	53,8%	53,4%	52,9%	52,1%
Variation percentile-bootstrap	71,4%	53,5%	51,9%	52,3%	53,7%	51,7%
Constante						
Coefficient	3,736	3,738	3,749	3,756	3,757	3,753
Variation t-bootstrap	14,3%	13,0%	12,4%	12,9%	12,9%	13,2%
Variation percentile-bootstrap	14,2%	13,9%	13,5%	13,5%	13,8%	13,9%
Part de marché TC estimée						
Coefficient	17,54%	17,56%	17,56%	17,55%	17,54%	17,55%
Variation t-bootstrap	7,0%	7,0%	7,0%	7,0%	7,0%	7,0%
Variation percentile-bootstrap	7,4%	7,3%	7,3%	7,0%	7,2%	7,0%

* variation = demi-amplitude de l'intervalle de confiance/coefficient en pourcentage

Logiquement les coefficients et la part de marché TC estimée sont très proches des valeurs obtenues pour un découpage en 196 zones et un seuil de 150 déplacements. Les intervalles de confiance pour les coefficients des variables faisant intervenir les temps généralisés VP et TC sont très larges. Toutefois, ceux obtenus par la méthode classique de calage sont deux fois plus importants en moyenne quel que soit le seuil en nombre de déplacements retenus pour le découpage en 25 zones. Les intervalles de confiances bootstrap sont également inférieurs à ceux obtenus pour le découpage en 7 zones. En revanche, les intervalles de confiance pour le coefficient de la densité et pour la constante sont plus réduits. Et surtout, celui de l'estimation de la part de marché est tout à fait satisfaisant avec un intervalle de confiance [16,3 % ; 18,8 %] qui contient la valeur observée de 17,9 %. Ces résultats sont atteints à partir de 200 à 300 échantillons bootstrap, ce qui limite fortement les temps de calcul.

5. Conclusions

Les méthodes habituellement utilisées pour estimer un modèle agrégé de

choix modal doivent faire face à deux objectifs contradictoires. D'une part, elles doivent s'appuyer sur un découpage le plus fin possible pour produire des données pertinentes pour les variables explicatives prises en compte dans les fonctions d'utilité. L'estimation des temps généralisés est effectivement d'autant plus fiable et homogène au sein de l'origine-destination que les zones d'origine et de destination sont suffisamment petites. D'autre part, elles doivent s'appuyer sur des effectifs enquêtés suffisants pour chacune des origines-destinations afin que la part de marché observée sur l'origine-destination possède une précision suffisante. Concrètement compte tenu des tailles d'échantillon généralement disponibles, cela signifie une forte agrégation zonale.

Cette contradiction débouche sur un certain nombre de problèmes lors de l'estimation des coefficients :
- choix d'un seuil suffisant en nombre de déplacements enquêtés afin de garantir un niveau de précision acceptable pour la matrice de part de marché des modes. Ce choix conduit à éliminer toutes les origines-destinations dont l'effectif enquêté sera inférieur à ce seuil. Un découpage trop fin élimine donc une partie importante de l'information de base disponible (Cf. Tableaux 2 à 5) ;
- nombre de zones du découpage. Il doit être le plus grand possible afin de s'assurer d'une pertinence suffisante pour la mesure des variables explicatives au niveau de l'origine-destination ;
- nombre de classes de la régression. Il est souhaitable que le nombre de classes soit le plus grand possible afin d'accroître le nombre de degrés de liberté de la régression et très probablement la variance initiale du jeu de données. Deux actions sont possibles pour cela : réduire le seuil du nombre de déplacements enquêtés, mais on réduit aussi la précision de la matrice de part de marché observée ou bien procéder à une agrégation zonale, mais on réduit alors la précision de la mesure des variables explicatives ;
- nombre de déplacements inclus dans les origines-destinations sélectionnées pour la régression. Il est souhaitable qu'il soit le plus élevé possible afin d'utiliser au mieux l'information disponible. Mais pour cela on doit faire face à la même contradiction que ci-dessus.

L'application sur le cas de l'agglomération lyonnaise illustre les conséquences de cette contradiction (Cf. Tableaux 2 à 7) :
- le nombre d'origines-destinations pour lesquelles l'effectif enquêté est suffisant est beaucoup trop faible pour les découpages en 87 et 196 zones. Ce nombre reste faible pour le découpage en 25 zones ;
- les coefficients de calage sont fortement dépendants du choix du zonage. Ils sont également fortement dépendants du choix du seuil de fiabilité pour le nombre de déplacements enquêtés, sauf pour le zonage le plus agrégé en 7 zones.

