# Impact of media investments on brands' market shares: a compositional data analysis approach

Joanna Morais

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *20/10/2017* **par :**

JOANNA MORAIS

---

**Impact of media investments on brands' market shares: a compositional data analysis approach**

---

**JURY**

| | | |
|---|---|---|
| RON KENETT | Professeur d'Université | Président du Jury |
| VERA PAWLOWSKY - GLAHN | Professeur d'Université | Membre du Jury |
| KAREL HRON | Professeur d'Université | Membre du Jury |
| ARMELLE GLERANT - GLIKSON | Maître de Conférences | Membre du Jury |
| PHILIPPE DEVAILLY | Dir. adjoint Revenue Management, Renault | Invité |
| HERVÉ TRANGER | Dir. Scientifique, BVA | Invité |

---

**École doctorale et spécialité :**
*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**
*Toulouse School of Economics (TSE-R)*

**Directeur(s) de Thèse :**
*Christine THOMAS-AGNAN* et *Michel SIMIONI*

**Rapporteurs :**
*Vera PAWLOWSKY-GLAHN* et *Karel HRON*

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *20/10/2017* **par :**

JOANNA MORAIS

---

**Impact of media investments on brands' market shares: a compositional data analysis approach**

---

**JURY**

| | | |
|---|---|---|
| RON KENETT | Professeur d'Université | Président du Jury |
| VERA PAWLOWSKY - GLAHN | Professeur d'Université | Membre du Jury |
| KAREL HRON | Professeur d'Université | Membre du Jury |
| ARMELLE GLERANT - GLIKSON | Maître de Conférences | Membre du Jury |
| PHILIPPE DEVAILLY | Dir. adjoint Revenue Management, Renault | Invité |
| HERVÉ TRANGER | Dir. Scientifique, BVA | Invité |

---

**École doctorale et spécialité :**
*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**
*Toulouse School of Economics (TSE-R)*

**Directeur(s) de Thèse :**
*Christine THOMAS-AGNAN* et *Michel SIMIONI*

**Rapporteurs :**
*Vera PAWLOWSKY-GLAHN* et *Karel HRON*

L'université n'entend ni approuver ni désapprouver les opinions particulières du candidat.

*A tous ces inconnus croisés par hasard,*
*qui ont aiguillé mon chemin sans le savoir.*

# Remerciements

Jamais je n'aurais cru que faire une thèse en mathématiques appliquées serait une expérience si excitante, enrichissante et épanouissante. A vrai dire je n'avais même jamais imaginé faire une thèse avant qu'on me la propose ! Ma bonne étoile... ou comme me l'a dit un jour Hervé, peut-être la chance qui sourit aux audacieux.

Je souhaite donc remercier en tout premier lieu Hervé Tranger : merci de m'avoir fait confiance et d'avoir été le tuteur d'une belle plante comme tu dis. Je parle évidemment de cette thèse, c'est avant tout grâce à toi. Un grand merci à Pascal Gaudin de m'avoir donné les moyens de réaliser cette thèse dans d'excellentes conditions et d'avoir toujours gardé un oeil bienveillant sur mon travail. Je remercie Mathieu, Thierry, Florence, Patrick et Eric d'avoir partagé avec moi leur expertise. Merci également à mes collègues de BVA qui m'ont permis de passer de très bons moments durant ces trois années.

Merci à Vera Pawlowsky-Glahn, Karel Hron, Armelle Glerant-Glikson et Ron Kenett d'avoir accepté de faire partie de mon jury de thèse. Vos travaux ont été très inspirants pour ma recherche, et soutenir ma thèse devant vous est un honneur pour moi.

Je me dois de rendre hommage comme il se doit à mes deux directeurs de thèse. Faire une thèse c'est aussi partager une vraie aventure humaine, et avec des gens comme vous ce fut un véritable plaisir, j'ai énormément appris. Christine Thomas-Agnan, être ton étudiante est une vraie fierté pour moi, tu es une personne exceptionnelle tant sur le plan professionnel que personnel. Michel Simioni, merci pour tes judicieux conseils, tes petites anecdotes qui nous ont bien fait rire et ta bonne humeur.

Je souhaite également remercier chaleureusement Marc-Henri Ambroise, Maristela Castanho, Philippe Devailly et Matthias Guegan qui m'ont fait confiance, m'ont permis d'accéder aux données tant convoitées, et m'ont abreuvée de leur connaissance du marché automobile. Huong, je suis ravie d'avoir collaboré avec toi, ce fut une vraie bouffée d'air frais dans ma thèse. Un grand merci également à la famille CODA (Marina Vives-Mestres, Juan José Egozcue, Josep Antoni Martin) de m'avoir pris sous leur aile.

Enfin je souhaite remercier mes proches. Pablo, tu m'as soutenue dans mes moments de joie intense comme dans mes moments de mauvaise humeure extrême. Simon, c'est aussi grâce à toi que cette thèse a pu se dérouler dans de bonnes conditions, tu m'as supportée au quotidien lors de mes passages à Toulouse. Merci Olivia pour ton aide en anglais. Et pour finir, je souhaite remercier la Société Française de Statistique qui, par deux fois, a mis sur mon chemin une personne qui vaut la peine de se compliquer la vie : merci Loïc pour tout le soutien et les conseils que tu m'as apportés.

**Financement :** Ce travail a été financé par la société d'études et conseil BVA et par l'ANRT dans le cadre d'une thèse CIFRE.

# Sommaire

# Introduction (version française)

Les constructeurs automobiles dépensent des montants considérables en publicité pour promouvoir leur image et pour stimuler leurs ventes. La mesure de l'impact des investissements media sur la performance des marques est donc un sujet d'intérêt pour le marché automobile. La société d'étude et conseil BVA a décidé d'aborder cette question, en finançant cette thèse CIFRE[1]. Cette question peut paraitre simple en apparence mais plusieurs particularités du marché automobile doivent être prises en compte pour y répondre.

Premièrement, il est important de tenir compte du fait que les investissements media ont peu de chance d'impacter significativement le volume du marché global. Le marché automobile est un marché de biens durables et chers, dont la taille est principalement déterminée par la demande (la taille et la richesse de la population, les habitudes en termes de mobilité et de transport, les aides gouvernementales pour l'achat de véhicules écoresponsables, etc.), plus que par l'offre. Par exemple, pendant la crise économique, les volumes de ventes ont chuté. Ainsi, les marques essaient d'avoir la plus grosse part du gâteau, dans un marché de taille donnée, en utilisant les investissements media notamment.

Deuxièmement, le marché des véhicules particuliers est généralement divisé en cinq segments principaux en Europe, nommés de A à E en fonction de la taille du châssis, et également appelés "urbaines", "citadines", "compactes", "berlines familiales", "berlines routières". Par exemple, les véhicules les plus populaires en France pour chacun de ces segments sont la Renault Twingo, la Renault Clio, la Renault Mégane, la Peugeot 508 et la Mercedes-Benz Classe E. Chacun de ces segments peut être considéré comme un ensemble de choix homogènes pour le consommateur. De plus, à l'intérieur d'un segment donné, une marque ne propose en principe qu'un véhicule phare. C'est pourquoi il est courant d'analyser le marché automobile au niveau "marque $\times$ segment".

Troisièmement, le marché automobile est un marché très compétitif où l'image des marques est capitale. Lorsque les consommateurs investissent dans un bien durable et coûteux, ils ont besoin d'être confiants en la marque qu'ils choisissent en plus d'être confiants en les caractéristiques techniques du véhicule. Cela justifie les millions d'euros dépensés par les marques pour promouvoir la qualité de leurs véhicules, mais également leur qualité de services, leurs valeurs, leur identité et leur image. Cependant, l'impact

[1]CIFRE signifie contrat industriel de formation par la recherche et désigne un type de contrat de thèse particulier en France, financé et réalisé en collaboration avec une entreprise.

publicitaire d'une marque dépend également de la publicité faite par ses concurrents. En effet, la situation dans laquelle Renault dépense 10 alors que ses concurrents dépensent 1, et la situation dans laquelle Renault dépense toujours 10 mais ses concurrents dépensent 100, n'ont pas le même impact sur Renault. De plus, certains concurrents sont plus néfastes pour une marque donnée que d'autres, et on peut imaginer que des synergies peuvent exister entre certains constructeurs.

Quatrièmement, l'impact des investissements media n'est pas seulement instantané mais il se diffuse au cours du temps, très probablement de manière décroissante. Chaque canal de communication (e.g. affichage, télévision, presse, radio) est susceptible d'avoir son propre "taux de rétention", et des interactions entre les canaux peuvent exister.

Pour prendre en compte ces éléments, nous allons considérer les hypothèses suivantes. Le marché automobile est tel que chaque marque essaye de maximiser sa part de marché, au sein d'un segment, en utilisant les investissements media comme outil. Nous allons donc favoriser la modélisation des parts de marché à celle des volumes de ventes. La concurrence et les effets croisés entre marques doivent être intégrés dans l'analyse de l'impact publicitaire. Chaque canal de communication doit être traité de façon propre, mais au sein d'une approche multivariée.

## L'approche marketing classique

Beaucoup de biens de consommation appartiennent à des marchés où la performance finale d'une marque ne dépend pas uniquement du produit proposé et des actions marketing de ladite marque, mais également des actions de ses concurrents (côté offre), et du contexte socio-économique (côté demande). Danaher et al. montrent que les effets d'interférence de la concurrence sur les ventes sont forts et diminuent l'élasticité[2] relative à la publicité. Il est donc indispensable de considérer les effets croisés de la publicité entre les marques.

L'effet des variables du mix marketing (publicité, prix, promotion, distribution) a été modélisé depuis les années 50 en utilisant les modèles dits de réponse au marché (*market response models*), où la variable de réponse est habituellement les ventes ou les parts de marché d'une marque ou d'un produit (voir Hanssens et al. [27] pour une revue des modèles existants). La mesure de ces effets se fait généralement en termes d'élasticités. Dans le cas du marché automobile par exemple, Glerant [21] mesure l'évolution de l'élasticité des ventes aux éléments du mix marketing, selon les phases du cycle de vie des véhicules. Modéliser les parts de marché plutôt que les ventes a l'avantage de prendre en considération la concurrence et les effets croisés entre marques.

Trois principales catégories de modèles de réponse au marché sont utilisées dans la pratique : les modèles linéaires, multiplicatifs et d'attraction. Les modèles d'attraction sont en général inspirés d'une version agrégée du modèle multinomial logit (MNL), très

[2]L'élasticité mesure la variation relative d'une variable (e.g. les ventes) provoquée par la variation relative d'une autre variable (e.g. la publicité). C'est une mesure de sensibilité souvent utilisée en économie et en marketing.

répandu en économétrie pour la modélisation de choix discrets (voir par exemple Train [64]). Cependant, d'après Hanssens et al. [27] (p.124), le modèle d'interaction concurrentielle multiplicative (MCI) qui est le plus connu des modèles d'attraction, est en principe préféré au MNL parce que ce dernier ne permet pas d'avoir des rendements d'échelle décroissants pour la publicité lorsque les marques ont moins de 50% de part de marché (voir Gruca et Sudharsan [25] pour une preuve mathématique). Parmi les trois catégories de modèles de réponse au marché, seuls les modèles d'attraction permettent de modéliser les parts de marché correctement parce qu'ils sont compatibles avec les contraintes de positivité et de somme unitaire des données de parts de marché, comme l'explicitent Cooper et Nakanishi [10] (p.28). Néanmoins, les modèles d'attraction ne sont pas systématiquement utilisés dans cette situation, et ce pour trois raisons.

1. La première raison est que certains auteurs ont montré empiriquement que les modèles d'attraction ne donnent pas des résultats significativement meilleurs que les autres modèles en termes de qualité d'ajustement et de prédiction (voir par exemple Ghosh et al. [20] et Leeflang et Reuyl [36]). Cependant, Naert et al. [51] ont montré l'inverse quelques années avant, suggérant que cette affirmation dépend de l'application considérée.

2. La deuxième raison est que l'estimation d'un modèle d'attraction n'est pas évidente : c'est un modèle non linéaire qui peut être linéarisé par une transformation, généralement la transformation log centrée, également appelée transformation log ratio centrée (CLR) dans la littérature de l'analyse des données de composition (voir Aitchison [1]). Une simple estimation par moindres carrés ordinaires est généralement utilisée sur les coordonnées ainsi obtenues, bien qu'il soit évident que les termes d'erreur log centrés ne peuvent être indépendamment distribués. Les moindres carrés généralisés (GLS) et les moindres carrés généralisés itératifs (IGLS) ont également été considérés par plusieurs auteurs, mais sans conclure à une amélioration significative de l'estimation (voir par exemple Ghosh et al. [20], Leeflang et Reuyl [36], et Cooper et Nakanishi [10], p.128).

3. La troisième raison est que ces modèles sont souvent surparamétrisés. Le modèle MCI classique suggère que l'impact d'un instrument marketing est le même pour toutes les marques, ce qui est souvent trop restrictif. Le modèle MCI différentiel (DMCI) incluant des paramètres spécifiques aux marques, donne lieu à la spécification de $D + KD$ paramètres, où $D$ est le nombre de marques et $K$ est le nombre de variables explicatives, mais il ignore les effets croisés entre les marques. La spécification supplémentaire de paramètres pour les effets croisés, faite dans le modèle appelé modèle MCI étendu (FEMCI), augmente le nombre de paramètres à $D(1 + DK)$. Avec l'estimation habituelle faite sur les coordonnées CLR, seules les versions centrées des paramètres du FEMCI sont identifiables.

Un autre sujet important qui a été abordé dans la littérature marketing est l'effet dynamique de la publicité. Certains auteurs ont mis en évidence l'existence d'effets de court terme et d'effets de long terme de la publicité sur les ventes (voir par exemple

Assmus et al. [2] et Lodish et al. [39]). Dans le cas de biens durables et coûteux comme l'automobile, on peut s'attendre à ce que l'impact de la publicité se propage sur plusieurs périodes, avec des rendements décroissants sur les ventes.

C'est ce que l'on appelle l'effet retard de la publicité (*advertising carryover effect*) et il est en général intégré dans les modèles de réponse de marché en utilisant une variable de stock, construite en utilisant un taux de rétention qui peut être estimé économétriquement. Dans le cas de la publicité, cette notion est appelée "adstock" et a été introduite par Broadbent [5] en 1979. Le modèle d'adstock le plus courant est le modèle de Koyck, défini comme $Q_t = \mu + \beta Adstock_t + \epsilon_t$, où la fonction d'adstock est égale à $Adstock_t = (1-\lambda)(M_t + \lambda M_{t-1} + \lambda^2 M_{t-2} + \dots)$, $Q_t$ est la quantité demandée au temps $t$, $M_t$ est l'investissement media au temps $t$, et $\lambda$ est son taux de rétention. $\beta(1-\lambda)$ peut alors être interprété comme l'effet courant (court terme) de la publicité et $\beta$ comme l'effet de report (long terme) de la publicité, et on peut dire que $\theta$% the l'impact publicitaire a lieu dans les $\log(1-\theta)/log(\lambda) - 1$ périodes après la diffusion de la publicité.

Vaktratsas et Ambler [66] déclarent que des études de type méta analyse, faites par Clarke [9] en 1976 et par Assmus, Farley, et Lehmann [2] en 1984, suggèrent que 90% des effets de la publicité se dissipent après trois à quinze mois. Ils ajoutent que dans une étude empirique, Leone [37] en 1995 suggère que cet intervalle peut être réduit à six à neuf mois. Cependant, Leone [37] met également en avant le fait que le paramètre de rétention $\lambda$ augmente lorsque le niveau d'agrégation augmente. Notons que les campagnes publicitaires sont le plus souvent analysées au niveau hebdomadaire et pour des produits de grande consommation (appelés FMCG[3] en anglais), tandis que dans notre application nous analysons les budgets publicitaires mensuels pour un bien durable. Nous pouvons alors nous attendre à obtenir de larges taux de rétention de la publicité. Habituellement, les effets publicitaires au cours du temps sont estimés sur les ventes et pour un seul instrument marketing (la publicité dans la plupart des cas), et non sur les parts de marché avec une publicité multicanale. Récemment, Zantedeschi et al. [72] ont modélisé l'impact des variables d'adstock en multicanal sur des volumes de ventes. Pour autant que nous sachions, il n'existe pas d'application considérant un modèle d'attraction incluant des effets de report de la publicité, en multicanal.

## L'approche par analyse des données de composition

Les données de parts de marché sont avant tout des données de parts, caractérisées par les contraintes suivantes : elles sont positives et somment à 1. Par définition, ce sont des "données de composition" : une composition est un vecteur de parts d'un certain ensemble qui porte une information relative. Pour une composition de $D$ parts de marché, si $D-1$ parts sont connues, la $D^{\text{ème}}$ part est simplement égale à 1 moins la somme des $D-1$ autres parts. Une D-composition appartient à un espace appelé le simplexe $\mathcal{S}^D$. Le simplexe peut être considéré comme une généralisation du triangle : une 2-composition

[3]Fast-Moving Consumer Goods.

peut être représentée sur un segment, une 3-composition peut être représentée dans un triangle, une 4-composition peut être représentée dans un tétraèdre, etc. A cause de ces contraintes, les modèles de régression classiques ne peuvent pas être directement utilisés pour la modélisation de données compositionnelles.

Beaucoup de champs d'application sont concernés par l'analyse des données de parts. En économie politique, Elff [15] étudie les comportements de vote et analyse les relations entre les parts de votes des partis politiques et leurs positions politiques dans différents groupes de votants. En géologie, Solano-Acosta et Dutta [62] se sont intéressés à la composition lithologique du grès en termes de quartz, de feldspath et de fragments rocheux. Pour les aménagements environnementaux, l'utilisation des sols est modélisée pour connaître la proportion des différents types d'usages (forêt, agriculture, zone urbaine, etc.) sur une parcelle donnée (voir Chakir et al. [7]).

Deux types de modèles statistiques sont adaptés au cas où la variable dépendante est une composition : le modèle de Dirichlet et le modèle de régression compositionnel.

Bien que connu dans la littérature marketing, le modèle de Dirichlet (DIR) est peu utilisé pour modéliser des parts de marché. Hanssens et al. [27] (p.128) déclare que la distribution de Dirichlet, qui est la distribution d'une composition obtenue par la clôture de variables indépendamment distribuées selon des lois Gamma, semble être adaptée aux cas où il n'y a pas d'erreur d'échantillonnage dans les données. Ce cas est peut-être rare en marketing, mais c'est le cas de notre application comme nous le verrons plus tard.

Au contraire, les modèles compositionnels provenant de l'analyse des données de composition et dénotés modèles CODA, semblent méconnus dans la littérature marketing. L'inverse est également vrai : l'analyse des données de composition, qui est un domaine de recherche récent en statistique dont le premier champ d'application a été la géologie, n'a pas été appliquée au marketing avant cette thèse.

L'analyse des données de composition a été initiée par Aitchison [1] qui a notamment développé la géométrie du simplexe, également appelée géométrie d'Aitchison depuis 2001 par Pawlowsky-Glahn et Egozcue [55]. Elle est basée sur une approche de transformation de type log ratio : une composition qui appartient à l'espace du simplexe est transformée en coordonnées en utilisant une transformation log ratio, de manière à pouvoir utiliser n'importe quelle méthode statistique classique (la régression linéaire par exemple) sur les coordonnées ainsi obtenues. Les résultats peuvent ensuite être retrouvés dans le simplexe par transformation inverse.

Les modèles de régression compositionnels ont été récemment explorés d'un point de vue théorique, dans les ouvrages suivants : Pawlowsky-Glahn et Buccianti [54] en 2011, Van Den Boogaart et Tolosana-Delgado [68] en 2013, et Pawlowsky-Glahn, Egozcue et Tolosana-Delgado [56] en 2015. Mais très peu d'articles les utilisent en pratique : Hron et al. [30] présentent un cas où les variables explicatives sont compositionnelles, et Egozcue et al. [13] se concentrent sur le cas où la variable dépendante est une composition. Pour autant que nous sachions, le cas où une variable dépendante compositionnelle

est expliquée par des variables explicatives prenant des valeurs différentes pour chaque composant (ou par des variables compositionnelles) comme dans notre situation[4], a uniquement été abordé dans deux articles récents : Wang et al. [70] en 2013, et Chen et al. [8] en 2016. Cependant, le premier article présente un modèle simplifié comparé au deuxième article, qui n'a pas été abordé dans les ouvrages précédemment cités. Notons que Kynclova et al. [35] ont aussi modélisé une composition par des compositions mais dans le cas particulier d'un modèle autorégressif.

Les principaux obstacles à l'utilisation des modèles compositionnels sont les suivants. Les notations des opérateurs dans la géométrie du simplexe peuvent sembler fastidieuses et sont inhabituelles. En effet, l'opération d'addition dans l'espace euclidien classique est "remplacée" par une opération de perturbation dénotée $\oplus$, et la multiplication est "remplacée" par l'opération puissance dénotée $\odot$. De plus, les modèles compositionnels sont difficilement interprétables puisqu'ils sont ajustés dans un espace transformé, et peu de recherche a été faite pour explorer ce sujet. Ils sont habituellement interprétés en termes d'effets marginaux sur les parts transformées, ce qui est compliqué à utiliser en pratique. Cependant, l'avantage des modèles compositionnels est qu'ils permettent d'introduire facilement des effets croisés entre les composants et qu'ils prennent en compte de façon rigoureuse la nature des données de parts.

## Notre contribution

Cette thèse traite la question suivante : "Quel est l'impact des investissements media sur les parts de marché des marques du marché automobile ?". Dans cette optique, nous allons combiner les atouts de l'approche marketing et de l'approche compositionnelle, de manière à élaborer un nouveau modèle d'attraction performant capable d'expliquer les parts de marché des marques d'un segment donné du marché, en fonction des investissements publicitaires multicanaux, et tenant compte des effets retard de la publicité et des effets croisés entre les marques. Le modèle final sera cohérent avec la géométrie dans le simplexe.

La première étape est de comparer les différents modèles en fonction de leurs propriétés. Nous montrons qu'ils peuvent être tous écrits sous une forme similaire, sous forme d'attraction, ce qui facilite leur comparaison, et est particulièrement appréciable pour le modèle CODA puisque cela permet de se débarrasser des notations complexes de la géométrie du simplexe. Nous expliquons pourquoi les modèles CODA et Dirichlet peuvent être plus performants que les modèles de parts de marché habituels. Nous prouvons également que le modèle MCI peut être considéré comme un cas particulier du modèle CODA, et nous mettons en avant les similarités entre le modèle CODA et le "modèle MCI étendu" (*fully extended MCI model*, un modèle MCI incluant des effets croisés entre les marques) utilisé en marketing.

[4]Les variables explicatives seront les investissements media qui prennent des valeurs différentes pour chaque marque.

Puis, nous prouvons que le modèle MCI et le modèle CODA peuvent être combiné et nous développons une procédure de sélection de modèles pour déterminer quelle spécification choisir. Nous proposons plusieurs types d'interprétations pour le modèle CODA, directement liés aux parts et inspirés de ceux utilisés en marketing. Les élasticités sont utiles pour isoler l'impact d'une variable explicative sur une part donnée, puisqu'elles correspondent à la variation relative d'un composant induite par la variation relative d'une variable explicative, *ceteris paribus* (au sens du simplexe). Nous prouvons que cette mesure est cohérente avec la géométrie du simplexe. Nous expliquons aussi que l'estimation des modèles d'attraction peut être améliorée en utilisant une transformation développée en analyse des données de composition.

Le modèle final que nous proposons inclut les effets retard de chaque canal de communication (affichage, presse, radio, télévision). Nous expliquons comment déterminer les taux de rétentions correspondants d'une manière multivariée, et comment interpréter ce modèle de manière à en tirer des enseignements pratiques sur les stratégies de mix marketing des constructeurs automobiles.

Pour finir, nous commençons à adapter le théorème de Dorfamn et Steiner sur l'optimisation du budget publicitaire au cas concurrentiel (attraction) et multicanal.

## Données

Les données nécessaires pour réaliser ce travail de recherche ont été mises à disposition par la direction marketing de Renault dans le cadre d'un accord de confidentialité entre Renault et BVA. Trois bases de données ont été fournies :

- la base mensuelle des immatriculations contenant les informations suivantes : le segment, la marque, le modèle, la version, le mois, l'année, et le nombre de véhicules immatriculés correspondant, pour tous les véhicules particuliers neufs immatriculés en France de janvier 2000 à août 2015,
- la base mensuelle des prix catalogue contenant les informations suivantes : la marque, le modèle, la version, les options, le mois, l'année, le nombre de véhicules immatriculés et le prix catalogue correspondants, pour tous les véhicules particuliers neufs immatriculés en France de janvier 2000 à août 2015,
- la base mensuelle des investissements media contenant les informations suivantes : la marque, le modèle, le type de media (canal), le mois, l'année, et le montant investi correspondant, pour tous les véhicules particuliers neufs ayant fait l'objet d'investissements publicitaires en France de janvier 2000 à août 2015.

Un travail minutieux de rapprochement des bases a été effectué, en collaboration avec l'équipe marketing de Renault (Philippe Devailly et Matthias Guegan). La base de référence est la base des immatriculations, contenant l'information du segment qui est une information importante dans le cadre de notre recherche. La base des prix a été agrégées au niveau "modèle $\times$ marque $\times$ segment $\times$ date" en pondérant les prix des différentes versions par les volumes d'immatriculations correspondants. Nous n'avons

conservé de la base des investissements media que ceux associés à un modèle de véhicule en particulier, les dépenses institutionnelles associées à une marque ou un groupe de marques sont omises.

Les principaux types de difficultés rencontrés lors de ce rapprochement de bases sont les suivants : les modèles ou les marques ne sont pas homogènes dans les différentes bases ; les modèles ou les marques ont changé de nom au cours du temps ; les modèles ont changé de segment au cours du temps ; les modèles étant la suite directe d'un modèle précédent doivent être considérés comme une seule lignée de véhicules (e.g. Peugeot 206, Peugeot 207, Peugeot 208) ; certains prix sont exprimés en francs et d'autres en euros sur la même période, etc.

Nous obtenons à la fin de cette étape une base mensuelle au niveau "modèle $\times$ marque $\times$ segment $\times$ date" contenant le volume d'immatriculations, le prix catalogue moyen et les dépenses publicitaires en affichage, presse, radio, télévision, cinéma et internet. Notre objectif n'étant pas de mesurer l'impact des campagnes publicitaires sur les ventes de chaque modèle de véhicules à un niveau micro, mais plutôt d'apprécier l'impact structurel des investissements media des marques sur leur niveau de part de marché à un niveau macro, nous agrégeons cette base au niveau "marque $\times$ segment $\times$ date". Cette agrégation a des avantages : elle nous permet d'être à un niveau où la confidentialité des données est moindre, et d'être sur une base de données avec beaucoup moins de "trous" (e.g. apparition ou disparition d'un modèle de véhicule, mois sans vente pour un modèle donné, etc.) ce qui sera appréciable pour la modélisation.

Par ailleurs, notre objectif étant de mesurer l'impact de la publicité sur les ventes relatives, il est important de tenir compte du fait que nous observons les immatriculations du mois $t$ et non les ventes du mois $t$. Il peut s'écouler plusieurs semaines voire plusieurs mois entre l'acte d'achat qui nous intéresse et l'immatriculation d'un véhicule, en grande partie à cause du délai de livraison du véhicule neuf. C'est pourquoi dans le Chapitre 1 nous avons considéré les investissements media du mois $t-4$ et dans le Chapitre 2 ceux des mois $t-3, t-4, t-5$. Dans le Chapitre 3, pour le modèle final, nous avons proposé un calage des investissements media sur les immatriculations en utilisant la distribution empirique des délais de livraison (cf. Section 3.2.3).

## Structure de la thèse

Cette thèse est écrite en anglais, à l'exception de la présente introduction, de la conclusion et des résumés long et court qui sont traduits en français. Les chapitres 1, 2 et 3 sont des adaptations d'articles en cours de publication, dont les coauteurs sont les professeurs Christine Thomas-Agnan and Michel Simioni, directeurs de cette thèse.

Le premier chapitre de cette thèse présente les modèles de parts de marché usuels provenant de la littérature marketing et économétrique, et d'autres modèles de régression, issus de l'analyse des données de composition principalement, permettant plus généralement de modéliser des données de parts, en tenant compte des particularités

de ces données qui sont positives et somment à 1 à chaque observation. Les quatres grands types de modèles considérés sont le modèle MNL, le modèle GMCI, le modèle DIR et le modèle que nous nommons CODA. Une comparaison théorique approfondie des modèles met en évidence leurs points communs et leurs différences, ainsi que leurs avantages et leurs inconvénients. Une application empirique est présentée pour le segment B du marché automobile français pour la sous-composition Dacia/Nissan/Renault, de juin 2005 (apparition de Dacia) à août 2015, en utilisant le prix catalogue moyen et les dépenses publicitaires en affichage, presse, radio, télévision, cinéma et internet du mois $t-4$, ainsi que la prime à la casse en variables explicatives. Les modèles sont comparés en terme de qualité d'ajustement en validation croisée de manière à identifier le modèle donnant les meilleures estimations des parts de marché.

Le deuxième chapitre se concentre sur l'interprétation des modèles de parts de marché, en particulier du modèle MCI et du modèle CODA, qui est complexifiée par le fait qu'une part de marché ne peut évoluer sans que les autres soient impactées. Plusieurs types d'interprétation sont proposés, dont les effets marginaux, les élasticités et les rapports de cotes. Les élasticités se révèlent être particulièrement intéressantes car elles sont directement liées aux dérivées dans le simplexe et offrent une interprétation des effets directs et croisés entre les marques qui est pertinente d'un point de vue pratique. Une application au segment B du marché automobile de 2003 à août 2015 est présentée, en considérant les trois leaders français Citroën, Peugeot et Renault séparément et en agrégeant les autres marques. Les variables explicatives utilisées sont cette fois-ci le total des dépenses publicitaires des mois $t-3, t-4, t-5$, le prix catalogue moyen et la prime à la casse. L'application met en avant d'intéressantes interactions entre les marques.

Le troisième chapitre présente le modèle final retenu pour répondre à notre problématique de départ, et les interprétations concrètes de celui-ci. Nous proposons un modèle de type CODA intégrant les principaux canaux de communication (affichage, presse, radio et télévision) séparément et tenant compte de l'effet de report de l'impact publicitaire à travers des variables dites d'adstock. On explique comment déterminer le taux de rétention de la publicité des différents canaux de façon simultanée. Le modèle final est choisi à l'issue d'une comparaison de différentes spécifications de modèles de types MCI, DIR ou CODA, sur des critères de qualités explicative et prédictive. Un diagnostic des résidus et des ellipsoïdes de confiance et de prédiction sont réalisés. L'application finale porte, comme dans le Chapitre 2 sur les marques Citroën, Peugeot, Renault, et le groupe des autres marques, mais sur la période de 2005 à août 2015. Les variables explicatives sont les adstocks des dépenses publicitaires par canal, et à nouveau le prix catalogue moyen et la prime à la casse. Les élasticités directes et croisées des différents canaux sont très différentes d'une marque et d'un media à l'autre, et donnent des informations sur de possibles stratégies de mix marketing pour les acteurs du marché.

Le quatrième et dernier chapitre de cette thèse ouvre sur les pistes à explorer pour améliorer le modèle proposé et apporter des réponses complémentaires à notre problé-

matique. Nous évoquons notamment la possibilité de réaliser des modélisations supplémentaires sur les autres segments et autres marques, ainsi que des modélisations complémentaires pour analyser par exemple l'impact du contexte socio-économique sur les parts de marché des différents segments du marché automobile, ou encore l'impact de la composition des dépenses publicitaires (mix marketing) sur les ventes d'un certain véhicule. La construction de variables d'adstock reflétant au mieux l'effet retard de la publicité peut probablement être améliorée. Pour finir, nous avons commencé à examiner l'utilisation des élasticités des parts de marché aux dépenses publicitaires dans l'optimisation du budget publicitaire en multicanal.

# Introduction (English version)

Car manufacturers spend a lot of money on advertising in order to enhance their image and to stimulate their sales. The measure of the impact of the media investments on the brand performance is therefore a subject of major interest for the automobile market. The consulting and research company BVA has decided to investigate this question, funding this CIFRE[5] thesis. This question may seem simple in appearance but some features of the automobile market need to be considered to answer it.

Firstly, it is important to understand that media investments have little chance to significantly impact the global market size. The automobile market is a market of expensive and durable goods, whose size is mainly determined by the demand (the size and the wealth of the population, people's transportation behavior and mobility habits, governmental incentives to purchase a new environmentally-friendly car, etc.), more than by the supply. For example, during an economic crisis, the sales volumes usually fall. Then, brands try to get the largest share of a given market, using media investments notably.

Secondly, the market of personal-use vehicles is usually divided into five main segments in Europe, denoted by A to E according to the size of the chassis, also called minicompact, subcompact, compact, mid-size and large segments in the USA. For example the most popular vehicles in France for each segment are the Renault Twingo, the Renault Clio, the Renault Megane, the Peugeot 508 and the Mercedes-Benz E-Class. Each of these segments can be considered as an homogeneous set of choices for customers. Moreover, inside a given segment, a brand usually proposes only one main vehicle. Then, it is usual to analyze the automobile market at the "brand $\times$ segment" level.

Thirdly, the automobile market is a very competitive market where the brand image is very important. For the purchase of an expensive and durable good, customers need to trust the brand they choose in addition to having confidence in the technical characteristics of the vehicle. This explains why car manufacturers spend millions of euros on outdoor advertising displays, television, radio, press and so on, in order to promote the quality of their vehicles, but also to enhance their quality of service, values, identity and image. However, the advertising impact of a brand depends also on the advertising of its competitors. Indeed, the situation where Renault spends 10 while its competitors spend

[5]CIFRE means industrial training contract by research, and corresponds to a particular French type of thesis contract, supported by a company.

1, and the situation where Renault still spends 10 but its competitors spend 100, will not have the same impact on Renault. Moreover, some competitors are more harmful than others, and we can imagine that synergies exist between some brands.

Fourthly, the impact of media investments is not only contemporaneous but spreads out over time, certainly in a decreasing way. Each advertising channel (e.g. outdoor, television, press, radio) is likely to have its own "retention rate", and interactions between channels may exist.

To take into account these elements, we consider the following assumptions. The automobile market is such that each brand tries to maximize its market share, inside a given segment, using media investments. Thus, we are going to promote a brands' market shares modeling instead of a classical sales volume modeling. Competition and cross effects between brands must be included in the advertising impact analysis. Each communication channel should be addressed individually but with a multivariate approach.

## The classical marketing approach

A lot of consumer goods belong to competitive markets where the final performance of a brand does not only depend on the supplied product and marketing actions of the brand but also on competitors actions (supply side), and on the socioeconomic context (demand side). Danaher et al. [11] show that the competitive interference effects on sales are strong and diminish the advertising elasticity[6]. It is therefore essential to consider advertising cross effects between brands.

The effect of marketing mix variables (advertising, price, promotion, distribution) have been modeled since the 50's using the so-called market response models, where the response variable is usually the sales or the market shares of products or brands (see Hanssens et al. [27] for a review of existing models). The measurement of these effects is generally done in terms of elasticities. In the case of the automobile market for example, Glerant [21] measures the evolution of the elasticity of sales relative to the elements of the marketing mix, according to the phases of a vehicle life cycle. Modeling market shares instead of sales has the advantage of taking into account the competition and the cross effects between brands.

Three main categories of market response models are used in practice: linear, multiplicative and attraction models. Attraction models are generally inspired from an aggregated version of the multinomial logit model (MNL), widely used in econometrics for discrete choice modeling (see for example Train [64]). However, according to Hanssens et al. ([27], p.124), the multiplicative competitive interaction (MCI) model, which is the most famous attraction model, is usually preferred to the MNL model because the latter does not allow "*for decreasing returns to scale for advertising [...] for any brand with less than a 50 percent share of the market*", as proved in Gruca and Sudharsan

[6]The elasticity measures the relative variation of a variable (e.g. sales) implied by the relative variation of another variable (e.g. advertising). It is a sensitivity measure often used in economics and marketing.

[25]. Among the three categories of market response models, only the attraction models allow to model market shares in a proper way because they comply with the constraints of positivity and summing up to one of market shares data, as emphasized by Cooper and Nakanishi [10] (p.28). Nevertheless, attraction models are not used systematically for market shares modeling, because of three main reasons.

1. The first one is that some authors have shown empirically that attraction models do not give significantly better results than the others in terms of fitting and prediction accuracy (see for example Ghosh et al. [20] and Leeflang et al. [36]). Nevertheless, Naert et al. [51] have made the opposite claim a few years ago, suggesting that the conclusion can depend on the considered application.

2. The second reason is that the estimation of an attraction model is not straightforward: it is a non-linear model which can be linearized by a transformation, generally the log-centering transformation, also called centered log-ratio transformation (CLR) in the compositional data analysis literature (see Aitchison [1]). A simple estimation by ordinary least squares is generally run on the resulting coordinates, while it is obvious that the log-centered error terms cannot be independently distributed. Generalized least squares (GLS) and iterative generalized least squares (IGLS) have also been considered by several authors, but without concluding to a significant improvement of the estimation (see for example Ghosh et al. [20], Leeflang and Reuyl [36], and Cooper and Nakanishi [10], p.128).

3. The third reason is that they are often overparametrized. The classical MCI model suggests that the impact of a marketing instrument is the same for all brands, which is often too restrictive. The differential MCI model (DMCI) includes brand specific parameters, leading to $D + DK$ parameters, where $D$ is the number of brands and $K$ the number of explanatory variables, but it ignores the potential cross effects between brands. The additional specification of cross effects, done in the so-called fully extended MCI model (FEMCI), leads to a huge number of parameters: $D(1 + DK)$. With the estimation on the CLR transformed model, only the centered version of these coefficients can be identified.

Another important concern which has been addressed in the marketing literature is the dynamic effect of the advertising. Some authors have emphasized the existence of short term and long term effect of advertising on sales (see for example Assmus et al. [2] and Lodish et al. [39]). In the case of durable and expensive goods like automobile, we can expect the advertising impact to be spread over several periods, with diminishing returns effect on sales.

This is called the carryover effect of advertising and it is usually integrated in market response models using a stock variable, built using a retention rate which can be estimated econometrically. For advertising, this notion is also called "adstock" variable and was initiated by Broadbent [5] in 1979. The most commonly used adstock model is the Koyck model, defined as $Q_t = \mu + \beta Adstock_t + \epsilon_t$ where the adstock function is equal to $Adstock_t = (1 - \lambda)(M_t + \lambda M_{t-1} + \lambda^2 M_{t-2} + \dots)$, $Q_t$ is the demand at time $t$,

$M_t$ is the media investment at time $t$, and $\lambda$ is its retention rate. Then, $\beta(1-\lambda)$ can be interpreted as the current (short term) effect of advertising and $\beta$ as the carryover (long term) effect of advertising, and we can say that $\theta\%$ of the advertising impact occurs in the $\log(1-\theta)/\log(\lambda)-1$ periods after advertising.

Vakratsas and Ambler [66] report that "*Clarke [9] (1976) and Assmus, Farley, and Lehmann [2] (1984), in meta-analytic studies, suggest that 90% of the advertising effects dissipate after three to fifteen months. Leone [37] (1995), in an empirical generalizations study, suggests that the range be narrowed to six to nine months*". However, Leone [37] also emphasizes the fact that the retention parameter $\lambda$ "*should increase as the level of aggregation increases*". Note that advertising campaigns are most often analyzed at the week level and for FMCGs (fast moving consumer goods), whereas in our application we are analysing the monthly advertising budgets for a durable good. Then we can expect to find larger carryover effects of advertising. Usually advertising carryover effects are estimated on sales and for only one marketing instrument (advertising in most of the cases), not on market share for multi-channel advertising. Recently, Zantedeschi et al. [72] have modeled the impact of multi-channel adstock variables on sales. As far as we know, there is no existing application of an attraction model including a multi-channel advertising carryover effects.

## The compositional data analysis approach

Above all, market shares are shares data characterized by the following constraints: they are positive and sum up to 1. By definition shares are "compositional data": a composition is a vector of shares of some whole which carries relative information. For a composition of $D$ market shares, if $D-1$ market shares are known the $D^{\text{th}}$ market shares is simply 1 minus the sum of the $D-1$ other parts. A $D$-composition lies in a space called the simplex $\mathcal{S}^D$. The simplex can be considered as a generalization of the triangle: a 2-composition can be represented in a segment, a 3-composition in a triangle, a 4-composition in a tetrahedral, and so on. Because of the relative information carried by compositions, classical regression models cannot be used directly to model compositional data.

A large number of fields are concerned by the analysis of share data. In political economy, Elff [15] studies voting behaviors and analyzes the relationship between the shares of political parties and their policy positions in different groups of voters. In geology, Solana-Acosta and Dutta [62] are interested in the lithologic composition of sandstone according to whether it is quartz, feldspar or rock fragments. For environmental planning purposes, land use models focus on the proportions of different types of uses (forest, agriculture, urban, etc.) on a given piece of land (see for example Chakir et al. [7]).

Two types of statistical models are adapted to the case where the dependent variable is a composition: the Dirichlet covariate model and the compositional regression model.

Although known in the marketing literature, the Dirichlet model (DIR) is rarely used

to model market shares. Hanssens et al. [27] (p.128) argue that the Dirichlet distribution, which is the distribution of a composition obtained as the closure of independent Gamma-distributed variables, seems to be adapted to the case of no sampling error in the data. This case may be rare in this type of application, but it is actually our case.

By contrast, compositional models coming from the compositional data analysis and thus denoted CODA models, seem to be totally ignored by the marketing literature. The inverse is also true: the compositional data analysis, which is a quite recent field in statistics where the initial application area was geology, has not been applied to marketing before this thesis.

Aitchison [1] can be considered as the father of compositional data analysis, having developed the simplicial geometry, also called Aitchison geometry since 2001 by Pawlowsky-Glahn and Egozcue [55]. Compositional data analysis is based on a log-ratio transformation approach: a composition lying in the simplex is transformed in coordinates using a log-ratio transformation, such that any classical statistical method (linear regression for example) can be used on coordinates, and then the results in the simplex can be recovered by inverse transformation.

The compositional regression models have been recently investigated from a theoretical perspective, in the following books: Pawlowsky-Glahn and Buccianti [54] in 2011, Van Den Boogaart and Tolosana-Delgado [68] in 2013, and Pawlowsky-Glahn, Egozcue and Tolosana-Delgado [56] in 2015. But very few articles are applying them in practice: Hron et al. [30] present a case where the explanatory variables are compositional, and Egozcue et al. [13] focus on the case where the dependent variable is a composition. To the best of our knowledge, the case where a compositional dependent variable is explained by component-dependent (or compositional) explanatory variables as in our situation[7], has only been addressed in two recent articles: Wang et al. [70] in 2013, and Chen et al. [8] in 2016. However, the former article presents a simplified model compared to the later, which has not been mentioned in the books we cite above. Note that Kynclova et al. [35] are also modeling a composition by compositions but in the particular case of an autoregressive model.

The main obstacles to use compositional regression models are the following. The notations of operators in the simplicial geometry can be confusing and cumbersome as they are unusual. Indeed, the addition operation in the classical Euclidean space is "replaced" by a perturbation operation denoted $\oplus$, and the multiplication is replaced by a powering operation denoted $\odot$. Moreover, the resulting compositional models are complicated to interpret as they are fitted in a transformed space, and little research has been carried out investigating this issue. They are usually interpreted in terms of marginal effects on the transformed shares, which are complicated to use in practice. However, the positive point of the compositional models is that they can easily introduce all cross effects between components and that they take into account rigorously the

[7]Explanatory variables will be media investments which have different values for each component of the dependent market shares composition, that is for each brand.

compositional nature of share data.

## Our contribution

This thesis addresses the following question: "What is the impact of media investments on the brands' market shares in the automobile market?". In order to do so, we are going to combine the best part of the marketing approach and of the compositional approach, in order to build a new performing attraction model able to explain brands' market shares of a given segment of the market, as a function of multi-channel advertising investments, accounting for the advertising carryover effect and the cross effects between brands. This final model will be consistent with the simplicial geometry.

The first step is to compare the different models according to their properties. We show that they can all be written in a similar formulation, the attraction formulation, which eases the comparison and is particularly valuable for the CODA model because it allows to get rid of the cumbersome notations of simplicial operations. We explain why the CODA and the Dirichlet models can outperform traditional market share models. We also prove that the MCI model can be considered as a particular case of the CODA model.

Then we prove that the MCI and the CODA models can be mixed and we develop a model selection procedure to determine which specification should be chosen. We propose several types of interpretations for the CODA model directly linked to the shares and inspired from those used in marketing. Elasticities are useful to isolate the impact of an explanatory variable on a particular share as they correspond to the relative variation of a component with respect to the relative variation of an explanatory variable, *ceteris paribus* (in a simplex sense). We prove that this measure is consistent with the simplicial geometry. We also explain how the estimation of attraction models can be improved using a transformation developed in the compositional data analysis.

The final model we propose includes the carryover effect of each advertising channel (outdoor, press, radio, television). We explain how to determine the corresponding retention rates in a multivariate way, and how to interpret this model in order to get practical findings on marketing mix strategies for automobile manufacturers.

Finally, we start to adapt the Dorfman-Steiner theorem about the advertising budgeting optimization to the multi-channel attraction case.

## Data

The data required for this research work have been provided by the marketing direction of Renault under a confidentiality agreement between Renault and BVA. Three databases have been used:

- the monthly registrations database containing the following information: the segment, the brand, the model, the version, the month, the year, and the number

of corresponding registered vehicles, for all new personal-use vehicles registered in France from January 2000 to August 2015,

- the monthly catalogue prices database containing the following information: the brand, the model, the version, the options, the month, the year, and the number of corresponding registered vehicles and their catalogue price, for all new personal-use vehicles registered in France from January 2000 to August 2015,

- the monthly media investments database containing the following information: the brand, the model, the media channel, the month, the year, and the corresponding advertising expenses, for all new personal-use vehicles which are subject to media investments in France from January 2000 to August 2015.

The merger of the three databases has been carefully done, in collaboration with the Renault marketing team (Philippe Devailly and Matthias Guegan). The reference database is the registrations database, including the segment level information, which is important in our case. The prices database has been aggregated at the "model $\times$ brand $\times$ segment $\times$ date" level, weighting the different versions' prices by the corresponding registration volumes. We only consider the media investments which are associated to a particular vehicle, not the institutional expenses linked to a brand or a group of brands.

The main difficulties we have met for this merger are the following: the vehicle models or the brands are not homogenous in the different data sources, the models or the brands have been renamed across time, the model's segment has changed across time, the models being the direct follow-up of one previous vehicle have to be considered as a unique vehicle line (e.g. Peugeot 206, Peugeot 207, Peugeot 208), some prices are in francs and others in euro over the same period, etc.

We obtain at the end of this merger a monthly database at the "model $\times$ brand $\times$ segment $\times$ date" level, containing the registrations volumes, the average catalogue price and the advertising expenses in outdoor, press, radio, television, cinema and internet. As our aim is not to measure the advertising campaigns' impact on sales for each vehicle model at a micro level, but rather to assess the structural impact of brands' media investments on their market shares at a macro level, we aggregate this database at the "brand $\times$ segment $\times$ date" level. This aggregation has several advantages: it allows us to deal with data with a lower level of confidentiality, and to get a database with less "holes" (e.g. creation or disappearance of a vehicle model, month without sales for a particular vehicle, etc.) which will be valuable for modeling.

Furthermore, our aim being to measure the impact of advertising on relative sales, it is important to take into account that we observe registrations at month $t$, not sales at month $t$. Several weeks, or several months can separate the purchase act and the registration, mainly because of delivery times. That is why in Chapter 1, we consider media investments at $t-4$ and in Chapter 2 those of $t-3, t-4, t-5$. In Chapter 3, for the final model, we propose to align media investments on registrations using the empirical delivery times distribution (see Section 3.2.3).

## Structure of the thesis

This thesis is written in English, with the exception of this introduction, the conclusion and long and short resumes which are translated into French. Chapters 1, 2 and 3 are adaptations of working papers under publication, whose coauthors are professors Christine Thomas-Agnan and Michel Simioni, advisors of this thesis.