Ces résultats nous conduisent à préconiser l'utilisation d'un découpage très

agrégé lors de l'utilisation de la méthode « classique » de calage du choix du mode. On peut toutefois s'interroger sur la fiabilité des valeurs calculées pour les variables explicatives sur un découpage en 7 zones. En particulier est-ce que le temps généralisé entre deux zones de ce découpage représente correctement la diversité des situations enquêtées ? C'est pour faire face à ce problème que nous avons proposé une nouvelle méthode de calage permettant d'une part de conserver la totalité de l'information disponible dans les enquêtes et d'autre part de travailler sur le découpage le plus fin (la finesse du zonage reste toutefois dépendante de la fiabilité des données zonales. Ces données sont cependant nettement plus fiables que les données d'origines-destinations). L'application de cette méthode permet de tirer les enseignements suivants :
- pour les deux découpages les plus fins, la valeur des coefficients est peu dépendante du choix du seuil en termes de nombre de déplacements enquêtés par classe d'origines-destinations. De plus, la part de marché estimée est proche de celle observée. En revanche, pour des zonages plus grossiers, les résultats divergent ;
- la convergence de la méthode ne s'observe qu'en moyenne. Il est donc nécessaire d'estimer les coefficients de calage en effectuant une moyenne sur les dernières itérations. En revanche, la convergence en moyenne est assez rapide. Quelques centaines d'itérations sont en général suffisantes ;
- si la convergence en moyenne est bonne, en revanche une forte variabilité subsiste entre les itérations successives.

Ainsi, sur le plan empirique, notre nouvelle méthode de calage présente les avantages majeurs suivants :
- les résultats du calage ne sont pas dépendants du choix du zonage dans la mesure où celui-ci est suffisamment fin ;
- les résultats du calage ne sont pas dépendants du choix du seuil relatif au nombre de déplacements nécessaires pour chaque origine-destination.

De plus, sur le plan théorique, les avantages sont évidents :
- possibilité de conserver un zonage fin voire très fin, ce qui permet de produire des données pour les variables explicatives beaucoup plus pertinentes ;
- la totalité de l'information disponible dans les enquêtes déplacements peut être conservée pour le calage du modèle.

La principale limite de cette méthode est de ne pas permettre le calcul direct de tests sur l'estimation des coefficients pour apprécier la qualité de l'estimation. C'est pourquoi, nous avons mis en œuvre les techniques de bootstrap pour estimer un intervalle de confiance sur les coefficients de calage ainsi que sur l'estimation de la part de marché de chacun des modes. Cette estimation nécessite toutefois des temps de calcul assez longs dans la mesure où le bootstrap nécessite un processus itératif avec quelques centaines d'échantillons et que notre méthode met également en œuvre un proces-

sus itératif. Le nombre total d'itérations est donc égal au produit des nombres d'itérations de chacun des processus. La mise en œuvre du bootstrap a permis de calculer des intervalles de confiance qui sont beaucoup plus resserrés que ceux obtenus à partir de la méthode classique. De plus, l'intervalle de confiance sur la part de marché TC estimée est très réduit et comprend la valeur observée.

Même s'il serait souhaitable de pouvoir tester cette méthode sur d'autres contextes afin de confirmer les conclusions empiriques auxquelles nous sommes parvenues, la méthode proposée répond aux objectifs que nous nous étions fixés. Il serait également intéressant d'affiner l'analyse de la convergence du processus itératif d'estimation du modèle et du nombre d'itérations nécessaires pour le bootstrap afin de réduire les temps de calcul, même si cette contrainte devient de moins en moins pressante avec le temps compte tenu de l'accroissement régulier des performances des ordinateurs.

Notre analyse a été conduite sur deux modes de transport uniquement car elle permet d'identifier très facilement une distance permettant l'agrégation des origines-destinations en utilisant la différence d'utilité des deux modes considérés. La généralisation de la méthode à un plus grand nombre d'alternatives impose la recherche d'une autre distance à partir des valeurs d'utilité de chacune des alternatives.

REMERCIEMENTS

L'auteur tient à remercier Jean-Baptiste FEREY pour le test des méthodes de calage réalisé à l'occasion de son mémoire de DEA (FEREY, 2002), ainsi que Dimitri LASTIER pour la mise en œuvre du bootstrap.

RÉFÉRENCES

ABRAHAM C., COQUAND R. (1961) La répartition du trafic entre itinéraires concurrents, réflexions sur le comportement des usagers–application au calcul des péages. **Revue Générale des Routes et Aérodromes**, n° 357, pp. 57-76.

BEN-AKIVA M., LERMAN S.R. (1985) **Discrete choice analysis, theory and application to travel demand**. Cambridge, Massachusetts, MIT Press, 390 p. (4ème éd.).

BHAT C (1997), Recent methodological advances relevant to activity and travel behavior analysis. In **Recent developments in travel behavior research**, Oxford, Pergamon.

BHAT C (2000) Flexible model structures for discrete choice analysis. In D.A. HENSHER, K.J. BUTTON (eds.) **Handbook of transport modelling**. Pergamon, Elsevier, pp. 71-90.

BONNEL P (2004) **Prévoir la demande de transport**. Paris, Presses Ponts et Chaussées, 425 p.

BONNEL P., BECHAR M., JULIEN H., ODENT P., SAINT-MARC L. (2002) **Didacticiel de formation à la modélisation transport**. Produit sous CD-Rom, Laboratoire d'Économie des Transports, ENTPE, IMTRANS, MVA pour le compte de l'ADEME, Lyon.