The first chapter of this thesis presents the usual market share models coming from the marketing and econometric literatures, and other regression models mainly coming from compositional data analysis, which allow more generally to model share data, taking into account the specificity of these positive and summing up to one data. The four types of models we considered are the MNL model, the GMCI model, the DIR model and what we call the CODA model. An in-depth theoretical comparison of these models highlights their common points and their differences, along with their strengths and their weaknesses. An empirical application is presented for the B segment of the French automobile market for the subcomposition Dacia/Nissan/Renault, from June 2005 to August 2015, using the average catalogue price, the media investments in outdoor, press, radio, television, cinema and internet at time $t-4$ and the scrapping incentive as explanatory variables. The models are compared in terms of cross-validated quality measures in order to identify the model giving the best fitted market shares.

The second chapter focuses on the interpretation of market share models, in particular for the MCI and the CODA models, which is complexified by the fact that a market share cannot evolve without affecting the others. Several types of interpretations are proposed: marginal effects, elasticities and odds ratios. Elasticities come out to be particularly interesting because they are directly linked to simplicial derivatives and because they provide an interpretation for direct and cross effects between brands, which is meaningful from a practical point of view. An application to the B segment of the French automobile market from January 2003 to August 2015 is displayed, considering separately the three French leaders Citroën, Peugeot and Renault, and aggregating the other brands. The explanatory variables we have used are the total media investments at time $t-3, t-4, t-5$, the average catalogue price and the scrapping incentive. The application notably reveals interesting positive interactions between brands.

The third chapter presents the final model selected to answer our initial question, and its practical interpretations. We propose a CODA type model integrating separately the different communication channels (outdoor, press, radio, television) and accounting for the carryover effect of the advertising impact through the so-called adstock functions. We explain how to determine the channels' advertising retention rate in a multivariate way. The final model is chosen according to a comparison of different model specifications (MCI, DIR or CODA types) using goodness-of-fit and prediction accuracy criteria. A residuals diagnostic is done, along with confidence and prediction ellipsoids. The final application concerns the brands Citroën, Peugeot, Renault, and the group of other

brands, as in Chapter 2, but on the period from January 2005 to August 2015. The explanatory variables are the adstock variables by channel, and again the average catalogue price and the scrapping incentive. The direct and cross elasticities of advertising channels differ a lot from one brand to another, and from one channel to another, and they give information on the possible mix marketing strategies for the market players.

The fourth and final chapter of this thesis opens a discussion on avenues to be explored in order to improve the proposed model and to bring complementary answers to our initial problem. We talk about the possibility to make complementary modelings on other segments and on other brands, or to build a model to measure the impact of the socioeconomic context on segments' market shares, or even to measure the impact of the advertising budget composition (marketing mix) on vehicle sales. Moreover, the construction of adstock variables that best reflect the carryover effect of the advertising could probably be improved. Models distributional assumptions could also be challenged. Lastly, we made preliminary attempts at investigating the use of advertising elasticities of market shares in the multi-channel advertising budget optimization.

# Communications and publications

## Conferences (oral presentations and posters)

- Market share models: an application to the automobile market. 6$^{\text{ème}}$ Rencontres des jeunes statisticiens (RJS2015), August 2015, Le Teich, FRANCE
- Modeling market shares or compositional data: an application to the automobile market. 48$^{\text{ème}}$ Journées de statistique (JDS2016), May 2016, Montpellier, FRANCE
- Impact of media investments on new car sales in a competitive context: a market-shares approach. 38$^{\text{th}}$ INFORMS Society for Marketing Science conference (ISMS2016), June 2016, Shanghai, CHINA
- Modeling market-shares or compositional data: An application to the automobile market (Poster). 31$^{\text{st}}$ International Workshop on Statistical Modelling (IWSM2016), July 2016, Rennes, FRANCE
- Interpreting the impact of explanatory variables in compositional models. 7$^{\text{th}}$ International Workshop on Compositional Data Analysis (CODAWORK2017), June 2017, Abbadia San Salvadore, ITALY
- Impact assessment of media investments on automobile brand market shares using compositional models. 1$^{\text{st}}$ Econometrics and statistics conference (ECOSTA2017), June 2017, Hong Kong, CHINA

## Prizes

- Best general oral presentation prize, on behalf of the CODAWORK2017 Scientific Committee chaired by Karel Hron and Raimon Tolosana-Delgado.

## Seminars (invitations for oral presentation)

- Modeling market shares or compositional data: An application to the automobile market. 16$^{\text{th}}$ February 2016, Seminari d'Estadística de Girona, Dept. d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, SPAIN
- Impact of media investments on new car sales in a competitive context: a market share approach. 17$^{\text{th}}$ October 2016, DART, Bentley University, USA
- Using elasticities to interpret compositional models. 26$^{\text{th}}$ April 2017, Seminari d'Estadística de Girona, Dept. d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, SPAIN

## Publications

- Extended version of Chapter 1: Morais, J., Thomas-Agnan, C., and Simioni, M. A tour of regression models for explaining shares. *TSE Working Paper*, 16-742 (2016).
- Short version of Chapter 1: (accepted on the 27th of September 2017); corresponding working paper on HAL: Morais, J., Thomas-Agnan, C., and Simioni, M. Using compositional and Dirichlet models for market-share regression. Working paper, July 2017 (will remain available after publication).
- Short version of Chapter 2 (submitted to a special issue of Austrian Journal of Statistics): Morais, J., Thomas-Agnan, C., and Simioni, M. Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper 17-805* (2017) (will be deleted after publication), and Morais, J., Thomas-Agnan, C., and Simioni, M. Interpretation of explanatory variables impacts in compositional regression models. Working paper, July 2017 (will remain available after publication).
- Trinh, H. T., and Morais, J. Impact of socioeconomic factors on nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis. *TSE Working Papers*, 17-825 (2017). Submission in progress.
- Morais, J., and Thomas-Agnan, C. Impact of economic conditions on automobile market segment shares: a compositional approach. *CSBIGS* (2018). In progress.

# Chapter 1

# A tour of regression models for explaining shares

The aim of the first chapter is to make a presentation and a comparison of the different models we found in the literature for modeling market shares. Some of these models, found in the marketing and in the econometrics literature, are directly intended to model market shares, whereas other regression models are simply adapted to fit a compositional response variable, that is a vector of shares data, for whatever application. As a result of this comparison, we highlight the links between existing models which have been developed contemporaneously in different literatures and for different applications. This chapter will be useful in the following chapters as it emphasizes the best aspect of each model, the theoretical side or the application side, that we will combine to built a performing market share model able to answer our question.

This chapter is linked with working papers which have already been published (see Morais, Thomas-Agnan and Simioni [46], [50] and [49]). A final version will be published soon in Journal of Applied Statistics (accepted on the 27th of September 2017).

## 1.1 Introduction

Share data are characterized by the following constraints: they are positive and sum up to 1. By definition shares are "compositional data": a composition is a vector of parts of some whole which carries relative information. For a composition of $D$ parts, if $D-1$ parts are known the $D^{\text{th}}$ part is simply 1 minus the sum of the $D-1$ other parts: $D$-compositions lie in a space called the simplex $\mathcal{S}^D$. Because of these constraints, classical regression models cannot be used directly.

A large number of fields are concerned by the analysis of share data. In political economy, Elff [15] studies voting behaviors and analyzes the relationship between the shares of political parties and their policy positions in different groups of voters. In geology, Solana-Acosta and Dutta [62] are interested in the lithologic composition of sandstone according to whether it is quartz, feldspar or rock fragments. For environmental planning purposes, land use models focus on what are the proportions of different types of uses (forest, agriculture, urban, etc.) on a given piece of land, see for example Chakir et al. [7].

When the aim is to model market shares as a function of explanatory variables (marketing factors like advertising or price for example), the marketing literature proposes some regression models which can be qualified as attraction models (Cooper and Nakanishi [10]). They are generally inspired from an aggregated version of the multinomial logit models, widely used in econometrics for discrete choice modeling. But aggregated multinomial logit models (MNL) and market share models (GMCI) are very simple and present some limitations: for example they are not including cross effects between brands.

In this chapter, we also present two other models adapted to model market shares: the Dirichlet covariate model (DIR) and the compositional model (CODA). These models consider the vector of shares as a "composition" lying in the simplex. DIR allows to estimate brand-specific parameters and CODA allows to estimate additionally cross effect parameters. We show that these last two models can be written in a similar fashion, called attraction form, as the MNL and the GMCI models. This is particularly valuable for the CODA model because it allows to get rid of the notations of simplicial operations. We compare the main properties of the models in order to explain why CODA and DIR models can outperform traditional market share models. We also prove that the GMCI model can be considered as a particular case of the CODA model, and we highlight the similarities between the CODA model and the fully extended MCI model used in marketing, which is a MCI model including cross effects.

Finally, an application to the automobile market is presented where we model brands' market shares as a function of media investments in six channels (television, press, radio, outdoor, digital, cinema), controlling for the brands' average price and a scrapping incentive dummy variable. We compare the goodness of fit of the various models by cross validation in terms of quality measures adapted to share data. The direct elasticities of market shares relative to the television advertising are computed for each model.

The present chapter is organized as follows: the models adapted to model share data are presented in Section 1.2, and theoretically compared in Section 1.3. Section 1.4 presents an application to an automobile market data set, along with an empirical comparison of the models in terms of cross-validated goodness-of-fit measures, and an example of elasticity interpretation. Finally, the last section concludes on the findings and on further directions to be investigated.

## 1.2 Models for explaining shares

### 1.2.1 Notations

The notations used in this thesis are standardized in Table 1.1 depending on whether the variables are considered in volume or in share, as either dependent or explanatory variable, and if they are component and/or observation dependent.

For example, in the case we use for illustration, we are interested to model the market shares $S_{jt}$ of $D$ brands, computed using the closure (see the definition below) of the corresponding sales volumes $\check{S}_{jt}$, for $j = 1, \ldots, D$. The media investments $M_{cjt}$ and the prices $P_{jt}$ are component and observation dependent because they have a different value for each brand at each time, while the scrapping incentive is only an observation dependent variable (the same for all brands). Advertising and price variables can be considered either in volume (in euro) or in shares (shares-of-voice and relative prices).

Table 1.1: Notations

| **Variable** | **Volumes** | **Shares** | **Coordinates** |
|---|---|---|---|
| Dependent | $\check{S}_{jt}$ | $\mathbf{S}_t = (S_{1t}, \ldots, S_{Dt})' = \mathcal{C}(\check{S}_{1t}, \ldots, \check{S}_{Dt})'$ | $ilr(\mathbf{S}) = \mathbf{S}_t^* = \check{\mathbf{S}}_t^*$ |
| Explanatory (observation and component characteristic) | $\check{X}_{jt}$ | $\mathbf{X}_t = (X_{1t}, \ldots, X_{Dt})' = \mathcal{C}(\check{X}_{1t}, \ldots, \check{X}_{Dt})'$ | $ilr(\mathbf{X}) = \mathbf{X}_t^* = \check{\mathbf{X}}_t^*$ |
| Explanatory (observation characteristic only) | $Z_t$ | | |
| **General notations** | | | |
| $D$ | Number of components (3 in this application) | | |
| $j, l, m = 1, \ldots, D$ | Index of components or coordinates (brands in this application) | | |
| $T$ | Number of observations (123 in this application) | | |
| $t = 1, \ldots, T$ | Index of observations (time in this application) | | |
| $K, K_X, K_Z$ | Number of explanatory variables / of type $X$ / of type $Z$ | | |
| $k = 1, \ldots, K$ | Index of explanatory variables (by default) | | |
| $k = 1, \ldots, K_X$ | Index of explanatory variables of type $X$ | | |
| $\kappa = 1, \ldots, K_Z$ | Index of explanatory variables of type $Z$ | | |
| $s_j$ | Theoretical mean share (expected value of $S_j$) | | |
| **Notations for the application** | | | |
| $C$ | Number of media channels (6 in this application) | | |
| $c = 1, \ldots, C$ | Index of media channels | | |
| $M_{cjt}$ | Media investment in channel $c$ at time $t$ for brand $j$ | | |
| $P_{jt}$ | Average price at time $t$ of brand $j$ | | |
| $SI_t$ | Scrapping incentive dummy at time $t$ | | |

A composition $\mathbf{S}$ is a vector of $D$ shares $S_j$ potentially coming from the closure of $D$ positive numbers $\check{S}_j$ and belongs to the simplex $\mathcal{S}^D$:

$$\mathcal{S}^D = \left\{ \mathbf{S} = (S_1, \ldots, S_D)' = \mathcal{C}(\check{S}_1, \ldots, \check{S}_D)' : S_j > 0,\ j = 1, \ldots, D;\ \sum_{j=1}^{D} S_j = \tau \right\}$$

where the closure operation $\mathcal{C}(y_1, \ldots, y_D)' = \left( \frac{\tau y_1}{\sum_{j=1}^{D} y_j}, \ldots, \frac{\tau y_D}{\sum_{j=1}^{D} y_j} \right)'$ normalizes any vector $\mathbf{y}$ to a constant sum $\tau$. In the case of market shares, $\tau = 1$.

### 1.2.2 Multinomial logit models

In econometrics, multinomial logit (MNL) models are widely used to model discrete choices of individuals, i.e. to model the probability that an individual $i$ chooses an alternative $j$, using individual data. Sometimes these data are aggregated using a group variable (e.g. time, space, age group) and then the counts for each alternative and the covariates are recorded for each group. We are going to describe how an individual-level MNL model can be adapted to aggregated data, provided that the explanatory variables are either describing the alternatives (and are constant for all decision makers in a group) or are group characteristics.

#### Discrete choice model: a random utility model for individual data

Multinomial logit models (MNL) are widely known by statisticians because they are a generalization of the famous binary logistic regression model. MNL is a particular case of discrete choice models, used to explain and predict polytomous, discrete or qualitative, response variable (a finite set of mutually exclusive and collectively exhaustive alternatives) by a set of explanatory variables (see Koppelman and Bhat [33]).

In econometrics, random utility models are based on the idea that decision makers are choosing the alternative that maximizes their utility. For an introduction to utility in econometrics, see for example McFadden [43]. Thus, the probability for decision maker $i$ to choose alternative $j$ at choice situation $t$ is defined as

$$p_{ijt} = \mathbb{P}(Choice_{it} = j) = \mathbb{P}[U_{ijt} \geq U_{ilt}, \quad \forall\, l \neq j], \tag{1.1}$$

where $Choice_{it}$ is the variable of choice of individual $i$ at choice situation $t$, and $U_{ijt}$ is the utility associated to alternative $j$ for decision maker $i$ at choice situation $t$.

Random utility models decompose the utility $U_{ijt}$ as a sum of a deterministic part $V_{ijt}$ and a random part $\epsilon_{ijt}$:

$$U_{ijt} = V_{ijt}(X_t) + \epsilon_{ijt},$$

where X is a set of explanatory variables for the deterministic part of the utility.

If error terms are extreme-value (Gumbel) distributed, the computations of probabilities from equation (1.1) have a closed form leading to the multinomial logit model, also called random coefficient logit model (see Koppelman and Bhat [33]):

$$p_{ijt} = \frac{\exp(U_{ijt})}{\sum_{l=1}^{D} \exp(U_{ilt})},$$

which can be estimated by maximum likelihood using the density of the multinomial distribution on individual-level data.

#### Conditional logit model: alternative-specific explanatory variables

If explanatory variables only characterize alternatives (and not individuals), MNL is called "conditional logit model". If alternative characteristics do not change across decision makers, the conditional logit model can be expressed in an aggregated way, using

count data instead of individual data, which means that only the numbers of individuals who have chosen each alternative are needed instead of the individual choices. This is the case for our illustration data: it allows us to estimate the market share of a brand (probability to be chosen) depending on the characteristics of this brand relatively to the characteristics of other brands in competition.

The expected share of alternative $j$ at choice situation $t$ (e.g. market share of brand $j$ at time $t$) corresponds actually to the probability of $j$ to be chosen by an individual, and is expressed as

$$s_{jt} = \mathbb{E}(S_{jt}|\check{\mathbf{X}}_t) = \frac{\exp(a_j + \sum_{k=1}^{K} b_k \check{X}_{kjt})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K} b_k \check{X}_{klt})}, \tag{1.2}$$

with $a_D = 0$ for identifiability reasons.

**Estimation by maximum likelihood for aggregated data**

The multinomial distribution is a generalization of the binomial distribution. For $\check{S}$ independent individuals who choose exactly one of $D$ alternatives, with each alternative having a given probability to be chosen $(s_1, \ldots, s_D)$, the multinomial distribution gives as a result a vector containing the volumes of choices for each alternative $(\check{S}_1, \ldots, \check{S}_D)$, where $\check{S} = \sum_{j=1}^{D} \check{S}_j$. We denote $(\check{S}_1, \ldots, \check{S}_D) \sim \mathcal{MN}(\check{S}; s_1, \ldots, s_D)$, such that:

$$\mathbb{E}(\check{S}_j) = \check{S}s_j, \quad Var(\check{S}_j) = \check{S}s_j(1 - s_j), \quad Cov(\check{S}_j, \check{S}_l) = -\check{S}s_j s_l$$

If the explanatory variables characterizing the alternatives do not change across individuals (for example the price of the vehicle $j$ is the same for all individuals $i$), then the utility for alternative $j$, and thus the probability to choose the alternative $j$, will be the same for all individuals $i$. Therefore, the log likelihood is only a function of the counts $\check{S}_{jt}$ of individuals for each alternative.

In the aggregated case, it is needed to observe several choice situations $t$ in order to estimate the model, that is to have a group variable. In our illustrative application, the different choice situations are the months of observation. The corresponding log-likelihood function (up to a constant) which has to be maximized is

$$\log L = \sum_{t=1}^{T} \sum_{j=1}^{D} \check{S}_{jt} \log(s_{jt}) = \left[ \sum_{t=1}^{T} \sum_{j=1}^{D} \check{S}_{jt} (\check{X}_{jt} b) \right] - \left[ \sum_{t=1}^{T} \check{S}_t \log \left( \sum_{j=1}^{D} \exp(\check{X}_{jt} b) \right) \right],$$

with $\check{X}_{jt} b = \sum_{k=1}^{K} b_k \check{X}_{kjt}$.

**Implementation in R:** the package **mclogit** developped by Martin Elff [16] allows to fit conditional logit models with count data, using the Fisher-scoring/IWLS algorithm[1].

[1] For details on IWLS algorithm, see for example Green [22].

### 1.2.3 Market share models

Market share models were developed in the 80's, mainly by Cooper and Nakanishi [10]. To take into account the competition between brands in a market, it is often of interest to model market shares instead of sales volumes directly. Thus, this type of models is widely used in marketing. The aim is to model market shares of $D$ brands using their marketing instruments (e.g. price, advertising) as explanatory variables, with aggregated data (market-level data rather than individual-level data). These models are called generalized multiplicative competitive interaction (GMCI) models. They are inspired from an aggregated version of the conditional multinomial logit (MNL) model (see Section 1.2.2).

#### GMCI attraction model

The concept of "attraction" of a brand is central in this literature, and is comparable to the "utility" concept in discrete choice models for individual data. The specification of the attraction of brand $j$ is a function of the explanatory variables (marketing variables usually, like price and media for example) describing this brand. The market share of brand $j$ is defined as its relative attraction compared to competitors, i.e. as its attraction divided by the sum of attractions of all the brands of the market:

$$0 < S_{jt} = \frac{\mathcal{A}_{jt}}{\sum_{l=1}^{D} \mathcal{A}_{lt}} < 1,$$

where $\mathcal{A}_{jt}$ is the attraction of brand $j$ at observation $t$ such that $\mathcal{A}_{jt} > 0$.

Cooper and Nakanishi [10] (p.36) defined a general model for market shares, called the generalized multiplicative competitive interaction model (GMCI). It is defined as follows:

$$\mathcal{A}_{jt} = \exp(a_j) \prod_{k=1}^{K} f_k(\check{X}_{kjt})^{b_k} \exp(\varepsilon_{jt}) \quad \text{and} \quad S_{jt} = \frac{\mathcal{A}_{jt}}{\sum_{l=1}^{D} \mathcal{A}_{lt}}, \tag{1.3}$$

where $\exp(\varepsilon_{jt})$ is a multiplicative random error term and $f_k$ is a monotonic transformation of $\check{X}_k$ such that $f_k(.) > 0$. If all $f_k$ are the identity function (resp. the exponential function), it leads to the MCI specification (resp. the MNL specification):

$$\textbf{MNL spec.: } S_{jt} = \frac{\exp(a_j + \sum_{k=1}^{K} b_k \check{X}_{kjt} + \varepsilon_{jt})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K} b_k \check{X}_{klt} + \varepsilon_{lt})} \tag{1.4}$$

$$\textbf{MCI spec.: } S_{jt} = \frac{\exp(a_j) \prod_{k=1}^{K} \check{X}_{kjt}^{b_k} \exp(\varepsilon_{jt})}{\sum_{l=1}^{D} \exp(a_l) \prod_{k=1}^{K} \check{X}_{klt}^{b_k} \exp(\varepsilon_{lt})} \tag{1.5}$$

The MNL specification of the GMCI is similar to the conditional multinomial logit model (MNL), except that in the MNL model an intercept has to be fixed to zero for identifiability reason (see equation 1.2). Note however that the attraction formulation of the MNL model differs from that of the GMCI models: the GMCI attraction contains

the random component $\varepsilon_{jt}$ whereas the MNL does not since the attraction form in that case corresponds to the expected share. We will further develop this aspect in Section 1.3.2.

**Estimation by OLS**

Contrary to the MNL model which is estimated by maximum likelihood based on the multinomial distribution, Nakanishi and Cooper [52] (p.109) propose an estimation method relying on a log linearization that they call "log-centering transformation" which is actually the log ratio between a share $S_{jt}$ and the geometric mean of all shares at observation $t$, $\widetilde{\mathbf{S}}_t$. This transformation is also called CLR (centered log-ratio) transformation in the compositional data analysis literature. The log-centered formulations are given by:

$$\textbf{MNL spec.:}\quad \log\left(\frac{S_{jt}}{\widetilde{\mathbf{S}}_t}\right) = a_1 + \sum_{l=2}^{D}(a_j - a_1)d_l + \sum_{k=1}^{K} b_k(\check{X}_{kjt} - \overline{\check{\mathbf{X}}}_{kt}) + (\varepsilon_{jt} - \overline{\varepsilon}_t),$$

$$\textbf{MCI spec.:}\quad \log\left(\frac{S_{jt}}{\widetilde{\mathbf{S}}_t}\right) = a_1 + \sum_{l=2}^{D}(a_j - a_1)d_l + \sum_{k=1}^{K} b_k \log\left(\frac{\check{X}_{kjt}}{\widetilde{\check{\mathbf{X}}}_{kt}}\right) + (\varepsilon_{jt} - \overline{\varepsilon}_t),$$

where $d_l = 1$ if $l = j$, 0 otherwise (brand dummy). $\overline{\mathbf{S}}_t$ and $\widetilde{\mathbf{S}}_t$ are respectively the arithmetic and the geometric means of $S_{jt}$.

This OLS estimation would be correct if error terms $\varepsilon^*_{jt} = (\varepsilon_{jt} - \overline{\varepsilon}_t)$ had a multivariate distribution with diagonal variance covariance matrix, but indeed the $\varepsilon^*_{jt}$ can only follow a degenerate multivariate normal distribution. This issue has not been clearly raised in the marketing literature. However, Nakanishi and Cooper [10] (p.125) suggest to use a generalized least squares (GLS) estimation instead of an OLS estimation because of the potential heteroscedasticity and/or correlation of error terms (if observations are time periods for example). But as stated in Cooper and Nakanishi [10] (p.128), we found that the GLS procedure, which is quite heavy in terms of implementation for this kind of models, does not give empirically better results than the OLS procedure. We will explain later in Chapter 2, Section 2.2.1 how to properly estimate this model using another transformation than the CLR transformation.

**Implementation in R:** the function lm() allows to fit the log-centered GMCI model by ordinary least squares.

### 1.2.4 Dirichlet covariate models

The Dirichlet distribution is the distribution of a composition obtained as the closure of a vector of $D$ independent gamma-distributed variables with the same scale parameter. Thus, it is another distribution adapted for variables lying in the simplex.

**Dirichlet distribution**

Let $\mathbf{S} = (S_1, \ldots, S_D) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_D)$ where $S_j > 0$ and $\sum_{j=1}^{D} S_j = 1$, $\alpha_j > 0$ and $\sum_{j=1}^{D} \alpha_j = \alpha_0$. $\alpha_0$ is called the precision parameter (when this value increases, the concentration around the expected value increases, the variance and covariance decrease). The density function of the Dirichlet distribution is defined by

$$f(\mathbf{S}) = \left( \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{D} \Gamma(\alpha_j)} \right) \prod_{j=1}^{D} S_j^{\alpha_j - 1},$$

with $\Gamma$ the Euler Gamma function. The expected value, the variance and the covariance of components are such that:

$$\mathbb{E}(S_j) = \frac{\alpha_j}{\alpha_0}, \quad Var(S_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \quad Cov(S_j, S_l) = -\frac{\alpha_j \alpha_l}{\alpha_0^2(\alpha_0 + 1)}$$

An alternative parametrization can be considered, using the parameters $\mu_j = \mathbb{E}(S_j)$ to account for the expected values of the shares, and $\phi = \alpha_0$ to account for the precision. The correspondence between this parametrization and the previous one is based on the fact that $\alpha_j = \mu_j \phi$ and $\alpha_0 = \phi$.

Two main criticisms of the Dirichlet model can be found in the literature: its negative covariance structure, and the strong independence between the initial gamma-distributed variables.

However, Campbell and Mosimann [6] show that the negative covariance structure is not an issue for Dirichlet covariate models, contrary to the simple Dirichlet distribution. Each observation indexed by $t$ follows a different Dirichlet distribution. The fact that the negative correlation happens between the shares of a same Dirichlet distribution does not imply that the vectors of shares coming from different Dirichlet distributions are negatively correlated. Indeed, the formula of generalized covariance proves that

$$Cov(X, Y) = \mathbb{E}[Cov(X, Y|Z)] + Cov[\mathbb{E}(X|Z), \mathbb{E}(Y|Z)]$$

Thus, if the covariance between the conditional expected values of two vectors of shares is positive and larger than the negative expected value of the conditional covariance between these two shares, then the unconditional covariance between the two shares can be positive[2].

In addition, Brehm et al. [4] show in a simulation study that the strong independence between the initial gamma-distributed variables (before closure) is not a problem: the Dirichlet covariate model successfully fits data with or without strong independence of variables before closure.

[2]The same argument can be used for the multinomial model.

### Dirichlet model

Campbell and Mosimann [6] developed the Dirichlet covariate model (called DIR hereafter) to explain a compositional dependent variable, supposed to be Dirichlet distributed, by classical (non compositional) covariates. Two parametrizations of this type of models exist. Hanssens et al. [27] (p.128) argue that the Dirichlet distribution seems to be adapted to the case of no sampling error in the data, which is true for our data base (we have the complete data base of registrations).

**Common parametrization** Under the common parametrization, the parameters of the Dirichlet distribution, the $\alpha_j$'s, are allowed to depend on the explanatory variables $\check{X}_k$ in a GLM (generalized linear model) fashion with a log link:

$$\log(\alpha_j(\check{\mathbf{X}}_t)) = a_j + \sum_{k=1}^{K} b_{kj}\check{X}_{kjt} \quad \text{and} \quad \mathbb{E}(S_j) = \frac{\alpha_j(\check{\mathbf{X}}_t)}{\sum_{l=1}^{D}\alpha_l(\check{\mathbf{X}}_t)} \tag{1.6}$$

The components, indexed by $j = 1, \ldots, D$, may have different explanatory variables (a different number of explanatory variables and/or explanatory variables which take different values for the different components), but for the sake of simplicity $\check{\mathbf{X}}$ denotes the vector of explanatory variables for all components.

**Alternative parametrization** Under the alternative parametrization, the model is defined by two equations:

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = a_j + \sum_{k=1}^{K} b_{kj}X_k \quad \text{and} \quad \log(\phi) = \gamma_0 + \sum_{k=1}^{K}\gamma_k Z_k$$

However, the alternative parametrization does not allow to use different explanatory variables for each component. Thus the common parametrization is preferred in our illustrative application.

### Estimation by maximum likelihood

As explained in Hijazi and Jernigan [29], "a different Dirichlet distribution is modeled for every value of the explanatory variables, resulting in a conditional Dirichlet distribution". The conditional distributions $\mathbf{S}_t|\check{\mathbf{X}}_t$ are mutually independent. We assume $\mathbf{S}_t|\check{\mathbf{X}}_t \sim \mathcal{D}(\alpha_1(\check{\mathbf{X}}_t), \ldots, \alpha_D(\check{\mathbf{X}}_t))$, with unknown parameters. The log-likelihood to maximize is:

$$\log L(\mathbf{S}|\alpha(\check{\mathbf{X}})) = \sum_{t=1}^{T}\left[\log\Gamma\left(\sum_{j=1}^{D}\alpha_j(\check{\mathbf{X}}_t)\right) - \sum_{j=1}^{D}\log\Gamma(\alpha_j(\check{\mathbf{X}}_t)) + \sum_{j=1}^{D}(\alpha_j(\check{\mathbf{X}}_t)-1)\log S_{jt}\right]$$

**Implementation in R:** the package **DirichReg** created by Maier [40] allows to fit Dirichlet model for the common or alternative parametrization, by maximum likelihood.

### 1.2.5 Compositional models

Compositional data analysis was developed in the 80's by John Aitchison [1]. The first applications were in the geological field, with the objective to analyze the composition of a rock sample in terms of the relative presence of different chemical elements. More generally, CODA aims to analyze relative information between the components (parts) of a composition, where the total of the components is not relevant or is not of interest, taking into account the constraints of the simplex space.

#### The log-ratio transformation approach

As it is not possible to use properly classical statistical methods (e.g. linear regression models) on constrained data like compositions, a log-ratio transformation of compositions can be used in order to obtain unbounded coordinates in $\mathbb{R}$. Then, usual tools can be used on coordinates, and results in the simplex can be recovered by inverse transformation, thus enforcing the simplex constraints. Several transformations are proposed: notably the ALR (additive log-ratio), the CLR (centered log-ratio) and the ILR (isometric log-ratio) transformations (see Egozcue et al. [13]).

- The ALR transformation is the first transformation proposed by Aitchison in 1986. It is defined as $alr(\mathbf{S}) = \left(\log \frac{S_1}{S_D}, \ldots, \log \frac{S_{D-1}}{S_D}\right)'$. Its inverse transformation is $\mathbf{S} = alr^{-1}(alr(\mathbf{S})) = \mathcal{C}(\exp(alr(\mathbf{S})_1), \ldots, \exp(alr(\mathbf{S})_{D-1}, 1)'$.

- The CLR transformation leads to $D$ coordinates which satisfy the constraint of zero sum (it is not reducing the dimension of the composition). It is defined as $clr(\mathbf{S}) = \left(\log \frac{S_1}{\widetilde{\mathbf{S}}}, \ldots, \log \frac{S_D}{\widetilde{\mathbf{S}}}\right)'$, where $\widetilde{\mathbf{S}}$ is the geometric mean of the $D$ components. Its inverse transformation is $\mathbf{S} = clr^{-1}(clr(\mathbf{S})) = \mathcal{C}(\exp(clr(\mathbf{S})_1), \ldots, \exp(clr(\mathbf{S})_D))'$.

- The ILR transformation consists in a projection of components in an orthonormal basis of $\mathcal{S}^D$ in order to obtain $D-1$ orthonormal coordinates. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ be an arbitrary orthonormal basis in $\mathbb{R}^{D-1}$, then $\mathbf{e}_l = clr^{-1}(\mathbf{v}_l)$, for $l = 1, \ldots D-1$, represent an orthonormal basis in the simplex $\mathcal{S}^D$ equipped with its "natural geometry" (see Pawlowsky-Glahn et al. [56]). Considering the $D \times (D-1)$ matrix $\mathbf{V}$ with columns $\mathbf{v}_l = clr(\mathbf{e}_l)$, the ILR coordinates are defined as $ilr(\mathbf{S}) = \mathbf{S}^* = \mathbf{V}' clr(\mathbf{S}) = \mathbf{V}' \log(\mathbf{S})$. The inverse transformation is given by $\mathbf{S} = ilr^{-1}(\mathbf{S}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{S}^*))'$.

**Example of ILR transformation:** A possible ILR transformation is defined by

$$S_l^* = \sqrt{\frac{D-l}{D-l+1}} \log \frac{S_l}{(\prod_{l'=l+1}^{D} S_{l'})^{\frac{1}{D-l}}}, \quad l = 1, \ldots, D-1,$$

where $S_1^*$ contains all the relative information of the share $S_1$ to the shares $S_2, \ldots, S_D$ ($S_1^*$ is the only coordinate which includes $S_1$ and compares it to the rest of the composition).

If $D = 3$ for example, it leads to $S_1^* = \sqrt{\frac{2}{3}} \log \frac{S_1}{\sqrt{S_2 S_3}} = \sqrt{\frac{2}{3}} \log S_1 - \frac{1}{\sqrt{6}} (\log S_2 + \log S_3)$ and $S_2^* = \sqrt{\frac{1}{2}} \log \frac{S_2}{S_3} = \frac{1}{\sqrt{2}} (\log S_2 - \log S_3)$. Thus, the balance matrix $\mathbf{V}$ is equal to

$$\mathbf{V} = \begin{bmatrix} \sqrt{2/3} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$$

For regression as well as for some other statistical analysis, the results are the same after inverse transformation regardless of the chosen transformation. However, as ALR is isomorphic but not isometric, and CLR introduces collinearity between coordinates, ILR is preferred for compositional regression models.

### CODA regression models

Compositional regression models are of different types depending on whether the response variable and/or the explanatory variables are compositional. We focus here on the case where the dependent as well as the explanatory variables are compositional and of same dimension $D$ (for example, market shares of $D$ brands are explained by the corresponding media investments)[3].

CODA models can be expressed either in terms of the initial compositional observations in the simplex (equation (1.7)) or alternatively in terms of the corresponding transformed coordinates in the Euclidean space (equation (1.8)), as explained below.

**- Linear CODA model in the simplex (in terms of compositions):**

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{k=1}^{K} \mathbf{B_k} \boxdot \mathbf{X_{k}}_t \oplus \boldsymbol{\varepsilon}_t, \tag{1.7}$$

with $\mathbf{S}, \mathbf{a}, \mathbf{X_k}, \boldsymbol{\varepsilon} \in \mathcal{S}^D$ and $\mathbf{B_k} \in \mathbb{R}_{D \times D}$ such that row and column sums are equal to zero[4]. The following operations are used in the simplex:

- $\oplus$ is the *perturbation operation*, corresponding to the addition operation in the real (Euclidean) geometry: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)' \quad$ with $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$
- $\odot$ is the *power transformation*, corresponding to the multiplication operation in the real geometry: $\mathbf{x} \odot \lambda = \mathcal{C}(x_1^\lambda, \ldots, x_D^\lambda)' \quad$ with $\lambda \in \mathbb{R}, \mathbf{x} \in \mathcal{S}^D$
- $\boxdot$ is the *compositional matrix product*, corresponding to the matrix product in the real geometry: $\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\big(\prod_{j=1}^D x_j^{b_{1j}}, \ldots, \prod_{j=1}^D x_j^{b_{Dj}}\big)' \quad$ with $\mathbf{B} \in \mathbb{R}_{D \times D}, \mathbf{x} \in \mathcal{S}^D$
- The *simplicial inner product* is given by: $< x, y >= \frac{1}{D} \sum_{j=1}^D \sum_{l=1}^D \log \frac{x_j}{x_l} \log \frac{y_j}{y_l}$

[3]The dependent and the explanatory compositions can be of different dimensions (Chen et al. [8]).

[4]Under these conditions, $\mathbf{B} \boxdot \mathbf{X}$ is an endomorphism of the simplex $\mathcal{S}^D$ (See Kynclova et al. [35]). Thus model (1.7) is a linear model in the simplex.

**- Linear CODA model in the Euclidean space (in terms of ILR coordinates):**

$$S^*_{jt} = a^*_j + \sum_{k=1}^{K} \sum_{m=1}^{D-1} b^*_{kjm} X^*_{kmt} + \varepsilon^*_{jt} \quad \forall\, j \in 1, \ldots, D-1, \tag{1.8}$$

where, the stars denote the ILR transformed version of variables, $j$ is the index of $\mathbf{S}$'s ILR coordinates, $m$ is the index of $\mathbf{X}$'s ILR coordinates and $\varepsilon^*_j \sim \mathcal{N}(0, \sigma^2_j)$. Equation (1.8) corresponds to a system of $D-1$ linear models, one for each ILR coordinate of $\mathbf{S}$. Note here that compositional explanatory variables coordinates can be equivalently calculated using $\check{\mathbf{X}}$ (volumes) or $\mathbf{X}$ (shares), idem for the dependent variable.

The second presentation of the CODA model (equation (1.8)) has the advantage to be a system of classical linear models but its connection with the original data is obscured by the ILR transformation. On the other hand, the first presentation in terms of the original share data (equation (1.7)) is obscured by the simplex operations involved in the model equation. However, we show in 1.3.2 that this model can be expressed in a so-called attraction formulation so that it is not needed to be familiar with simplex notations detailed above to understand and use this compositional model.

**Estimation by OLS**

After log-ratio transformation, the estimation is usually done with the OLS method separately on the $D-1$ linear models expressed in coordinates (equation (1.8)). The orthonormality of coordinates allows us to estimate the $D-1$ models separately. Then, the estimated model can be back transformed into the simplex using the inverse transformation which transforms $\mathbf{a}^*$ into $\mathbf{a}$, $\mathbf{b}^*$ into $\mathbf{b}$, $\mathbf{S}^*$ into $\mathbf{S}$ and $\mathbf{X}^*$ into $\mathbf{X}$:

$$\begin{aligned} \mathbf{a} &= ilr^{-1}(a^*_1, \ldots, a^*_{D-1}) = \mathcal{C}(\exp(\mathbf{V}\mathbf{a}^*)) \\ \mathbf{B}_{D,D} &= \mathbf{V}\mathbf{B}^*_{D-1,D-1}\mathbf{V}' \\ \mathbf{S} &= ilr^{-1}(S^*_1, \ldots, S^*_{D-1}) = \mathcal{C}(\exp(\mathbf{V}\mathbf{S}^*)) \end{aligned}$$

with $\mathbf{B}^* = \begin{bmatrix} b^*_{1,1} & \ldots & b^*_{1,D-1} \\ \ldots & b^*_{j,l} & \ldots \\ b^*_{D-1,1} & \ldots & b^*_{D-1,D-1} \end{bmatrix}$, and $\mathbf{B} = \begin{bmatrix} b_{1,1} & \ldots & b_{1,D} \\ \ldots & b_{j,l} & \ldots \\ b_{D,1} & \ldots & b_{D,D} \end{bmatrix}$ where $b^*_{j,l}$ is the parameter corresponding to the impact of $X^*_l$ on $S^*_j$, and $b_{j,l}$ is the parameter corresponding to the impact of $X_l$ on $S_j$.

**Implementation in R:** the packages **compositions** [67] and "robCompositions" [63] allow to transform compositional data, to fit the compositional model by OLS on the coordinates and to back transform the results into compositions. Implementation of the CODA model using R is presented in the book of Van den Boogaart and Tolosana-Delgado [68].

### 1.2.6 Alternative models

In the literature, some articles mix compositional data analysis and aggregated choice models. In Bechtel [3] and in Fry and Chong [19], the shares are specified according to a nested multinomial logit model which does not embody the IIA property (see Section 1.3.3). They use an additive log-ratio transformation of their model (ALR) as can be found in the compositional analysis, in order to be able to estimate the model by OLS or GLS.

Some authors propose to transform compositional data to directional data by the square root transformation mapping the simplex into the unit hypersphere. Wang et al. [70] further use classical regression models for the polar coordinates, whereas Scealy and Welsh [59] use the additive Kent regression model.

## 1.3 Theoretical comparison of share models

In this section, we highlight the similarities and differences of the presented models from a theoretical perspective. Because these models are deeply linked with the type of applications they have been proposed for, the following comparison refers not only to statistical properties, but also to econometric and marketing properties. Table 1.2 summarizes the distributional assumptions, the estimation methods, the properties and the complexity of each model[5]. These items are discussed in detail below. Finally we highlight the relationship between the GMCI and the CODA models, notably the fact that GMCI can be expressed in a compositional way and that it is a particular case of the CODA model.

### 1.3.1 Distributional assumptions

In the MNL model the dependent variable is a vector of positive numbers $\check{S}_j$ which follow a multinomial distribution. In the other three models the dependent variable is directly the vector of shares $S_j$ which are Dirichlet distributed in the case of DIR and Gaussian in the simplex distributed for GMCI and CODA (the coordinates are Gaussian in the transformed space). Note that the MNL model differs from the MNL specification of the GMCI model by its underlying distributional assumptions.

MNL and Dirichlet models belong to the family of GLM (generalized linear models): see Peyhardi et al. [57] for MNL and Maier [40] for DIR. GMCI and CODA models belong to the family of transformation models (TRM hereafter) in which a classical linear model is postulated in the transformed space.

### 1.3.2 Expected shares and attraction formulation

#### Expected value of shares

Let us notice that the model formulation of the two GLM models - MNL (1.2) and DIR (1.6) - involves the expected shares $\mathbb{E}(S_{jt}|\check{\mathbf{X}}_t)$, while the two transformation models formulation - GMCI (1.3) and CODA (1.7) - involves the random shares $S_{jt}$ and a random error term. The usual expected value cannot be analitically computed for the GMCI and the CODA models. For this reason, we turn attention to the "expected value in the simplex", defined as follows (see Theorem 6.10 p.109 in Pawlowsky-Glahn et al. [56]):

$$\mathbb{E}^{\oplus}\mathbf{S} = \mathcal{C}(\exp(\mathbb{E}\log\mathbf{S})) = clr^{-1}(\mathbb{E}clr(\mathbf{S})) = ilr^{-1}(\mathbb{E}ilr(\mathbf{S})) = ilr^{-1}(\mathbb{E}\mathbf{S}^*)$$

This means that the expected value in the simplex of the composition of shares, $\mathbb{E}^{\oplus}\mathbf{S}$, coincides with the ILR back transformation of expected values of the random coordinates, $\mathbb{E}\mathbf{S}^*$.

[5]Here the GMCI model is presented with the MCI specification. Note that if $\check{X}$ is replaced by $\exp\check{X}$, it corresponds to the MNL specification.

**Remark:** If the explanatory variables only consist of intercepts, what we call "constant model" in Chapter 3, the fitted shares are not the same across the four models. In the case of the CODA and the GMCI models, they correspond to the center of the compositional data (the closed vector of geometric means of each component), while in the case of the MNL and the DIR models, fitted shares are linked to the arithmetic means of components (weighted in the case of MNL). The geometric mean, which is coherent with the simplex geometry, can be more adapted than the arithmetic mean to summarize shares data, as illustrated in Figure 1.1. This is an argument in favor of CODA and GMCI models.

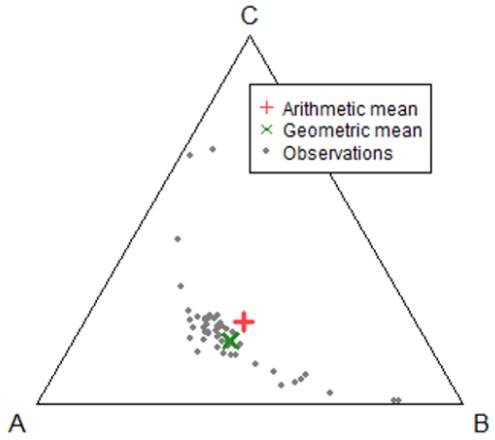


Figure 1.1: Arithmetic and geometric means of a compositional data set in a ternary diagram

### Attraction formulation of share models

As seen before, the attraction formulations in MNL and GMCI are different (in GMCI it includes a random error term). In order to unify the presentation, we introduce a deterministic attraction $A_{jt}$ and a random attraction $u_{jt}$ such that $\mathcal{A}_{jt} = A_{jt}u_{jt}$. According to equations (1.2), (1.4), (1.5), the deterministic attraction formulations of MNL and of the two GMCI models ($G_{MNL}$ for the MNL specification and $G_{MCI}$ for the MCI specification) are

$$A_{jt}^{MNL} = \exp(a_j + \sum_{k=1}^{K} b_k \check{X}_{kjt}) \quad \text{with } a_D = 0 \quad \Leftrightarrow \quad \mathbb{E}S_{jt} = \frac{A_{jt}^{MNL}}{\sum_{m=1}^{D} A_{mt}^{MNL}}$$

$$A_{jt}^{G_{MNL}} = \exp(a_j + \sum_{k=1}^{K} b_k \check{X}_{kjt}) \quad \Leftrightarrow \quad \mathbb{E}^{\oplus}S_{jt} = \frac{A_{jt}^{G_{MNL}}}{\sum_{m=1}^{D} A_{mt}^{G_{MNL}}}$$

$$A_{jt}^{G_{MCI}} = \exp(a_j) \prod_{k=1}^{K} \check{X}_{kjt}^{b_k} \quad \Leftrightarrow \quad \mathbb{E}^{\oplus}S_{jt} = \frac{A_{jt}^{G_{MCI}}}{\sum_{m=1}^{D} A_{mt}^{G_{MCI}}}$$

This emphasizes the fact that the type of expected shares involved in the attraction formulation are different between MNL and the MNL specification of GMCI.

The Dirichlet model can also be expressed with an attraction formulation:

$$A_{jt}^{DIR} = \exp(a_j + \sum_{k=1}^{K} b_{kj} \check{X}_{kjt}) = \alpha_{jt} \qquad \Leftrightarrow \quad \mathbb{E}S_{jt} = \frac{A_{jt}^{DIR}}{\sum_{m=1}^{D} A_{mt}^{DIR}}$$

This highlights the fact that the parameters of the DIR model are alternative-specific (they depend on $j$), contrary to the GMCI and MNL models.

We now derive the attraction form of the CODA model, using equation (1.7). We first express the market share of brand $j$ in the CODA model as

$$\mathbf{S}_t = \mathbf{a}_t \bigoplus_{k=1}^{K} \mathbf{B_k} \boxdot \mathbf{X_k}_t \oplus \boldsymbol{\varepsilon}_t = \mathcal{C}\left( a_1 \prod_{k=1}^{K} \prod_{l=1}^{D} \check{X}_{klt}^{b_{k1l}} \varepsilon_{1t}, \ldots, a_D \prod_{k=1}^{K} \prod_{l=1}^{D} \check{X}_{klt}^{b_{kDl}} \varepsilon_{Dt} \right)$$

Note that the market share $S_{jt}$ is expressed as a function of volumes $\check{X}_{klt}$ directly and not as a function of shares $X_{klt}$: it turns out that it is exactly equivalent because $S_{jt}$ is obtained by a closure operation.

Let us now define the deterministic attraction of the CODA model:

$$A_{jt}^{CODA} = a_j \prod_{k=1}^{K} \prod_{l=1}^{D} \check{X}_{klt}^{b_{kjl}} = \exp(\log(a_j)) \prod_{k=1}^{K} \prod_{l=1}^{D} \check{X}_{klt}^{b_{kjl}} \tag{1.9}$$

Then, we have:

$$\mathbb{E}^{\oplus} S_{jt} = \frac{A_{jt}^{CODA}}{\sum_{m=1}^{D} A_{mt}^{CODA}}$$

The attraction formulation of the CODA model is very close to the MCI attraction, but it includes cross effects between components: in the CODA model, the attraction of the component $j$ does not depend only on explanatory variables relative to the component $j$ but also on explanatory variables of other components $l \neq j$. See Section 1.3.5 for a deeper comparison of the GMCI and the CODA models.

### 1.3.3 Properties

We now discuss whether the properties that have been introduced and established in the literature for a given model are valid for the other ones.

#### IIA and subcompositional coherence

In the econometric literature, an important question often discussed is whether or not a choice model satisfies the IIA (independence from irrelevant alternatives) property. IIA means that the ratio of shares of an alternative $j$ with respect to an alternative $l$ only depends on the characteristics of $j$ and $l$ and is not affected by the presence or absence of

irrelevant alternatives. This property allows to simplify the models but it is not always realistic (see the famous red bus - blue bus example of McFadden [43]). MNL, GMCI and DIR models satisfy IIA but the CODA model does not because of the cross effects between brands.

In the CODA literature, the subcompositional coherence property (see Pawlowsky-Glahn [56]) means that the results of an analysis made on a subcomposition (i.e. remove some alternatives) should not contradict the results of the analysis made on the whole composition. This is coming from the fact that compositional data analysis is based on the use of log ratios. However, if we look at equation (1.9), we can see that the market share of brand $j$ is determined by the explanatory variables of all the brands. Thus, subcompositional coherence does not imply IIA, but the reciprocal is true. In the econometrics literature, it is considered that IIA can be a severe limitation, which is a positive point for CODA models.