BONNEL P., CABANNE I. (2000) A method for breaking down and measuring the effects of correlative explanatory variables. An application to the effects of urban sprawl, car ownership and transport supply on change in the market share of public transport. **Proceedings of the European Transport Conference**, Cambridge, 11-13 September, Seminar K, pp. 205-226.

BOUYAUX P. (1988) Modélisation désagrégée des transports urbains : une application à la ville de Rennes. **Revue d'Économie Régionale et Urbaine**, n° 5, pp. 783-809.

CERTU (1998a) **L'enquête ménages déplacements « méthode standard »**. Lyon, éditions du CERTU, 295 p. (Collections du CERTU).

CERTU (1998b) **Comportements de déplacement en milieu urbain : les modèles de choix discrets, vers une approche désagrégée et multimodale**. Lyon, éditions du CERTU, 133 p. (Dossier du CERTU mobilité-Transport).

CETE DE LYON, CETE DE L'OUEST, CETUR, MELATT (1986) **Modèles désagrégés, principes généraux, méthodologie, applications (Grenoble, Nantes). Journées de rencontre sur les modèles désagrégés**. Paris, 10-11 juin, 76 p.

CETE DE LYON, INSEE, SYTRAL (1995) **Enquête « déplacements auprès des ménages de l'agglomération lyonnaise ». Document technique**. Lyon, 119 p.

CETUR (1985) **Les déplacements domicile-travail et domicile-école, modèles désagrégés de choix modal, application au cas de l'agglomération grenobloise**. Bagneux, 71 p.

DALY A. (1985) **A study of transferability of disaggregate mode choice models from Grenoble to Nantes**. Cambridge Systematics, 19 p.

DOMENCICH T.A., MCFADDEN D. (1975) **Urban travel demand: a behavioural analysis, North-Holland**. Amsterdam, Elsevier.

EFRON B (1979) Bootstrap methods: Another look at the jackknife. **The Annals of Statistics**, Vol. 7, pp. 1-26.

EFRON B, TIBSHIRANI RJ (1993) **An Introduction to the Bootstrap**. London, Chapman & Hall.

FEREY J.-B. (2002) **Analyse des méthodes d'estimation des modèles de répartition modale**. Lyon, ENTPE, Université Lumière Lyon 2, 137 p. (mémoire de DEA).

HALL P. (1992) **The Bootstrap and Edgeworth Expansion**. New-York, Springer-Verlag.

HENDERSON J.M., QUANDT R.E. (1980) **Microeconomic theory a mathematical approach**. McGraw-Hill International editions, economics series, 420 p. (3ème éd.).

HIVERT L., ORFEUIL J.-P., TROULAY P. (1988) **Modèles désagrégées de choix modal : réflexions méthodologiques autour d'une prévision de trafic**. Arcueil, INRETS, Rapport n° 67, 65 p.

LICHÈRE V., RAUX Ch. (1997a) **Développement d'un modèle stratégique de simulation des déplacements, présentation générale**. Lyon, SEMALY, LET, 28 p. + annexes.

LICHÈRE V., RAUX Ch. (1997b) **Développement d'un modèle stratégique de simulation des déplacements, guide de l'utilisateur**. Lyon, SEMALY, LET, 33 p.

LUCE R. (1959) **Individual choice behaviour: a theoretical analysis**. New York, Wiley.

LUCE R., SUPPES P. (1965) Preference, utility and subjective probability. In R. LUCE, R. BUSH, E. GALANTER (eds.) **Handbook of Mathematical psychology**. **Vol. 3**. New York, Wiley.

MCFADDEN D. (2000) Disaggregate behavioral travel demand's RUM side, a 30-year retrospective. In **pre-print 9th International Association for Travel Behaviour research Conference**, Gold Coast, Queensland, Australia, 2-7 July, Vol. 1, 38 p.

ORTUZAR J. DE D., WILLUMSEN L.G. (2011) **Modelling transport**. John Wiley & Sons (4ème éd.).

PTV-ISIS (2001) **Manuel de l'utilisateur Davisum (30/07/01), version 7.0**. Lyon, PTV-Isis.

RATP, CAMBRIDGE SYSTEMATICS (1982) **Études de politiques de transports en région Île-de-France, mise au point et utilisation de modèles désagrégés de choix modal**. Paris, 91 p.

RICHARDSON A.J., AMPT E.S., MEYBURG A.H. (1995) **Survey methods for transport planning**. University of Melbourne, Eucalyptus press, 459 p.

ROUSSEAU J., SAUT C. (1997) Un outil de simulation de politiques de transport : impact 3. **Revue Générale des Chemins de Fer**, pp. 77-83.

SEMALY (2000) **TERESE, modèle d'affectation de voyageurs dans les études de transport collectif. Document pédagogique, cours analyse et prévision de la demande de transport du DESS Transports Urbains et Régionaux de Personnes**. Lyon, ENTPE, Université Lumière Lyon 2.