#### Invariance

The **scale invariance** is the fact that multiplying the count data by a constant does not affect the estimation results. It is a desirable property satisfied by the four models.
The **permutation invariance** is a desirable property corresponding to invariance through a permutation of the components of a composition. It is clearly satisfied by all the described models.
The **perturbation invariance** corresponds to coherence when performing a change of units possibly different for each component of a composition. For example, we can model brands' market shares in terms of sales volumes or in terms of sales values (that is sales volumes perturbed by the vector of prices). The estimated market shares and parameters from the "volume" model should be equal to those of the "value" model after perturbation by the vector of prices. This property is satisfied by CODA and GMCI models. We can show empirically that it is not satisfied by MNL and DIR.

### 1.3.4 Model complexity

In MNL, GMCI and DIR models, the deterministic attraction $A_{jt}$ is a function of the explanatory variables characterizing the alternative $j$ only, leading to the absence of cross effects. However in the DIR model, parameters are alternative-specific, which increases the complexity of the model. In the CODA model, the attraction may depend on all alternative characteristics, inducing alternative-specific and cross effect parameters. This is why CODA is the most complex model with the higher number of parameters.

It is not possible to estimate all cross effects in the MNL model (see So and Kuhfeld [61]). Cross effects can be incorporated in the GMCI model (in a so-called fully extended MCI model, see Cooper and Nakanishi [10], p.61) and in the Dirichlet models but the number of parameters to be estimated dramatically increases. CODA is relatively parsimonious in the sense that it allows to incorporate all cross effects with a number of parameters relatively lower than the other models (proportional to $(D-1)^2$ versus $D^2$ for others), thanks to the dimension reduction of the ILR transformation.

It is interesting to see that using the same dependent and explanatory variables, the complexity is totally different from one model to another. For example (as in our application, see Section 1.4), if the number of components of the dependent variable is $D = 3$, explained by $K_X = 7$ compositions of size $D = 3$ and $K_Z = 1$ time-dependent variable, the number of estimated parameters are the following: 11 for MNL, 13 for GMCI, 27 for DIR and 32 for CODA. With 32 parameters, the CODA model reflects all the cross effects between shares whereas the DIR and the GMCI models with cross effects would require 69 parameters ($D(1+D\times K_X+K_Z)$). Note also that the number of parameters increases dramatically with the number of components (brands), especially in the CODA model. For example if $D$ becomes equal to 5 (with $K_X$ and $K_Z$ fixed), the number of parameters become 15, 17, 45, and 120, which can be a serious limitation for the CODA model.

### 1.3.5 Relationship between GMCI and CODA models

#### Compositional form of the GMCI model

Even though the GMCI estimation procedure uses a log-ratio transformation as the CODA model, the two models are different and we are now going to express the GMCI model in a compositional form, which will reveal this difference.

Wang et al. [70] propose in 2013 a compositional regression model for the case where both dependent and explanatory variables are compositional, but their model is simpler than the CODA model presented in paragraph 1.2.5: this model does not include cross effects between components contrary to the CODA model.

Actually Wang et al.'s model is exactly similar to the MCI model proposed by Cooper and Nakanishi in 1988 [10] (p.6), except that they use ILR coordinates for the estimation while CLR coordinates are used in the MCI model. From this correspondence we derive a compositional form for the GMCI model:

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{k=1}^{K} b_k \odot \mathbf{X}_{\mathbf{k}t} \oplus \boldsymbol{\varepsilon}_t \tag{1.10}$$

$$\Leftrightarrow S_{jt} = \frac{a_j \prod_{k=1}^{K} \check{X}_{kjt}^{b_k} \varepsilon_{jt}}{\sum_{l=1}^{D} a_l \prod_{k=1}^{K} \check{X}_{klt}^{b_k} \varepsilon_{lt}} = \frac{\exp(\log a_j + \sum_{k=1}^{K} b_k \log \check{X}_{kjt} + \log \varepsilon_{jt})}{\sum_{l=1}^{D} \exp(\log a_l + \sum_{k=1}^{K} b_k \log \check{X}_{klt} + \log \varepsilon_{lt})}$$

Note that the previous equation is derived for the MCI specification, but the expression can be recovered for the MNL specification if $\check{X} = \exp \check{x}$.

Equation (1.10) highlights the similarities and differences between GMCI and CODA models: in place of the $B_k$ matrix in equation (1.7) of the CODA model, we have a single $b_k$ parameter for all components of the explanatory composition in the GMCI model.

#### GMCI model: a particular case of the CODA model

One can show that the MCI model, when estimated using the ILR transformation, is a particular case of the CODA model where the dimension of the dependent variable and

those of compositional explanatory variables are equal ($D_S = D_X$), and where $B^*$ is a diagonal matrix with $b^* = b$ on the diagonal and 0 otherwise, that is where only the $j^{\text{th}}$ ILR coordinates of compositional explanatory variables are relevant to explain the $j^{\text{th}}$ ILR coordinates of the dependent variable (see the appendix A.1.1 for a proof in the case of $D = 3$).

However, the so-called differential MCI (DMCI) model in marketing (see Cooper and Nakanishi [10], p.58), where brand specific parameters without cross effect are specified, is not a particular case of the CODA model and we can prove that it is not scale invariant (see the appendix A.1.2). This means that fitting the DMCI model using explanatory variables in euro or in thousands euros does not give the same estimated market shares, which is not acceptable. Then, the DMCI model is not consistent with the simplicial geometry and we strongly recommend not to use it.

Concerning the fully extended MCI model (FEMCI) which is a MCI model with cross effects between brands (see Cooper and Nakanishi [10], p.61), it is interesting to highlight the fact that it is similar to the CODA model in the sense that the deterministic attraction of the two models is the same (equation (1.9)). However, the estimating method is different: they suggest to estimate the FEMCI model by OLS on the CLR coordinates using dummy variables (see equations (5.27) and (5.28) in [10], p.144) implying constant variance of error terms across the CLR coordinates, whereas the CODA model is usually estimated by OLS, separately on the different ILR coordinates, allowing non constant variance across the ILR coordinates. Note that the issue raised in Section 1.2.3 concerning the independence of error terms when CLR coordinates are used still holds here. Moreover, Cooper and Nakanishi pointed out the fact that their method can only lead to the estimation of centered coefficients, $b^*_{kjl} = b_{kjl} - \bar{b}_{k.l}$, but they argued that the $b^*_{kjl}$ are sufficient for interpreting the model. In the CODA model, the estimation allows to obtain the estimated coefficients directly. Thus, we strongly support the use of the CODA model when cross effects are considered.

Table 1.2: Benchmark of models for explaining shares

| Name | Expected shares | Distribution | Estimation | Properties** | Nb param. |
|---|---|---|---|---|---|
| **MNL** GLM* type | $\mathbb{E}S_{jt} = \frac{\exp(a_j + \sum_{k=1}^{K_X} b_k \check{X}_{kjt} + \sum_{\kappa=1}^{K_Z} b_{\kappa j} Z_{\kappa t})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K_X} b_k \check{X}_{klt} + \sum_{\kappa=1}^{K_Z} b_{\kappa l} Z_{\kappa t})}$<br>with $a_1 = 0$ for identifiability reasons. | $(\check{S}_{1t}, \dots \check{S}_{Dt}) \sim \mathcal{MN}(\check{S}_t, s_{1t}, \dots, s_{Dt})$<br>Indep. distributed across $t$ | Maximum Likelihood | Permutation invariance, Scale invariance, Perturbation invariance, IIA, Subcompo. coherence | $(D-1) \times (1+K_Z) + K_X$ |
| **GMCI** (MCI spec.) TRM* type | Share: $\mathbb{E}^{\oplus} S_{jt} = \frac{a_j \prod_{k=1}^{K_X} \check{X}_{kjt}^{b_k} \prod_{\kappa=1}^{K_Z} b_{\kappa j}^{Z_{\kappa t}}}{\sum_{m=1}^{D} a_m \prod_{k=1}^{K_X} \check{X}_{kmt}^{b_k} \prod_{\kappa=1}^{K_Z} b_{\kappa m}^{Z_{\kappa t}}}$<br>Equivalently in terms of CLR coordinate: $\mathbb{E}\log\left(S_{jt}/\widetilde{\mathbf{S}}_t\right) = \mathbb{E}clr(\mathbf{S})_{jt} = a_1 + \sum_{j'=2}^{D} a'_{j'} d_{j'} + \sum_{k=1}^{K_X} b_k \log\left(\check{X}_{kjt}/\widetilde{\check{\mathbf{X}}}_{kt}\right) + \sum_{\kappa=1}^{K_Z} \left(b_{\kappa 1} Z_{\kappa t} + \sum_{j'=2}^{D} b'_{\kappa j'} Z_{\kappa t} d'_j\right)$ | $clr(\mathbf{S}_t) \sim \mathcal{N}_D(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$<br>with $\mathcal{N}_D$ the multivariate normal distribution (degenerate here).<br>Indep. distributed across $t$ | OLS on coordinates | Permutation invariance, Scale invariance, Perturbation invariance, IIA, Subcompo. coherence | $D \times (1 + K_Z) + K_X$ |
| **DIR** GLM type | With the common parametrization: $\mathbb{E}S_{jt} = \frac{\exp(a_j + \sum_{k=1}^{K_X} b_{kj} \check{X}_{kjt} + \sum_{\kappa=1}^{K_Z} b_{\kappa j} Z_{\kappa t})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K_X} b_{kj} \check{X}_{klt} + \sum_{\kappa=1}^{K_Z} b_{\kappa l} Z_{\kappa t})}$<br>$\log \alpha_{jt} = a_j + \sum_{k=1}^{K_X} b_{kj} \check{X}_{kjt} + \sum_{\kappa=1}^{K_Z} b_{\kappa j} Z_{\kappa t}$ | $(S_{1t}, \dots S_{Dt}) \sim \mathcal{D}(\alpha_{1t}, \dots, \alpha_{Dt})$<br>Indep. distributed across $t$ | Maximum likelihood | Permutation invariance, Scale invariance, Perturbation invariance, IIA, Subcompo. coherence | $D \times (1 + K_X + K_Z)$ |
| **CODA** TRM type | Composition in the simplex: $\mathbb{E}^{\oplus} \mathbf{S}_t = \mathbf{a} \bigoplus_{k=1}^{K_X} \mathbf{B_k} \boxdot \mathbf{X}_{kt} \bigoplus_{\kappa=1}^{K_Z} Z_{\kappa t} \odot \mathbf{b}_\kappa$<br>Equivalently in terms of share in the simplex: $\mathbb{E}^{\oplus} S_{jt} = \frac{a_j \prod_{l=1}^{D} \prod_{k=1}^{K_X} \check{X}_{klt}^{b_{kjl}} \prod_{\kappa=1}^{K_Z} b_{\kappa j}^{Z_{\kappa t}}}{\sum_{m=1}^{D} a_m \prod_{l=1}^{D} \prod_{k=1}^{K_X} \check{X}_{klt}^{b_{kml}} \prod_{\kappa=1}^{K_Z} b_{\kappa m}^{Z_{\kappa t}}}$<br>Equivalently in terms of ILR coordinates: $\mathbb{E}S^*_{jt} = a_j + \sum_{k=1}^{K_X} \sum_{m=1}^{D-1} b^*_{kjm} X^*_{kmt} + \sum_{\kappa=1}^{K_Z} b^*_{\kappa,j} Z_{\kappa t}$ | $\mathbf{S}_t \sim \mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$<br>with $\mathcal{N}_{\mathcal{S}}$ the normal distribution on the simplex, $\mu$ a mean vector, $\boldsymbol{\Sigma}$ a diagonal variance matrix.<br>$\mathbf{S}^* = ilr(\mathbf{S}_t) \sim \mathcal{N}_{D-1}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ | OLS on coordinates | Permutation invariance, Scale invariance, Perturbation invariance, IIA, Subcompo. coherence | $(D-1) \times ((D-1)K_X + K_Z + 1)$ |

*GLM: generalized linear model; TRM: transformation model.

** Properties in black are satisfied and properties in grey are not.

## 1.4 Empirical comparison of share models

In this section, we use the MNL, GMCI, DIR and CODA models for a concrete case study and we demonstrate that DIR and CODA models can outperform usual market share models. After presenting the application and the data of our illustrative example, a cross-validation process is proposed based on quality measures adapted for shares models for the four types of models. Finally, we compare the interpretation of the parameters of the four models in terms of elasticities.

### 1.4.1 Application and data

The main objective of this application is to understand the impact of media investments on brands' market shares controlling for other factors like price and scrapping incentive. In each model specification, the interest is on the marginal impact of each media channel on relative sales, that is on the elasticities of market shares relative to media investments by channel.

The French automobile market is segmented in five segments, from A to E, according to the size of the chassis. We focus here on the B segment, which represents half of the sales in France in terms of volume. The B segment corresponds to small mainstream vehicles, like the Renault Clio which is the most famous vehicle of this segment in France. More precisely, following the subcompositional coherence property of the compositional data analysis, we focus on three particular brands of this segment: Renault, Nissan and Dacia ($D = 3$).

The studied period, running from June 2005 to August 2015, is characterized by the birth of Dacia on the French automobile market, a low-cost brand belonging to Renault, at the beginning of 2005. It is also characterized by the economic crisis which has hurt the French automobile market a lot from 2008 to 2012. The French government tried to help this market setting up a scrapping incentive from December 2008 to December 2010, to promote the replacement of old vehicles with new environmentally-friendly vehicles, which has artificially boosted the sales during 2009 and 2010. Note that Dacia increased a lot its market share during the crisis thanks to its low price. These facts have to be kept in mind in order to understand the evolution of market shares, and it justifies the use of a scrapping incentive dummy as control variable.

The ternary diagram allows to represent compositions of 3 components in the simplex (see Van den Boogaart and Tolosana-Delgado [68]). Figure 1.2 represents for example the annual market shares of Dacia, Nissan and Renault from 2005 to 2014. We can see easily that Dacia increases its market share easily at the expense of Renault from 2005 to 2010.

The four models are applied to an automobile market data set containing for each brand of the B segment the sales volume in units $\check{S}_{jt}$, the catalog price in euro $P_{jt}$, the media investments by channel in euro at time $t-4$, $M_{cj,t-4}$ (television, press, radio, outdoor, digital, cinema), and the periods of scrapping incentive $SI_t$ (dummy variable), monthly from June 2005 to August 2015 ($T = 123$ periods of observation). According

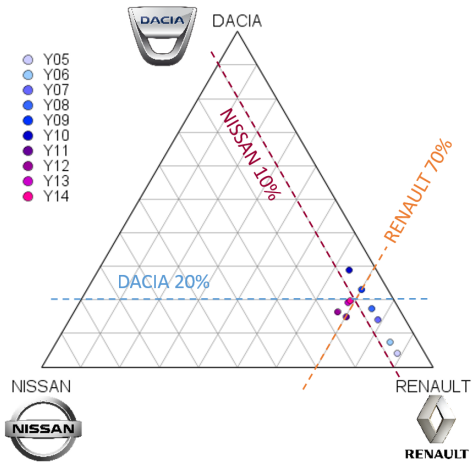


Figure 1.2: Ternary diagram of annual market shares of Dacia, Nissan and Renault

to the marketing literature, it is preferable to use the logarithm of price instead of the raw price. The reason of that is linked to the shape of the elasticity of market shares to the price. Moreover, to keep the market shares equal, the logged variables have to increase in the same proportion while the non-logged variables have to increase by the same amount. Indeed, for our application, using the log of price instead of the price gives best in-sample fits for the four models. The media investments have to be considered with a lag with respect to sales. Statistically in this application, a lag of four months gave the best results on the four models. To avoid the problem of zeros due to the use of logarithm, when media investments are equal to zero, we replace them by one euro, which is a very small amount compared to the non-zero investments (see A.1.3 in the appendix). Table 1.3 summarizes the four models.

### 1.4.2 A cross-validation comparison

The repeated random sub-sampling cross-validation process described below is used to compute out-of-sample prediction accuracy measures on the four considered models, in order to avoid an over-fitting effect and to compare the considered models which do not have the same number of parameters.

1. Randomly draw a sub-sample of 100 observations among 123, resulting in 81% (100) in-sample observations and 19% (23) out-of-sample observations
2. Fit the 4 models to the sub-sample, store the fitted parameters
3. Apply the 4 models to the out-of-sample observations, store the fitted values of the shares
4. Compute the quality measures using the out-of-sample predicted share values
5. Iterate 100 times steps 1 to 4
6. Compute the average quality measures using the out-of-sample predicted share values over the 100 iterations

Table 1.3: MNL, GMCI, DIR and CODA fitted models

| Name | Model |
|---|---|
| **MNL** | Estimation by ML using the $\check{S}_{jt}$ |
| | $$\mathbb{E}(S_{jt}\|M_{t-4},P_t,SI_t)=\frac{\exp(a_j+\sum_{c=1}^{C}b_cM_{cj,t-4}+b_P\log P_{jt}+b_{SIj}SI_t)}{\sum_{l=1}^{D}\exp(a_l+\sum_{c=1}^{C}b_cM_{cl,t-4}+b_P\log P_{lt}+b_{SIl}SI_t)}$$ |
| | with $a_1=b_{SI1}=0$ for identifiability reasons |
| **GMCI** | Estimation by OLS on the CLR coordinates |
| | $$S_{jt}=\frac{\exp(a_j+\sum_{c=1}^{C}b_cM_{cj,t-4}+b_P\log P_{jt}+b_{SIj}SI_t+\varepsilon_{jt})}{\sum_{l=1}^{D}\exp(a_l+\sum_{c=1}^{C}b_cM_{cl,t-4}+b_P\log P_{lt}+b_{SIl}SI_t+\varepsilon_{lt})}$$ |
| | CLR coordinates: |
| | $$\log\left(\frac{S_{jt}}{\widetilde{S}_t}\right)=a_1+\sum_{j'=2}^{D}a'_{j'}d_{j'}+\sum_{c=1}^{C}b_c(M_{cj,t-4}-\overline{M_{c,t-4}})$$ $$+b_P\log\left(\frac{P_{jt}}{\widetilde{P}_t}\right)+b_{SI1}SI_t+\sum_{l=2}^{D}b'_{Il}SI_td_l+(\varepsilon_{jt}-\overline{\varepsilon_t})$$ |
| **DIR** | Estimation by ML using the $S_{jt}$ (common parametrization) |
| | $$\mathbb{E}(S_{jt}\|M_{t-4},P_t,SI_t)=\frac{\exp(a_j+\sum_{c=1}^{C}b_{cj}M_{cj,t-4}+b_{Pj}\log P_{jt}+b_{SIj}SI_t)}{\sum_{l=1}^{D}\exp(a_l+\sum_{c=1}^{C}b_{cl}M_{cl,t-4}+b_{Pl}\log P_{lt}+b_{SIl}SI_t)}$$ |
| | $$\log\alpha_{jt}=a_j+\sum_{c=1}^{C}b_{cj}M_{cj,t-4}+b_{Pj}\log P_{jt}+b_{SIj}SI_t$$ |
| **CODA** | Estimation by OLS on the $(D-1)$ ILR coordinates separately |
| | $$\mathbf{S}_t=\mathbf{a}\bigoplus_{c=1}^{C}\mathbf{B_c}\boxdot\mathbf{M_c}_{t-4}\oplus\mathbf{b_P}\boxdot\log\mathbf{P}_t\oplus\mathbf{b}_{SI}\odot SI_t\oplus\varepsilon_t$$ |
| | $$\Leftrightarrow S_{jt}=\frac{a_j\prod_{l=1}^{D}\prod_{c=1}^{C}M_{cl,t-4}^{b_{cjl}}\log P_{lt}^{b_{Pjl}}b_{SIj}^{SI_t}\varepsilon_{jt}}{\sum_{m=1}^{D}a_m\prod_{l=1}^{D}\prod_{c=1}^{C}M_{cl,t-4}^{b_{cml}}\log P_{lt}^{b_{Pml}}b_{SIm}^{SI_t}\varepsilon_{mt}}$$ |
| | ILR coordinates: |
| | $$S^*_{j't}=a^*_{j'}+\sum_{c=1}^{C}\sum_{l'=1}^{D-1}b^*_{cj'l'}M^*_{cl',t-4}+\sum_{l'=1}^{D-1}b^*_{Pj'l'}(\log P)^*_{l't}$$ $$+b^*_{SIj'}SI_t+\varepsilon^*_{j't}\qquad\forall j'=1,\dots,D-1$$ |

Note that here we want to have an efficient model throughout the studied period, the aim is not to have a good predictive model for the future. Moreover the presented models are not taking into account the time dimension. That explains why we choose an iterative random draw of observations, instead of dividing the sample into a learning sample and a validation sample according to time.

### 1.4.3 Out-of-sample accuracy

The out-of-sample accuracy of the four models is compared according to different indicators adapted to share data that we found in the literature. The considered quality indicators are the following:

- $R^2_T$: the R-squared based on the total variability, widely used in the compositional literature (positive but can be larger that 1; should be larger as possible),
- $R^2_A$: the R-squared based on Aitchison distance, used in Hijazi [28] and Monti et al. [44] (it can be smaller than 0 and larger than 1; should be larger as possible),
- $KL_C$: the compositional Kullback-Leibler divergence defined by Martin-Fernandez et al. [42] (positive; should be lower as possible)

For more information about these quality measures, see the appendix A.1.4.

Table 1.4 presents the averages, over the 100 cross validation trials, of the out-of-sample quality measures for the MNL, GMCI, DIR and CODA models. The best result for each measure is in bold.

Table 1.4: Average out-of-sample quality measures

| | MNL | GMCI | DIR | CODA |
|---|---|---|---|---|
| $R^2_T$ | 0.425 (0.164) | 0.462 (0.179) | 0.622 (0.224) | **0.647** (0.227) |
| $R^2_A$ | 0.196 (0.270) | 0.155 (0.325) | **0.373** (0.235) | 0.084 (0.433) |
| $KL_C$ | 0.139 (0.034) | 0.137 (0.032) | **0.117** (0.071) | 0.134 (0.034) |

Figures in parentheses indicate the standard deviations.

The out-of-sample average quality measures suggest that DIR is the most adapted model to fit our data (27 parameters). However, according to the $R^2$ based on total variability ($R^2_T$), CODA (32 parameters) is better than the Dirichlet model. The GMCI model and the MNL model without cross effects are almost systematically the worst models, certainly due to their simplicity and low number of parameters.

### 1.4.4 Interpretation of parameters

MNL and GMCI models are usually interpreted in terms of direct and cross elasticities (see Cooper and Nakanishi [10], p.34). In Section 2.3.2 in Chapter 2, we adapt this notion to compositional models using the attraction formulations presented in Section 1.3.2. The (direct) elasticity of the share $S_{jt}$ relative to the media $\check{M}_{kj,t-4}$ is equal to $(1 - S_{jt})b_k\check{M}_{kj,t-4}$ in the MNL model, DIR model and MNL specification of the GMCI models, whereas it is equal to $(1 - S_{jt})b_k$ in the MCI specification for the GMCI, and to $b_{kjj} - \sum_{m=1}^{D} S_{mt}b_{kmj}$ for the CODA model.
For example, the direct elasticities of market shares of the three considered brands are computed for the television. As elasticities are time dependent, they are computed for the 123 observed periods, and the average is presented in Table 1.5. These elasticities

Table 1.5: Average direct elasticities for TV investments

| | MNL | GMCI | DIR | CODA |
|---|---|---|---|---|
| DACIA | 0.0019 | 0.0028 | -0.0068 | -0.0046 |
| NISSAN | **0.0101** | **0.0152** | **0.0389** | **-0.0022** |
| RENAULT | 0.0058 | 0.0088 | 0.0145 | -0.0038 |

can be interpreted as the average relative impact on the brand $j$ market share, $S_j$, of a 1% increase of the brand $j$ advertising investments in television.

We observe that elasticities are not the same across models, and can even be of opposite sign. For example, the DIR model concludes that, on average over the period 2005-2015, if Nissan increases its TV investment by 1% in $t-4$, it will increase its market share by 0.04% in $t$, whereas in CODA, it will have a small negative impact. The CODA model, which includes all cross effects, suggests that the impacts of TV investments of Dacia, Nissan and Renault tend to "cancel each other", in the sense that all impacts are very close to zero. However, all models except CODA agree on the fact that Nissan has the highest TV's elasticities (in bold in the table).

## 1.5 Conclusion

Because of the constraints of shares data, classical regression models cannot be used directly to model market shares. Market share models have been developed in the marketing literature, but other models can be adapted to this type of applications.

In this chapter, we present four types of models adapted to model market shares, or share data in general: the multinomial logit model (MNL), the generalized multiplicative competitive interaction model (GMCI), the Dirichlet model (DIR) and the linear compositional model (CODA). We express all of them in attraction form to ease their comparison. We highlight the similarities and the differences of these models from a theoretical point of view. MNL and DIR are generalized linear models estimated by maximum likelihood and centered on the arithmetic mean shares, whereas GMCI and CODA are transformation models estimated by OLS, centered on the geometric mean shares. We prove that GMCI can be written as a particular compositional model, and that it can be considered as a particular case of the CODA model. The CODA model comes out to be similar to the fully extended attraction model used in marketing, but with several advantages: for example, it manages to capture all cross effects with a relative parsimony, thanks to the isometric log-ratio (ILR) transformation involved in the estimation. All these models can be implemented using R, and can be interpreted in terms of elasticities.

We use these models to understand the impact of media investments by channel on brands' market shares in the automobile market, controlling for price and scrapping incentive. We base our model choice on cross-validation using quality measures adapted for shares data. In our application, DIR and CODA models, which are not usually used in this context, outperform the usual market share models, thanks to their higher flexibility. Indeed, MNL and GMCI models are very parsimonious models and they fail to capture the variability of the considered data.

In the following chapters, we are going to focus on the interpretability of the CODA model, especially on the direct and cross effects, and on the relationship between the CODA and the MCI model. As the latter is a particular case of the former, we can imagine that an intermediate specification is possible (see Chapter 2). Concerning our application, it would be more useful in a practical sense to consider an example such that the market shares are the real market shares in the B segment, not the market shares inside a subcomposition of the B segment. Moreover, it would be valuable to consider the carryover effect of advertising, using several lags of media investments or creating a cumulative variable of media investments, such as the so-called adstock variables (see Chapter 3).

# Chapter 2

# Interpretation of market share models

The second chapter of this thesis aims to improve the interpretability of CODA models. In order to do that, we adapt the types of interpretation measures which are used in marketing: the marginal effects, the elasticities and the odds ratios. We also explain how the MCI model can be estimated in a proper manner using the ILR transformation, and how we can combine the CODA and the MCI in order to get an intermediate specification, more parsimonious than the CODA model, what we call the MCODA model. We develop an adapted Fisher test allowing to test whether the unconstrained model (CODA) is better than the constrained models (MCI and MCODA).

This chapter is linked to two working papers which have already been published (see Morais, Thomas-Agnan and Simioni [48] and [47]). A final version of [47] is under submission to the Austrian Journal of Statistics (first version sent on the 19th of July 2017).

## 2.1 Introduction

In the existing literature, we find different types of models to explain shares data, also called compositional data (see Chapter 1 for a comparison). On the one hand, the compositional models, what we call CODA models, come out to be very flexible as they introduce component-specific and cross-effect parameters, resulting in a high complexity. However, their interpretation is not straightforward and little research has been carried out investigating this issue in the dedicated compositional data analysis literature. They are usually interpreted in terms of marginal effects on the transformed shares, which is complicated to use in practice. On the other hand in the marketing literature, market share models, among which the most commonly used is the MCI model[1], are usually interpreted in terms of elasticities. But these models are very simple and the estimation process is questionable.

In this chapter, we combine the best part of each type of models, in order to improve the interpretability of CODA models and the estimation procedure of MCI models. As we prove in Chapter 1 that the MCI model is a particular case of the CODA model, we develop here an intermediate specification, the MCODA model, allowing to have a simple specification for some explanatory variables and a complex specification for others. A model selection procedure is proposed using an adapted Fisher test, considering that the CODA model is the unconstrained model to be compared to the constrained models, the MCI model or the MCODA model.

We propose several types of interpretations directly linked to the shares, in terms of marginal effects, elasticities and odds ratios. We show that marginal effects on shares may not be well adapted to interpret these models because they depend a lot on the considered observation. Elasticities are useful to isolate the impact of an explanatory variable on a particular share as they correspond to the relative variation of a component with respect to the relative variation of an explanatory variable, *ceteris paribus* (in a simplex sense). We show that they can be computed from the transformed model or equivalently from the model in the simplex, and that they are consistent with the simplicial derivatives. Other types of elasticities and odds ratios can be computed for ratios of shares, having the advantage to be observation independent, but they can be complicated to interpret in practice.

The MCI model, the CODA model and the MCODA model are applied to the B segment of the French automobile market from 2003 to 2015. The aim is to explain the brands' market shares of the three leaders: Citroën, Peugeot and Renault, against the group of other brands, using the brands' media investments (all channels taken together), the price and the scrapping incentive. The models are interpreted using marginal effects, elasticities and odds ratios, and they are compared using the Fisher test and in terms of out-of-sample quality measures.

[1]Note that in this chapter we are going to focus on the MCI model which is one of the specifications of the GMCI models presented in Chapter 1 because it is directly a particular case of the CODA model.

This chapter is organized as follows. The second section presents the two types of models (CODA and MCI), their intermediate specification (MCODA), along with the adapted Fisher test for model selection. The third section explains the different ways to interpret them. The fourth section presents the results of the estimation of the models for the application along with interpretations, Fisher tests and quality measures. Finally, the last section concludes on the findings and on further directions to be investigated.

## 2.2 Compositional regression models

### 2.2.1 Two types of compositional models

We prove in Chapter 1 that the MCI model can be written in a compositional way. From now on, we will use the denomination compositional models for the CODA model and the MCI model. They are adapted to model a compositional dependent variable using compositional explanatory variables denoted $\mathbf{X}$ (and potentially classical variables denoted $Z$) as explanatory variables. The difference between the two models is about the specification of the relationship between compositional explanatory and dependent variables: in contrast with the CODA model, the MCI model does not allow for component-specific and cross effect parameters associated to a compositional explanatory variable. There is no difference between MCI and CODA with regard to classical variables: component-specific parameters are specified in both cases.

Table 2.1 reminds the characteristics of the MCI and CODA models. For simplicity, models are presented with a single explanatory random compositional $X$ and a single explanatory real random variable $Z$, but of course several ones can be used like in the examples presented in Section 2.4.

#### The MCI model (without component-specific and cross-effect parameters)

The classical MCI model, as presented in the marketing literature, suffers from an estimation procedure which is questionable. Indeed, as highlighted in Chapter 1, this model is usually estimated by OLS after a CLR transformation, which prevents the error terms to be independently distributed or orthonormal, as required by the OLS estimation. In 2013, Wang et al. [70] present a model which we have demonstrated to be similar to the MCI model in the sense that they have the same attraction formulation (see Section 1.3.5), but they use an ILR transformation for the estimation, which is a proper way to estimate it. Therefore, in this chapter, contrary to Chapter 1, we use the ILR transformation for the MCI model, as well as for the CODA model.

In the MCI model, a compositional explanatory variable is associated to a unique parameter $b \in \mathbb{R}$ (see Table 2.1, equation (2.1)). Thus, cross-effects[2] are not modeled directly, but indirectly through the shares closure. Indeed, we show in Chapter 1 that the MCI model in equation (2.1) can be written in attraction form like in equation (2.3). This equation contains a closure, and we can see that a change of $X_l$ will have an indirect impact on $S_j$ through the denominator. We remind here that equation (2.3) can be expressed either in terms of shares $X_j$ or in terms of volumes $\check{X}_j$ thanks to the scale invariance property, as constants cancel out.

If a classical explanatory variable $Z$ is used in the MCI model, it is associated to a composition of parameters $\mathbf{c}$. It can be surprising to see that in the attraction form of the MCI model, the variable $Z$ is powering the intercept $c_j$, but this corresponds to the term $Z_t \odot \mathbf{c}$. The intercept of the MCI model is also a composition, $\mathbf{a}$.

[2]We denote by cross-effect the effect of a variation of $X_l$ on $S_j$, where $l \neq j$.

The ILR transformation of the MCI model is presented in equation (2.5). Assuming that the transformed error terms are normal (implying that the non-transformed compositional error terms are "normal in the simplex"), we can use OLS to estimate the model.

An important feature of the MCI model is that compositional explanatory variables $\mathbf{X}$ have to be of the same dimension that the compositional dependent variable $\mathbf{S}$, such that $\mathbf{S}, \mathbf{X} \in \mathcal{S}^D$. This model is adapted when compositions $\mathbf{X}$ and $\mathbf{S}$ refer to variables associated to the same components in the same order, for example $\mathbf{S}$ can be the composition of brands' market shares and $\mathbf{X}$ the composition of brand media investments (where brands are in the same order in $\mathbf{S}$ and $\mathbf{X}$) (see Section 2.4), or $\mathbf{S}$ can be the composition of GDP from three sectors and $\mathbf{X}$ the composition of labor force of these three sectors. Otherwise, this model makes no sense. Then, equation (2.5) is estimated using $(D-1) \times T$ observations (the number of ILR coordinates $D-1$ times the number of observations $T$).

**The CODA model (with component-specific and cross-effect parameters)**

The CODA model where dependent and explanatory variables are compositional is used by Van Den Boogaart and Tolosana-Delgado [68] in a toy example, and by Chen et al. [8] in an article, but to our knowledge it has not been applied elsewhere. Using exactly the same dependent and explanatory variables as the MCI model (see equation (2.2)), it allows each component $X_l$ of $\mathbf{X}$ to have a specific impact on each component $S_j$ of $\mathbf{S}$. This is particularly visible in the attraction form of the CODA model (equation (2.4)): instead of having a unique parameter $b \in \mathbb{R}$ associated to $\mathbf{X}$, we have a matrix of parameters $\mathbf{B} \in \mathbb{R}_{D_S, D_X}$. If the dimensions of the dependent composition $\mathbf{S}$ and of the explanatory composition $\mathbf{X}$ are equal ($D_S = D_X$) and if they refer to the same components in the same order, then $\mathbf{B}$ is a square matrix with direct effects on the diagonal and cross-effects outside of the diagonal. There is no difference between the MCI model and the CODA model for the specification of the intercept and classical explanatory variables. The same remark as for the MCI model can be done concerning the attraction form of the CODA model: equation (2.4) can be expressed either in terms of shares $X_j$ or in terms of volumes $\breve{X}_j$ thanks to the closure operation.

As in the MCI model, in order to estimate the CODA model, we transform it using the ILR transformation (see equation (2.6)). But here, $D_S - 1$ equations are estimated separately (one for each coordinate of $\mathbf{S}$) with $T$ observations each. The complexity of the CODA model is reflected by a large number of parameters. This can be an issue if the number of observations $T$ is too small.

Note that in the CODA model, $\mathbf{X} \in \mathcal{S}^{D_X}$ and $\mathbf{S} \in \mathcal{S}^{D_S}$ may have different dimensions. For example, $\mathbf{S}$ can be the composition of GDP from three sectors and $\mathbf{X}$ the composition of labor force for six occupation categories. In our application, $D_S = D_X$: $\mathbf{S}$ is the composition of brands' market shares and $\mathbf{X}$ is the composition of brand media investments (see Section 2.4).

Table 2.1: Two kinds of models for compositional dependent and explanatory variables

| | MCI | CODA |
|---|---|---|
| In compositions | $\mathbf{S}_t = \mathbf{a} \oplus b \odot \mathbf{X}_t \oplus Z_t \odot \mathbf{c} \oplus \epsilon$ (2.1) | $\mathbf{S}_t = \mathbf{a} \oplus \mathbf{B} \boxdot \mathbf{X}_t \oplus Z_t \odot \mathbf{c} \oplus \epsilon$ (2.2) |
| In attraction form | $S_{jt} = \frac{a_j X_{jt}^b c_j^{Z_t} \epsilon_{jt}}{\sum_{m=1}^D a_m X_{mt}^b c_m^{Z_t} \epsilon_{mt}}$ (2.3) | $S_{jt} = \frac{a_j \prod_{l=1}^D X_{lt}^{b_{jl}} c_j^{Z_t} \epsilon_{jt}}{\sum_{m=1}^D a_m \prod_{l=1}^D X_{lt}^{b_{ml}} c_m^{Z_t} \epsilon_{mt}}$ (2.4) |
| In coordinates | $\mathbf{S}_t^* = \mathbf{a}^* + \mathbf{X}_t^* b + \mathbf{c}^* Z_t + \epsilon_t^*$ (2.5) | $\mathbf{S}_t^* = \mathbf{a}^* + \mathbf{X}_t^* \mathbf{B}_\mathbf{k}^* + \mathbf{c}^* Z_t + \epsilon_t^*$ (2.6) |
| | with $\epsilon_{jt}^* \sim \mathcal{N}(0, \sigma^2) \quad \forall j, \forall t$ | with $\epsilon_{jt}^* \sim \mathcal{N}(0, \sigma_j^2) \quad \forall t$ |
| Component-specific parameters for $X$ | No | Yes |
| Cross-effects for $X$ | No | Yes |
| Dimension | $D$ for $\mathbf{S}$ and $\mathbf{X}$ | $D_S$ for $\mathbf{S}$ ; $D_X$ for $\mathbf{X}$ |
| Nb. parameters | $(D-1)(1+K_Z)+K_X$ | $(D_S-1)(1+K_Z+\sum_{k=1}^{K_X}(D_k-1))$ |

$\mathbf{X}_t$: compositional explanatory variable; $Z_t$: classical explanatory variable.
$D_S$: number of components of $\mathbf{S}$; $D_X$ or $D_k$: number of components of $\mathbf{X}_\mathbf{k}$.
$\mathbf{S}, \mathbf{a}, \mathbf{b}, \mathbf{X}, \epsilon \in \mathcal{S}^D$; $b, X \in \mathbb{R}$; $\mathbf{B} \in \mathbb{R}^{D_S \times D_X}$; $\mathbf{S}^*, \mathbf{a}^*, \mathbf{b}^*, \mathbf{B}^*, \mathbf{X}^*, \epsilon^*$: ILR coordinates.
$\epsilon$: normal in the simplex distributed error terms ; $\epsilon^*$: normal distributed error terms.
$K_X$ and $K_Z$: number of compositional and classical explanatory variables ($K_X = K_Z = 1$ in the table).
$\mathbb{E}^\oplus$: expected value in the simplex.

### 2.2.2 Intermediate specification (MCODA model) and model selection

We prove in Chapter 1 that the MCI model is a particular case of the CODA model. Then, in a given model it is possible to mix the two specifications, if and only if $D_S = D_X$ for the explanatory variable with the MCI specification. This model, what we call MCODA model, is defined as follows in the simplex:

$$\mathbf{S}_t = \mathbf{a} \oplus \ \beta \odot \mathbf{X}_t \ \oplus \ \mathbf{B} \boxdot \boldsymbol{\chi}_t \ \oplus \ Z_t \odot \mathbf{c} \ \oplus \epsilon, \tag{2.7}$$

where $\mathbf{X}$ are the compositional explanatory variables with a MCI specification and $\boldsymbol{\chi}$ are those with a CODA specification.

The MCODA model can be estimated by OLS using its expression in ILR coordinates:

$$S_{jt}^* = \sum_{d=1}^{D-1} \mathbb{1}_{d=j} a_d^* + \beta X_{jt}^* + \sum_{l=1}^{D-1} \sum_{d=1}^{D-1} \mathbb{1}_{d=j} b_{dl}^* \chi_{lt}^* + \sum_{d=1}^{D-1} \mathbb{1}_{d=j} c_d^* Z_t + \epsilon_{jt}^* \tag{2.8}$$

with $\mathbb{1}_{d=j} = 1$ if $d = j$ and 0 otherwise, and $\epsilon_{jt}^* \sim \mathcal{N}(0, \sigma^2) \quad \forall j = 1, \ldots, D-1, \forall t = 1, \ldots, T$. However, the induced constant variance of transformed error terms across coordinates, as in the MCI model, is questionable.

Note that the CODA model can also be estimated using dummy variables as in equation (2.8) leading to the same estimated coefficients but not to the same standard errors

than in equation (2.6) because of the assumption on error terms.

As the MCI model and the MCODA model are constrained versions of the CODA model, a model selection can be done using a Fisher test. We consider to test the following null hypothesis : $H_0 : b^*_{j,j} = b^*, \forall j$ and $b^*_{j,l} = 0, \forall j \neq l$. The associated test statistic is:

$$\begin{aligned} F &= \frac{SSE_0 - SSE_1}{SSE_1} \times \frac{N-K}{p} \quad \sim \mathcal{F}(p, N-K) \text{ under } H_0 \\ &= \frac{SSE_0 - SSE_1}{SSE_1} \times \frac{(D-1)[T - K(D-1) - K_Z - 1]}{p} \end{aligned}$$

with $SSE_0$ and $SSE_1$ the sum of squared errors of the constrained and unconstrained models, $T$ the number of observations, $K$ the number of compositional explanatory variables, $K_Z$ the number of classical explanatory variables, and $p$ the number of constraints.

## 2.3 Interpretation of compositional models

As the estimation of compositional models is performed in the coordinate space, the interpretation of the fitted parameters is difficult because parameters are linked to the log-ratio transformation of shares, not directly to the shares. It is possible to derive the coefficients in the simplex associated to shares using the inverse transformation, but their interpretation is not straightforward either.

We are going to show that relative impacts, like elasticities or odds ratios, are more natural than marginal effects, to interpret impacts on shares in compositional model (as is the case in the classical logistic model).

Table 2.2 provides a summary of the different types of interpretation of compositional and classical explanatory variables' impacts, in the MCI and in the CODA models, which are detailed below. Note that it is not possible to measure the impact of the $X_{lt}$'s share, but only of the corresponding volume of $\check{X}_{lt}$. Indeed, a share cannot increase *ceteris paribus* because it implies a change in other shares. However, we can consider a change in the volume of $\check{X}_{lt}$, with all other volumes $\check{X}_{mt}, \forall m \neq l$ fixed.

Table 2.2: Impact assessment measures for compositional model

| Var | Measure | Effect | MCI | CODA |
|---|---|---|---|---|
| $X$ | $me(S_{jt}, \check{X}_{lt})$ | Direct | $b(1-S_{jt})\frac{S_{jt}}{\check{X}_{lt}}$ | $(b_{jl} - \sum_{m=1}^{D} S_{mt} b_{ml})\frac{S_{jt}}{\check{X}_{lt}}$ |
| | | Indirect | $(-bS_{lt})\frac{S_{jt}}{\check{X}_{lt}}$ | |
| | $ME(\mathbf{S}_t, \check{\mathbf{X}}_t)$ | Matrix | $[\mathbf{S_{jt}}] \odot \mathbf{W_t} b \odot [\mathbf{1}/\check{\mathbf{X}}_{\mathbf{lt}}]$ | $[\mathbf{S_{jt}}] \odot \mathbf{W_t B} \odot [\mathbf{1}/\check{\mathbf{X}}_{\mathbf{lt}}]$ |
| | $e(S_{jt}, \check{X}_{lt})$ | Direct | $b(1-S_{jt})$ | $(b_{jl} - \sum_{m=1}^{D} S_{mt} b_{ml})$ |
| | | Indirect | $-bS_{lt}$ | |
| | $E(\mathbf{S}_t, \check{\mathbf{X}}_t)$ | Matrix | $\mathbf{W_t} b$ | $\mathbf{W_t B}$ |
| | $e\left(\frac{S_{jt}}{S_{j't}}, \check{X}_{lt}\right)$ | Direct | $b$ | $(b_{jl} - b_{j'l})$ |
| | | Indirect | 0 | |
| | $\mathrm{OR}\left(\frac{S_{jt}}{S_{j't}}, \check{X}_{lt}, \Delta\right)$ | Direct | $(1+\Delta)^b$ | $(1+\Delta)^{(b_{jl}-b_{j'l})}$ |
| | | Indirect | 0 | |
| | $e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{X}_{lt}\right)$ | Direct | $b$ | $b_{11}^{*(j,l)}\sqrt{\frac{D_X-1}{D_X}}/\sqrt{\frac{D_S-1}{D_S}}$ |
| | | Indirect | 0 | |
| $Z$ | $me(S_{jt}, Z_t)$ | | $(\log c_j - \sum_{m=1}^{D} S_{mt} \log c_m) S_{jt}$ | |
| | $ME(\mathbf{S}_t, Z_t)$ | Vector | $[\mathbf{S_{jt}}] \odot \mathbf{W}_t \log \mathbf{c}$ | |
| | $e(S_{jt}, Z_t)$ | | $(\log c_j - \sum_{m=1}^{D} S_{mt} \log c_m) Z_t$ | |
| | $E(\mathbf{S}_t, Z_t)$ | Vector | $\mathbf{W}_t \log \mathbf{c} Z_t$ | |
| | $e(\frac{S_{jt}}{S_{j't}}, Z_t)$ | | $\log(c_j/c_{j'}) Z_t$ | |
| | $\mathrm{OR}\left(\frac{S_{jt}}{S_{j't}}, Z_t, \Delta\right)$ | | $(c_j/c_{j'})^{\Delta Z_t}$ | |

In this table, $\mathbb{E}^{\oplus} S_{jt}$ is denoted by $S_{jt}$ to shorten notations, and $\odot$ denotes the Hadamard product. Moreover, these measures are estimated using observed shares $S_{jt}$ in practice, not fitted shares.
Direct effect when $l = j$; indirect effect when $l \neq j$.
$\mathbf{W_t}$ contains $1 - S_{it}$ on the diagonal and $-S_{it}$ otherwise.

### 2.3.1 Marginal effect of a component

In classical linear models, coefficients are usually interpreted in terms of marginal effects: if the explanatory variable increases by one unit, then the dependent variable increases by the value of the coefficient. In the case of compositional models, we prove in this chapter that it is possible to compute marginal effects, but it is not straightforward. The marginal effect of the component $\check{X}_{lt}$ (in volume) on the dependent share $S_{jt}$ is

$$me(\mathbb{E}^{\oplus}S_{jt}, \check{X}_{lt}) = \frac{\partial \mathbb{E}^{\oplus}S_{jt}}{\partial \check{X}_{lt}},$$

where $\mathbb{E}^{\oplus}S_{jt}$ is the "expected value in the simplex" of $S_{jt}$ (see Chapter 1, Section 1.3.2), such that $\mathbb{E}^{\oplus}S_{jt} = \frac{a_j X_{jt}^b c_j^{Z_t}}{\sum_{m=1}^{D} a_m X_{jt}^b c_m^{Z_t}}$ for the MCI model and $\mathbb{E}^{\oplus}S_{jt} = \frac{a_j \prod_{l=1}^{D} X_{lt}^{b_{jl}} c_j^{Z_t}}{\sum_{m=1}^{D} a_m \prod_{l=1}^{D} X_{lt}^{b_{ml}} c_m^{Z_t}}$ for the CODA model.

For the CODA model, we show that marginal effects are equal to

$$\begin{aligned} me(\mathbb{E}^{\oplus}S_{jt}, \check{X}_{lt}) &= \frac{\partial \mathbb{E}^{\oplus}S_{jt}}{\partial \log \mathbb{E}^{\oplus}S_{jt}} \frac{\partial \log \mathbb{E}^{\oplus}S_{jt}}{\partial \log \check{X}_{lt}} \frac{\partial \log \check{X}_{lt}}{\partial \check{X}_{lt}} \\ &= \left( b_{jl} - \sum_{m=1}^{D} S_{mt} b_{ml} \right) \frac{\mathbb{E}^{\oplus}S_{jt}}{\check{X}_{lt}} \end{aligned} \tag{2.9}$$

Let us define $ME(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t)$ the $D_S \times D_X$ matrix containing all marginal effects. It can be computed as follows:

$$ME(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t) = \left[\mathbf{S_{jt}}\right] \odot \mathbf{W}_t \mathbf{B} \odot \left[\frac{\mathbf{1}}{\check{\mathbf{X}}_{\mathbf{lt}}}\right] = \left[\mathbf{S_{jt}}\right] \odot \mathbf{W}_t \mathbf{V} \mathbf{B}^* \mathbf{V}' \odot \left[\frac{\mathbf{1}}{\check{\mathbf{X}}_{\mathbf{lt}}}\right],$$

where $\odot$ denotes here the Hadamard product (term by term product)[3], $\left[\mathbf{S_{jt}}\right]$ is a $D_S \times D_S$ matrix with $S_{jt}$ on the $j^{\text{th}}$ row, $\left[\frac{\mathbf{1}}{\check{\mathbf{X}}_{\mathbf{lt}}}\right]$ is a $D_X \times D_X$ matrix with $1/\check{X}_{lt}$ on the $l^{\text{th}}$ column, $\mathbf{B}^*$ and $\mathbf{B}$ denote respectively the parameters in the transformed space and in the simplex, and $\mathbf{W}_t$ is a $D_S \times D_S$ matrix composed of diagonal terms equal to $1 - \mathbb{E}^{\oplus}S_j$ and non-diagonal terms in column $j$ equal to $-\mathbb{E}^{\oplus}S_j$. Similar results can be found for the MCI model in Table 2.2, where $\mathbf{B}$ is replaced by $b$.

This marginal effect matrix can also be computed using ILR coordinates and Jacobian matrices instead of using the attraction form of the model (see detail in the appendix A.2.1).

### 2.3.2 Elasticity of a dependent share relative to a component

The marginal effect $me(\mathbb{E}^{\oplus}S_{jt}, \check{X}_{lt})$ depends on all shares $S_{mt}$ and on volumes $\check{X}_{lt}$. We can see in our application that it can vary a lot across observations (see Section 2.4.3),

[3]Note that $\odot$ in bold denotes the Hadamard product whereas $\odot$ denotes the power transformation. It is in fact a perturbation of the matrices involved considered as compositions in a larger space.

and therefore it is not a good measure to summarize the impact of a component $\check{X}_{lt}$ on a share $S_{jt}$. We are going to show that elasticities are more natural to interpret compositional models.

The first elasticity we may want to compute is the elasticity of the share $S_{jt}$ relative to the volume of $\check{X}_{lt}$. It corresponds to the relative variation of $S_{jt}$ induced by a relative variation of 1% of the volume $\check{X}_{lt}$ (keeping all other volumes constant) or alternatively a relative variation of 1% of the share $X_{lt}$ (holding constant the ratios $\frac{X_{it}}{X_{jt}}$ of the remaining components):

$$e_{jlt} = e(\mathbb{E}^{\oplus} S_{jt}, \check{X}_{lt}) = \frac{\frac{\partial \mathbb{E}^{\oplus} S_{jt}}{\mathbb{E}^{\oplus} S_{jt}}}{\frac{\partial \check{X}_{lt}}{\check{X}_{lt}}} = \frac{\partial \log \mathbb{E}^{\oplus} S_{jt}}{\partial \log \check{X}_{lt}} \tag{2.10}$$

Since both variables (dependent and independent) are compositions, we should consider the notion of derivative of a simplex-valued function with respect to a compositional variable. Egozcue et al. (in Pawlowsky-Glahn and Buccianti [54], chapter 12) treat the case of the derivative of a simplex-valued function with respect to a real variable, and Barcelo-Vidal et al. (in Pawlowsky-Glahn and Buccianti [54], chapter 13) the case of the derivative of a vector-valued function with respect to a compositional variable. Combining the two notions, let us denote by $\partial^{\oplus}\mathbf{h}/\partial^{\oplus} X_l$ the directional simplicial derivatives of a function $\mathbf{h}$ from the simplex $\mathcal{S}^{D_X}$ of $\mathbb{R}^{D_X}$ to the simplex $\mathcal{S}^{D_S}$ of $\mathbb{R}^{D_S}$. Using a result linking the directional simplicial derivatives of the function $\mathbf{h}$ of shares with the semi-log derivatives of the corresponding function of volumes (see appendix A.2.2), we can then derive the relationship between the directional simplicial derivatives of the composition $\mathbf{S}_t$ with respect to the shares $X_{lt}$ and the above elasticities as follows:

$$e^{\oplus}_{lt} = \frac{\partial^{\oplus} \mathbb{E}^{\oplus} \mathbf{S}_t}{\partial^{\oplus} X_{lt}} = \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^{\oplus} \mathbf{S}_t}{\partial \log \check{X}_{lt}}\right)\right)' = \mathcal{C}\left(\exp(e_{1lt}), \ldots, \exp(e_{Dlt})\right)'$$

We call $e^{\oplus}_{lt}$ the simplicial elasticity of $\mathbf{S}_t$ relative to $X_{lt}$. The elasticities $e_{jlt}$ from equation (2.10) are easy to compute from the attraction form of $\mathbb{E}^{\oplus} S_{jt}$, in a similar way than marginal effects (see equation (2.9)). They can also be expressed in a matrix form $E(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t)$, as can be seen in Table 2.2. The relationship between marginal effects and elasticities is the following:

$$ME(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t) = \left[\mathbf{S_{jt}}\right] \odot E(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t) \odot \left[\mathbf{1}/\check{\mathbf{X}}_{\mathbf{lt}}\right],$$

where $\odot$ denotes the Hadamard product.

These elasticities allow to isolate the impact of the $\check{\boldsymbol{X}}$ component on the $\mathbf{S}$ component which is very useful. This impact is understood as the impact of a relative variation of the volume (resp: the share) keeping all other volumes constant (resp: holding constant the ratios of the remaining components). Compared to marginal effects, the $e(\mathbb{E}^{\oplus} S_{jt}, \check{X}_{lt})$ still depend on observations through the shares $S_{mt}$, but do not depend anymore on the volumes $\check{X}_{lt}$. Then, if shares are not varying too much, as it is the case in our example (see Section 2.4.3), they can be a good measure of impact.

As for marginal effects, the elasticity matrix can also be computed from ILR coordinates (see detail in the appendix A.2.1).

Let us now consider making a first order Taylor approximation of the vector of shares for a small relative change in volume $X_l$. For a small $\delta = \frac{\Delta \check{X}_{lt}}{\check{X}_{lt}}$, we could write this first approximation of the share:

$$S'_{jt} = S_{jt}(1 + \delta e_{jlt}) \tag{2.11}$$

and it is easy to see that $\mathbf{S}'_t = (S'_{1t}, \ldots, S'_{Dt})'$ does belong to the simplex. Indeed, they are summing up to one because $\sum_{m=1}^{D} \mathbb{E}^{\oplus} S_{mt} e_{jlt} = 0$ and $\sum_{m=1}^{D} \mathbb{E}^{\oplus} S_{mt} = 1$ (see the proof in the appendix A.2.3).

Another first order Taylor approximation of the vector of shares (see equation (12.13), p.168, in Pawlowsky-Glahn and Buccianti[54]) denoted $\mathbf{S}''_t$ is

$$\mathbf{S}''_t = \mathbf{S}_t \oplus \delta \odot e^{\oplus}_{lt} = \mathcal{C}\left(S_{1t}\exp(\delta e_{1lt}), \ldots, S_{Dt}\exp(\delta e_{Dlt})\right)' \tag{2.12}$$

Note that when $\delta \to 0$, since $\exp(\delta e_{jlt}) \simeq 1 + \delta e_{jlt}$, these two approximations are equivalent at the first order, although the first one is simpler as it does not use the Aitchison geometry:

$$\mathbf{S}''_t \simeq \mathcal{C}\left(S_{1t}(1 + \delta e_{1lt}), \ldots, S_{Dt}(1 + \delta e_{Dlt})\right)' = \mathcal{C}\left(S'_{1t}, \ldots, S'_{Dt}\right)' = \left(S'_{1t}, \ldots, S'_{Dt}\right)' \tag{2.13}$$

### 2.3.3 Elasticity and odds ratio of a ratio of dependent shares relative to a component

In order to avoid being observation dependent, other measures can be computed for interpreting the MCI and the CODA models. However, they are concerning ratios of shares, not directly a single share. Then, they can be complicated to interpret in practice.

**Elasticity of a ratio of dependent shares** As compositional models are based on a log-ratio approach, elasticities of ratios are easy to compute. We can be interested in the elasticity of a ratio of shares (or volumes) $\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}$ relative to an infinitesimal change in the volume of $\check{X}_{lt}$ for example:

$$e\left(\frac{\mathbb{E}^{\oplus} S_{jt}}{\mathbb{E}^{\oplus} S_{j't}}, \check{X}_{lt}\right) = \frac{\partial \log(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't})}{\partial \log \check{X}_{lt}}$$

We see in Table 2.2 that the result for both compositional models is constant across observations because it only depends on parameters. Note here that the MCI model respects the IIA (independence from irrelevant alternatives) property, meaning that the ratio of two shares $\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}$ only depends on the corresponding components $j$ and $j'$ of $\check{\mathbf{X}}$. Then, $e(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}) = 0$ if $l \neq j, j'$. Moreover, the elasticity of the ratio between the share $j$ and the share $j'$ relative to a change in $\check{X}_{jt}$ is the same for all considered shares $j'$. This is a lack of flexibility of the MCI model, because it implies that an increase of $\check{X}_{jt}$ will reduce proportionally all other shares. The CODA model does not satisfy the IIA property, and then this model is able to take into account possible synergies (positive cross effects) between brands.

**Odds ratio of a ratio of dependent shares** Another type of interpretation which can be used for shares is the odds ratio. The advantage of this measure is that it is a measure of impact of a discrete change of $\check{X}_l$ ($\check{X}_l$ is increased by $\Delta \times 100\%$ between situations $t = t_1$ and $t = t_2$) on the ratio $\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}$, as opposed to an infinitesimal change for marginal effects and elasticities. The odds ratio for a couple of shares $\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}$ relative to $\check{X}_{lt}$ is

$$OR\left(\frac{\mathbb{E}^{\oplus} S_{jt}}{\mathbb{E}^{\oplus} S_{j't}}, \check{X}_{lt}, \Delta\right) = \frac{(\mathbb{E}^{\oplus} S_{j,t_2} / \mathbb{E}^{\oplus} S_{j',t_2}) | \check{X}_{l,t_2}}{(\mathbb{E}^{\oplus} S_{j,t_1} / \mathbb{E}^{\oplus} S_{j',t_1}) | \check{X}_{l,t_1}},$$

where $\check{X}_{l,t_2} = (1 + \Delta)\check{X}_{l,t_1}$ and $\Delta \geq 0$.

Note that $e(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt})$ and $OR(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}, \Delta)$ are more or less measuring the same thing differently, if $\Delta$ is small:

$$\begin{aligned} e(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}) &\simeq \frac{(\mathbb{E}^{\oplus} S_{jt_2} / \mathbb{E}^{\oplus} S_{j't_2}) - (\mathbb{E}^{\oplus} S_{jt_1} / \mathbb{E}^{\oplus} S_{j't_1})}{(\mathbb{E}^{\oplus} S_{jt_1} / \mathbb{E}^{\oplus} S_{j't_1})} / \frac{\check{X}_{lt_2} - \check{X}_{lt_1}}{\check{X}_{lt_1}} \\ &\simeq \frac{OR(\mathbb{E}^{\oplus} S_{jt} / \mathbb{E}^{\oplus} S_{j't}, \check{X}_{lt}, \Delta) - 1}{(\check{X}_{lt_2} - \check{X}_{lt_1}) / (\check{X}_{lt_1})} \end{aligned}$$

### 2.3.4 Elasticity of a particular ratio of dependent shares relative to a particular ratio of components

In the compositional data analysis literature, compositional models are interpreted with marginal effects directly on ILR coordinates, which corresponds to interpret marginal effects on particular log ratios of shares. Thus, it is advised to choose an appropriate ILR transformation in order to have ILR coordinates which make sense for the considered application, using a sequential binary partition for example (see Hron et al. [30]). We show here that we can go one step further and make an interpretation in terms of elasticity for the ratio of shares directly.

For example, in Chen et al. [8], the chosen ILR transformation is the following:

$$ilr(\mathbf{x})_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{x_i}{(\prod_{j=1+i}^{D} x_j)^{1/(D-i)}}, \quad i = 1, \ldots, D-1$$

With this transformation, the expected value of the first coordinate of $\mathbf{S}$ according to the MCI model is equal to

$$\mathbb{E} ilr(\mathbf{S})_1 = \sqrt{\frac{D-1}{D}} \log \frac{\mathbb{E}^{\oplus} S_{1t}}{g(\mathbb{E}^{\oplus} S_{-1t})} = a_1^* + b^* \sqrt{\frac{D-1}{D}} \log \frac{\check{X}_{1t}}{g(\check{X}_{-1t})} + c_1^* Z_t,$$

and the expected value of the first coordinate of $\mathbf{S}$ according to the CODA model is

equal to

$$\mathbb{E}ilr(\mathbf{S})_1 = \sqrt{\frac{D_S-1}{D_S}}\log\frac{\mathbb{E}^{\oplus}S_{1t}}{g(\mathbb{E}^{\oplus}S_{-1t})} = a_1^{*(j,l)} + b_{11}^{*(j,l)}\sqrt{\frac{D_X-1}{D_X}}\log\frac{\check{X}_{1t}}{g(\check{X}_{-1t})} + b_{12}^{*(j,l)}\sqrt{\frac{D_X-2}{D_X-1}}\log\frac{\check{X}_{2t}}{g(\check{X}_{-1-2t})} + \dots,$$

where the indexes $j$ and $l$ denote the fact that the $j^{\text{th}}$ component of $\mathbf{S}$ and the $l^{\text{th}}$ component of $\mathbf{X}$ are in the first position in $\mathbf{S}$ and $\mathbf{X}$ before the ILR transformation.

In order to interpret their model, Chen et al. [8] (who are considering a CODA model) compute the marginal effect of $ilr(\mathbf{X})_1^{(l)}$ on $ilr(\mathbf{S})_1^{(j)}$:

$$me(\mathbb{E}ilr(\mathbf{S})_1^{(j)}, ilr(\mathbf{X})_1^{(l)}) = \frac{\partial\sqrt{\frac{D_S-1}{D_S}}\log(\mathbb{E}^{\oplus}S_{jt}/g(\mathbb{E}^{\oplus}S_{-jt}))}{\partial\sqrt{\frac{D_X-1}{D_X}}\log(\check{X}_{lt}/g(\check{X}_{-lt}))} = b_{11}^{*(j,l)},$$

so that an increase of one unit of $ilr(\mathbf{X})_1^{(l)}$ implies an increase of $b_{11}^{*(j,l)}$ units of $\mathbb{E}ilr(\mathbf{S})_1^{(j)}$. Note that this is true if and only if $ilr(\mathbf{X})_1^{(l)} = \sqrt{\frac{D_X-1}{D_X}}\log(\check{X}_{lt}/g(\check{X}_{-lt}))$ moves because $\check{X}_{lt}$ moves, while other $\check{X}_{jt}$ remain constant. Otherwise, other ILR coordinates in the right part of the equation would be moving and the marginal effect should take it into account. However, for the MCI model, we do not have this problem because other ILR coordinates of $\mathbf{X}$ are not used.

We show that this is equivalent to compute the following elasticity (multiplying by a factor if $D_S \neq D_X$):

$$e\left(\frac{\mathbb{E}^{\oplus}S_{jt}}{g(\mathbb{E}^{\oplus}S_{-jt})}, \check{X}_{lt}\right) = \frac{\partial\log(\mathbb{E}^{\oplus}S_{jt}/g(\mathbb{E}^{\oplus}S_{-jt}))}{\partial\log\check{X}_{lt}} = \begin{cases} b_{11}^{*(j,l)} & \text{if } D_S = D_X \\ \sqrt{\frac{(D_X-1)/D_X}{(D_S-1)/D_S}}\, b_{11}^{*(j,l)} & \text{otherwise} \end{cases}$$

Thus, instead of saying that when $ilr(\mathbf{X})_1^{(l)}$ increases by 1 unit, $\mathbb{E}ilr(\mathbf{S})_1^{(j)}$ increases by $b_{11}^{*(j,l)}$ units, one can say that when $\check{X}_{lt}$ increases by 1%, $\mathbb{E}^{\oplus}S_{jt}/g(\mathbb{E}^{\oplus}S_{-jt})$ increases by $b_{11}^{*(j,l)}$% (in the case where $D_S = D_X$). The term $\mathbb{E}^{\oplus}S_{jt}/g(\mathbb{E}^{\oplus}S_{-jt})$ can be interpreted as the "predominance of share $j$ over the average of other shares".

Note that this $b_{11}^{*(j,l)}$ will be different for each permutation (i.e. each couple $j, l$). Chen et al. [8] show how one can determine in one step the first coefficient of $B^{*(j,l)}$, that is the $b_{11}^{*(j,l)}$ which is used to compute the above elasticity, for all possible permutations without fitting several times the model.

### 2.3.5 Elasticities and odds ratios relative to a classical variable

The same kind of interpretations can be done for classical variables $Z$, as presented in Table 2.2, except for the elasticity including the geometric mean. Indeed, this would allow to measure the marginal effect (not the elasticity) of $Z_t$ over $\mathbb{E}ilr(\mathbf{S})_1 = \sqrt{\frac{D_S-1}{D_S}} \log \frac{\mathbb{E}^{\oplus} S_{1t}}{g(\mathbb{E}^{\oplus} S_{-1t})}$. This marginal effect would be equal to $c_1^*$ for the MCI model and the CODA model, but this kind of interpretation is not useful to understand the impact of $Z$ on the final shares. Thus, we do not show this measure in Table 2.2.

Note that in practice, elasticities and other measures depending on $\mathbb{E}^{\oplus} S_{jt}$ are estimated using the observed shares $S_{jt}$, not the fitted shares $\widehat{S_{jt}}$.

## 2.4 Application

As explained in the introduction of this thesis, the automobile market is usually segmented in 5 segments, from A to E, according to the size of the vehicle chassis. Inside a segment, a brand generally supplies only one main vehicle. Thus, we can consider that the alternatives for a consumer inside a particular segment coincide with the available brands in this segment. One can suppose that consumers intending to buy new cars make their choice partly in accordance with the price and the "image" of the brand. Car manufacturers spend millions of euros in media investments to enhance their image, giving rise to the following question: do the media investments have an impact on brands' market shares?

In this chapter, we focus on the B segment of the French automobile market, and more precisely on the three leaders Citroën, Peugeot and Renault (see Figure 2.1). Other brands are aggregated in a group called Others (also denoted "ZZZ"). In order to answer the previous question, we model the corresponding market shares as a function of brands' total media investments, brands' average catalogue prices and the scrapping incentive dummy variable.

The considered media investments are the sum of the investments made by a brand for its vehicle of the B segment, in television, radio, press, outdoor, internet and cinema, in euro (see Figure 2.1). They do not include advertising budget for the brand itself. In order to take into account the carryover effect of the advertising, we use the media investments at time $t-3, t-4$ and $t-5$ before the registration time.

The brand's average catalogue price (average of catalogue prices weighted by corresponding sales at the vehicle level) is also used as an explanatory variable (see Figure 2.1). It does not include potential promotions made in the car dealership at the time of purchase. Even if they do not vary a lot across time, prices are used to position brands within the segment.

We also control for scrapping incentive periods, from December 2008 to December 2010. The corresponding dummy variable is a "classical" variable (not compositional) which varies across time only, not across brands.

In the case of this study, the MCI model considers that the effect of media investments and price are the same for all brands, whereas the CODA model implies cross-effects and brand-specific impacts of media investments and price on market shares. As our interest is on the impact of media investments, we also consider a MCODA model which contains cross-effects and brand-specific parameters for media investments, but a unique parameter for all brands for the composition of prices.

This section presents the results of this application. We interpret the models in terms of marginal effects, elasticities and odds ratios of shares, and we compare them in terms of goodness-of-fit measures. The Fisher tests comparing the unconstrained CODA model to the constrained MCI and MCODA models are also computed.

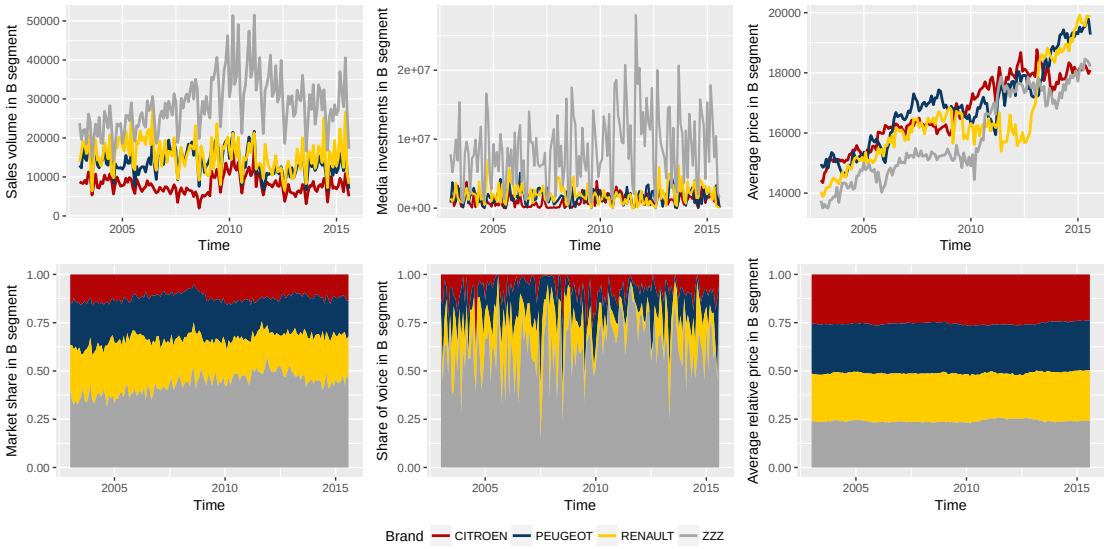


Figure 2.1: Sales, media and average price of brands, in volume and in share, in the B segment

### 2.4.1 Non brand-specific impact of media investments (MCI model)

**Model** Assuming that brand media investments and brand prices have the same effect for all brands, the following equations correspond to the model in the simplex and the attraction formulation of the model:

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{\tau=3}^{5} b_\tau \odot \mathbf{M}_{t-\tau} \oplus b_P \odot \mathbf{P}_t \oplus SI_t \odot \mathbf{c} \oplus \boldsymbol{\varepsilon}_t$$

$$\Leftrightarrow S_{jt} = \frac{a_j \prod_{\tau=3}^{5} M_{t-\tau,j}^{b_\tau} P_{t,j}^{b_P} c_j^{SI} \varepsilon_{jt}}{\sum_{m=1}^{4} a_m \prod_{\tau=3}^{5} M_{t-\tau,m}^{b_\tau} P_{t,m}^{b_P} c_m^{SI} \varepsilon_{mt}},$$

where $\mathbf{S}, \mathbf{M_{t-\tau}}, \mathbf{P} \in \mathcal{S}^4$ are the compositions of brand sales, of brand media investments at time $t-3$, $t-4$ and $t-5$, and of brand prices. $b_\tau, b_P \in \mathbb{R}$ are the parameters associated to compositional explanatory variables and $\mathbf{c} \in \mathcal{S}^4$ is a composition of parameters associated to the dummy variable $SI$ (scrapping incentive).

The ILR transformed version of the model is

$$\mathbf{S}_t^* = \mathbf{a}^* + \sum_{\tau=3}^{5} b_\tau \mathbf{M}_{t-\tau}^* + b_P \mathbf{P}_t^* + \mathbf{c}^* SI_t + \boldsymbol{\varepsilon}_t^*$$

$$\Leftrightarrow S_{jt}^* = a_j^* + \sum_{\tau=3}^{5} b_\tau^* M_{j,t-\tau}^* + b_P^* P_{jt}^* + c_j^* SI_t + \varepsilon_{jt}^* \quad \text{for } j = 1, 2, 3,$$

where $\boldsymbol{\varepsilon}^*$ is supposed to be a Gaussian distributed error term. The balance matrix used

for the ILR transformation is the default matrix in the R software:

$$V_{ILR,4} = \begin{bmatrix} -\sqrt{1/2} & -\sqrt{1/6} & -\sqrt{1/12} \\ \sqrt{1/2} & -\sqrt{1/6} & -\sqrt{1/12} \\ 0 & \sqrt{2/3} & -\sqrt{1/12} \\ 0 & 0 & \sqrt{3/4} \end{bmatrix}$$

**Results** All explanatory variables are significant at 0.1% according to the analysis of variance (ANOVA). Figure 2.2 compares observed and fitted shares. It confirms that the model succeeds in fitting the main trends of brands' market shares. However, the model underestimates the market share of "Others" at the beginning of the period, and overestimates it in the end.

The parameters estimated with the ILR transformed model are presented in Table 2.3. The corresponding parameters for the model in the simplex are in Table 2.4. We remark that the coefficient associated to price is positive, which can be surprising, but price here is correlated with the quality image of the brand, which is very important for the customer who buys a durable and expensive good like a car.

Table 2.3: Estimated parameters on ILR coordinates - MCI model

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| $a_1^*$ | 0.3439 | 0.0151 | 22.84 | 0.0000*** |
| $a_2^*$ | 0.3363 | 0.0159 | 21.19 | 0.0000*** |
| $a_3^*$ | 0.6620 | 0.0263 | 25.14 | 0.0000*** |
| $b_1$ | 0.0267 | 0.0071 | 3.79 | 0.0002*** |
| $b_2$ | 0.0241 | 0.0062 | 3.90 | 0.0001*** |
| $b_3$ | 0.0264 | 0.0062 | 4.26 | 0.0000*** |
| $b_P$ | 1.2217 | 0.2313 | 5.28 | 0.0000*** |
| $c_1^*$ | -0.0241 | 0.0338 | -0.71 | 0.4758 |
| $c_2^*$ | -0.1690 | 0.0334 | -5.05 | 0.0000*** |
| $c_3^*$ | 0.1292 | 0.0336 | 3.84 | 0.0001*** |
| Nb param. | 10 | | | |
| Signif. codes: | 0 '***' | 0.001 '**' | 0.01 '*' | 0.05 '.' 0.1 ' ' 1 |

Table 2.4: Estimated parameters in the simplex - MCI model

| | $S_1$ (Citroën) | $S_2$ (Peugeot) | $S_3$ (Renault) | $S_4$ (Others) |
|---|---|---|---|---|
| (Intercept) | 0.1300 | 0.2114 | 0.2502 | 0.4084 |
| $M_{t-3}$ | | 0.0267 | | |
| $M_{t-4}$ | | 0.0241 | | |
| $M_{t-5}$ | | 0.0264 | | |
| $P_t$ | | 1.2217 | | |
| $SI$ | 0.2610 | 0.2523 | 0.2086 | 0.2780 |

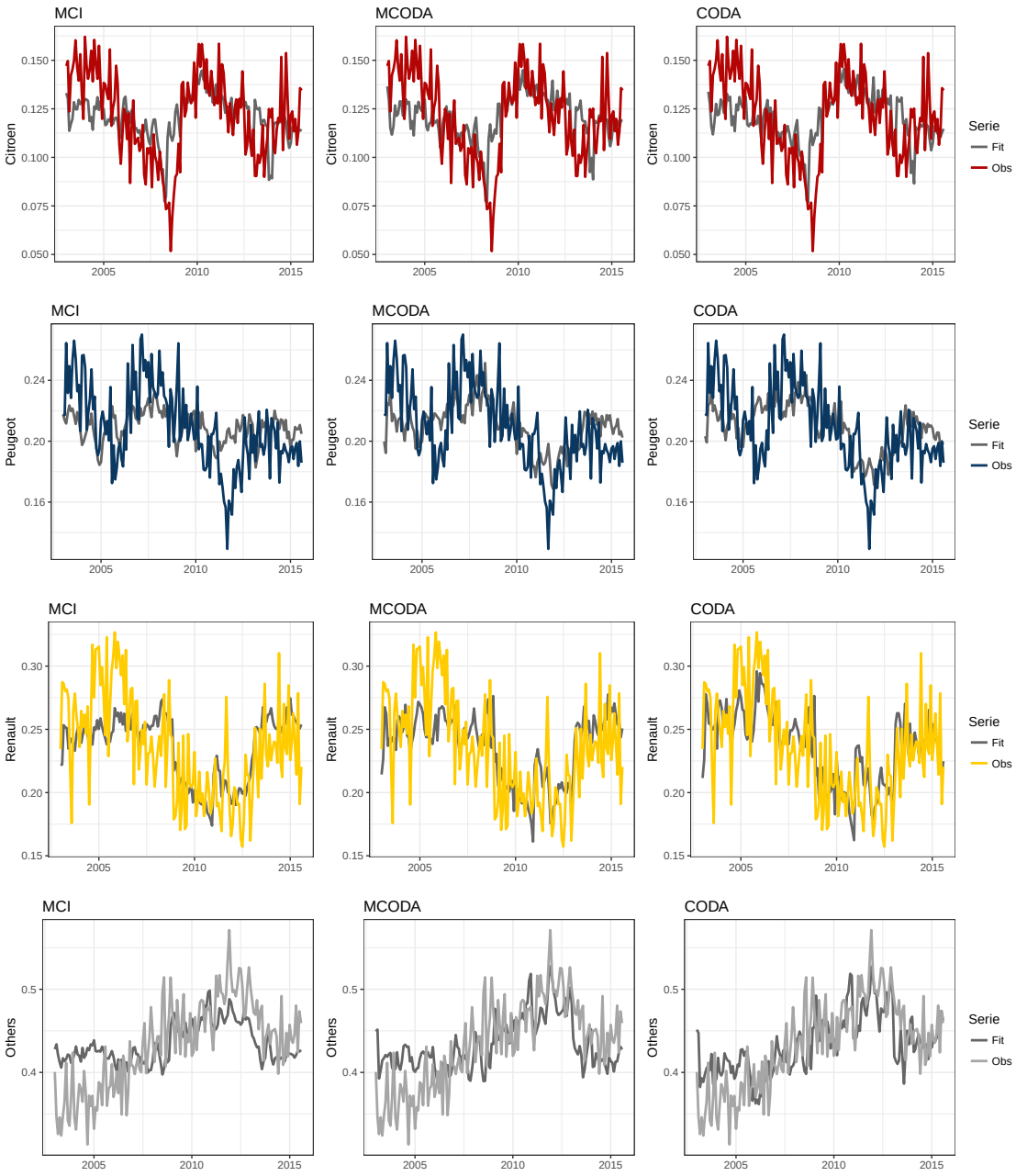


Figure 2.2: Observed (color) and predicted (grey) brands' market shares

### 2.4.2 Brand-specific impact of media investments (CODA model)

**Model** Now, let us look at a different specification of the model (dependent and explanatory variables are the same as in the MCI model) with brand-specific coefficients

and cross effects. It corresponds to the following model:

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{\tau=3}^{5} \mathbf{B}_\tau \boxdot \mathbf{M}_{t-\tau} \oplus \mathbf{B_P} \boxdot \mathbf{P}_t \oplus SI_t \odot \mathbf{c} \oplus \boldsymbol{\varepsilon}_t$$

$$\Leftrightarrow S_{jt} = \frac{a_j \prod_{\tau=3}^{5} \prod_{l=1}^{4} M_{t-\tau,l}^{b_{\tau,jl}} \prod_{l=1}^{4} P_{t,l}^{b_{P,jl}} c_j^{SI} \varepsilon_{jt}}{\sum_{m=1}^{4} a_m \prod_{\tau=3}^{5} \prod_{l=1}^{4} M_{t-\tau,l}^{b_{\tau,ml}} \prod_{l=1}^{4} P_{t,l}^{b_{P,ml}} c_m^{SI} \varepsilon_{mt}},$$

where $\mathbf{B}_\tau, \mathbf{B_P} \in \mathbb{R}^{D \times D}$ are the matrices of parameters associated to compositional explanatory variables.

The corresponding ILR transformed model is:

$$\mathbf{S}_t^* = \mathbf{a}^* + \sum_{\tau=3}^{5} \mathbf{B}_\tau^* \mathbf{M}_{t-\tau}^* + \mathbf{B}_\mathbf{P}^* \mathbf{P}_t^* + \mathbf{c}^* SI_t + \boldsymbol{\varepsilon}_t^*$$

$$\Leftrightarrow S_{jt}^* = a_j^* + \sum_{\tau=3}^{5} \sum_{l=1}^{3} b_{\tau,jl}^* M_{l,t-\tau}^* + \sum_{l=1}^{3} b_{P,jl}^* P_{lt}^* + c_j^* SI_t + \varepsilon_{jt}^* \quad \text{for } j = 1, 2, 3,$$

where $\boldsymbol{\varepsilon}^*$ is supposed to be a Gaussian distributed error term. The same balance matrix $V_{ILR,4}$ is used.

**Results** All variables of the model are significant at 0.1% according to the ANOVA, except for price which is significant at 1%. According to Figure 2.2, the CODA model seems to fit better than the MCI model (see Section 2.4.4 for the goodness-of-fit measures). The estimated parameters of the model are given in Table 2.5 and Table 2.6.

### 2.4.3 Interpretation of MCI and CODA models

**Marginal effect of media investments** We calculate the marginal effects of media investments at time $t-3$ on market shares at time $t$. The average marginal effects are reported in Table 2.7. They are quite consistent between the MCI model and the CODA model, with positive direct marginal effects and negative cross marginal effects. However, these measures are not really adapted to summarize an impact as they fluctuate a lot across time, as we can see in Figure 2.3 (marginal effects can be larger than 6e-08 but we voluntarily cropped the graph). The marginal effects of Citroën media investments are especially very high when these investments are very low, for example between 2007 and 2009.

**Elasticity of the share $S_j$ relative to $X_l$** For the MCI model, cross elasticities are necessarily negative and direct elasticities are necessarily positive if the parameter $b$ is positive. Moreover, cross elasticities of market shares $S_j$ with respect to a particular media budget $M_{l,t-3}$ are equal for any brand $j \neq l$. This is a lack of flexibility of the MCI model compared to the CODA model: it does not allow positive interaction between brands, and it considers that if a brand increases its media investments of 1% it impacts in the same way all competitors market shares $S_j$ (they will all decrease by $b$%).

Table 2.5: Estimated parameters on ILR coordinates - CODA model

| | $S_1^*$ (Peu. vs Cit.) | $S_2^*$ (Reu. vs Cit.,Peu.) | $S_3^*$ (Oth. vs Cit.,Peu.,Reu.) |
|---|---|---|---|
| (Intercept) | 0.3686*** | 0.3637*** | 0.6940*** |
| $M^*_{t-3,1}$ | 0.0193. | -0.0052 | 0.0081 |
| $M^*_{t-3,2}$ | 0.0162 | 0.0319* | -0.0245 |
| $M^*_{t-3,3}$ | -0.0069 | 0.0009 | 0.0279 |
| $M^*_{t-4,1}$ | 0.0208. | -0.0093 | 0.0205. |
| $M^*_{t-4,2}$ | 0.0151 | 0.0361** | -0.0259. |
| $M^*_{t-4,3}$ | -0.0197 | -0.0338· | 0.0278 |
| $M^*_{t-5,1}$ | 0.0289** | -0.0115 | 0.0278* |
| $M^*_{t-5,2}$ | 0.0104 | 0.0206* | -0.0274. |
| $M^*_{t-5,3}$ | -0.0114 | 0.0064 | 0.0323. |
| $P_1^*$ | 0.8854. | -0.5981 | 1.9138*** |
| $P_2^*$ | 0.0151 | 0.2615 | 0.6509 |
| $P_3^*$ | -0.6442 | -0.3729 | 2.4717*** |
| $SI^*$ | -0.0394 | -0.2088*** | 0.2070*** |
| Adjusted R2 | 0.3353 | 0.3255 | 0.3269 |
| Nb param. | 42 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2.6: Estimated parameters of $M_{t-1}$ in the simplex - CODA model

| | $S_1$ (Citroën) | $S_2$ (Peugeot) | $S_3$ (Renault) | $S_4$ (Others) |
|---|---|---|---|---|
| $M_{t-3,1}$ | 0.0179 | -0.0079 | -0.0067 | -0.0032 |
| $M_{t-3,2}$ | -0.0016 | 0.0111 | -0.0161 | 0.0066 |
| $M_{t-3,3}$ | -0.0132 | 0.0084 | 0.0292 | -0.0243 |
| $M_{t-3,4}$ | -0.0030 | -0.0115 | -0.0064 | 0.0209 |

Table 2.7: Average marginal effects of media investments $\check{M}_{t-3}$ on market shares

| $me(S_{jt}, \check{M}_{l,t-3})$ | MCI | | | | CODA | | | |
|---|---|---|---|---|---|---|---|---|
| | $\check{M}_{C,t-3}$ | $\check{M}_{P,t-3}$ | $\check{M}_{R,t-3}$ | $\check{M}_{Z,t-3}$ | $\check{M}_{C,t-3}$ | $\check{M}_{P,t-3}$ | $\check{M}_{R,t-3}$ | $\check{M}_{Z,t-3}$ |
| $S_{Citroën,t}$ | **1.93e-05** | -1.65e-09 | -2.13e-09 | -3.01e-10 | **1.68e-05** | -7.20e-10 | -2.82e-09 | -2.00e-10 |
| $S_{Peugeot,t}$ | -4.58e-06 | **1.14e-08** | -3.09e-09 | -5.30e-10 | -7.67e-06 | **5.51e-09** | 7.72e-09 | -7.52e-10 |
| $S_{Renault,t}$ | -4.88e-06 | -3.64e-09 | **1.35e-08** | -5.96e-10 | -6.43e-06 | -1.14e-08 | **2.23e-08** | -5.71e-10 |
| $S_{Others,t}$ | -9.89e-06 | -6.10e-09 | -8.24e-09 | **1.43e-09** | -2.66e-06 | 6.60e-09 | -2.72e-08 | **1.52e-09** |

C: Citroën; P: Peugeot; R: Renault; Z: Others.

Figures in bold: direct elasticities.

Let us consider a situation where the market shares of Citroën, Peugeot, Renault and Others in the B segment are respectively 10%, 25%, 25% and 40%. According to Table 2.8, if Renault increases its media investments $\check{M}_{t-3}$ of about 1%, the average elasticity of the MCI model on the studied period suggests that its market share should increase by 0.0204% to reach 25.005% and that competitors market shares should decrease by

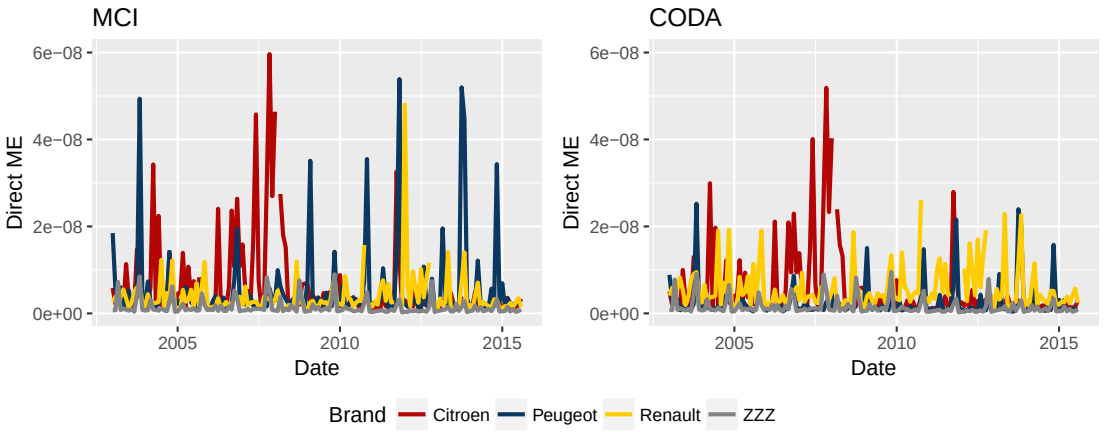


Figure 2.3: Direct marginal effects of $M_{j,t-3}$ on $S_{jt}$ across time

0.0204% to reach respectively 9.998%, 24.995% and 39.992%. Note that here we take an example for an arbitrary share of 25% using the average elasticity. However, the only way to ensure that the sum of the modified shares $\sum_{m=1}^{D} S'_{mt}$ is equal to 1 is to use the corresponding elasticities calculated at the same time $t$, not the average elasticities.

In the CODA model, when brand-specific effects and cross-effects are taken into account, the direct elasticity of Renault market share in the B segment relative to its corresponding media investments (0.0327) is much higher than for other brands, see for example Peugeot which has the lowest (0.0099). Note that positive cross effects (synergies) are possible in the CODA model: for example when Renault invests more in media, it tends to help its own market share a lot, but also to raise a little bit the share of Peugeot, and to have a negative impact on Citroën and Others. Then, after closure and depending on the considered values of $S_j$, an increase in Renault media investments in the B segment can increase or decrease the Peugeot market share.

Taking the same example as previously, according to the CODA model, if Renault increases its media investments $\check{M}_{t-3}$ of about 1%, the average elasticity on the studied period suggests that its market share should increase by 0.0327% to reach 25.008% and that competitors market shares should respectively decrease by 0.0097%, increase by 0.0119% and decrease by 0.0208% to reach respectively 9.999%, 25.003% and 39.992%.

As shown in Figure 2.4, the estimated direct elasticities are quite stable across time. However, as elasticities in the MCI model are computed using the same parameter $b$ for all brands, they are closer to each other than in the CODA model where they are computed using different parameters $b_{jl}$. The direct elasticity of Renault is larger than those of other brands during the whole studied period.

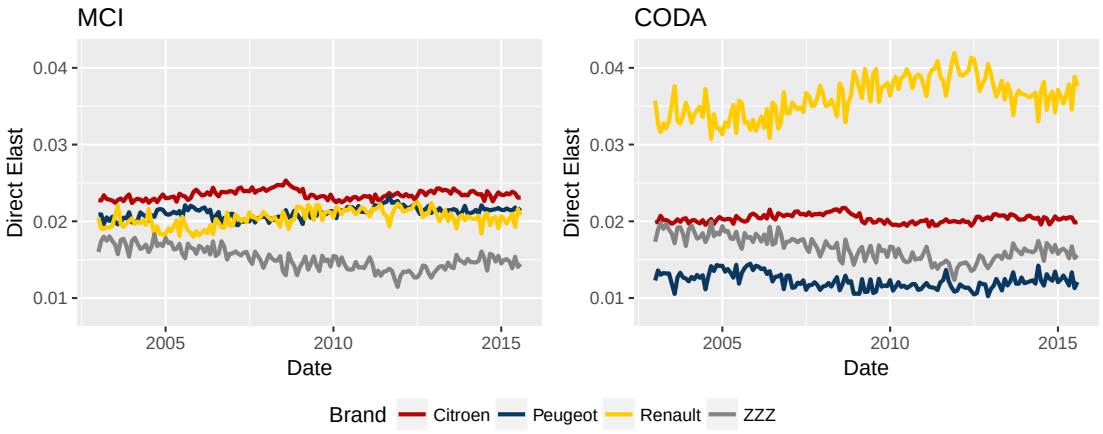


Figure 2.4: Direct elasticity of $S_{jt}$ relative to $M_{j,t-3}$ across time

Table 2.8: Average elasticity of market shares relative to media investments $\check{M}_{t-3}$

| $e(S_{jt}, \check{M}_{l,t-3})$ | MCI | | | | CODA | | | |
|---|---|---|---|---|---|---|---|---|
| | $\check{M}_{C,t-3}$ | $\check{M}_{P,t-3}$ | $\check{M}_{R,t-3}$ | $\check{M}_{Z,t-3}$ | $\check{M}_{C,t-3}$ | $\check{M}_{P,t-3}$ | $\check{M}_{R,t-3}$ | $\check{M}_{Z,t-3}$ |
| $S_{Citroën,t}$ | **0.0235** | -0.0056 | -0.0063 | -0.0116 | **0.0204** | -0.0028 | -0.0097 | -0.0078 |
| $S_{Peugeot,t}$ | -0.0032 | **0.0211** | -0.0063 | -0.0116 | -0.0054 | **0.0099** | 0.0119 | -0.0163 |
| $S_{Renault,t}$ | -0.0032 | -0.0056 | **0.0204** | -0.0116 | -0.0043 | -0.0173 | **0.0327** | -0.0111 |
| $S_{Others,t}$ | -0.0032 | -0.0056 | -0.0063 | **0.0151** | -0.0008 | 0.0054 | -0.0208 | **0.0161** |

C: Citroën; P: Peugeot; R: Renault; Z: Others.
Figures in bold: direct elasticities.

**Elasticity of the ratio $\frac{S_j}{S_{j'}}$ relative to $\check{X}_l$** (see Table A.1 in the appendix A.2.4)

In the MCI model, the elasticity of a ratio $S_j/S_{j'}$ relative to $\check{X}_j$ is equal to 0.0267, whereas in the CODA model it can be smaller or larger according to the considered brands: the largest elasticity is for $S_R/S_Z$ relative to $\check{X}_R$ which is equal to 0.0535. In general, ratios between the market shares of Renault and other brands are quite positively sensitive to media investments of Renault. For example, if the ratio $S_R/S_Z$ is equal to $25/40 = 0.6250$ and Renault increases by 1% its media investments, then the ratio will increase up to 0.6253. Let us remind that this measure does not depend on the considered period. This evolution is consistent with the fact that the market share of Renault is very positively elastic and the market share of "Others" is very negatively elastic to Renault media investments, as seen in Table 2.8.

**Odds ratio of $\frac{S_j}{S_{j'}}$ relative to a change of $\check{X}_l$** (see Table A.2 in the appendix A.2.4)

As expected, this measure is consistent with the previous one. In the MCI model, the odds ratio of any couple of brands' market shares $S_j/S_{j'}$ relative to a change of 10% of $\check{M}_{j,t-3}$ is equal to 1.0025, whereas it can reach 1.0054 in the CODA model for the ratio $S_R/S_Z$ for a change of 10% in $\check{M}_{R,t-3}$. It means that if the ratio of market shares

of Renault over Others is equal to $25/40 = 0.6250$ and if Renault decides to increase its media budget by 10%, then this ratio will increase to 0.6266 according to the MCI model and to 0.6284 according to the CODA model.

**Elasticity of $\frac{S_j}{g(S_{-j})}$ relative to $\check{X}_l$** (see Table A.3 in the appendix A.2.4)
As in the MCI model, the parameter $b_1$ will be the same for any chosen ILR transformation, then we obtain that $e\left(\frac{S_{jt}}{g(S_{-jt})}, \frac{M_{j,t-3}}{g(M_{-j,t-3})}\right) = e\left(\frac{S_{jt}}{S_{j't}}, M_{j,t-3}\right) = e\left(\frac{S_{jt}}{S_{j't}}, \frac{M_{j,t-3}}{M_{j',t-3}}\right)$. Moreover, these elasticities are consistent with previous impact measures, and the largest one concerns the ratio $\frac{S_R}{g(S_{-R})}$ relatively to the ratio $\frac{M_R}{g(M_{-R})}$, which is equal to 0.0389%. For example, let us consider a situation where the market shares are the following: $(S_C, S_P, S_R, S_Z)' = (13, 22, 25, 40)'$, inducing that $\frac{S_R}{g(S_{-R})} = 1.1095$. Then, if Renault increases its media investments by 1% of the geometric average of other brands media investments, we can expect its market share to move from 110.95% to 110.99% of the geometric average market share of others.

### 2.4.4 Complexity and goodness of fit

We have seen that the MCI model and the CODA model can be used for the same type of application. The CODA model is more complex than the MCI model because it allows to have component-specific parameters for each explanatory variables along with cross-effects parameters. We have also fitted an intermediate MCODA model without component-specific and cross-effects parameters for the price. The number of parameters to fit for the CODA model can be a serious limitation when the number of components $D$ and the number of explanatory compositions $K$ increase. For example, in our application the MCI model involves 10 parameters whereas the MCODA model and the CODA model involve respectively 34 and 42 parameters.

However, the CODA model is also more flexible than the MCI model in the sense that it allows to have positive synergies (positive interactions) between some shares, whereas cross elasticities of the MCI model are necessarily negative as long as the direct elasticity is positive (the cross elasticity is of opposite sign of the direct elasticity by construction). For example, we see in Table 2.8 that when the media investments of Citroën increase, it tends to benefit also to "Others", and when the media investments of Renault increase, it tends to benefit to Peugeot.

Is the complexity of the CODA model useful to explain brands' market shares of the B segment? According to the Fisher tests of the MCI model against the CODA model, and of the MCODA model against the CODA model, for which the estimated statistics are respectively 2.22 and 3.72 to be compared to the 99% quantiles, respectively 0.51 and 0.56, we conclude that the CODA model is significantly better than the MCI and MCODA models. This means that the brand-specific and cross-effect parameters for media investments and prices are necessary to reflect the complexity of the competitive interaction in the automobile market.

We also compare cross-validated quality measures, recorded in Table 2.9: the adjusted $R^2$ calculated on the transformed model with coordinates used for the estimation (for the CODA model, the adjusted $R^2$ is computed on the transformed model which uses dummy variables for estimations, as in the MCI and MCODA models), the $R^2$ based on the total variance as defined in compositional data analysis (see Section A.1.4 in the appendix for more details), and the RMSE. The out-of-sample computation process is the same than in Chapter 1 (see Section 1.4.2). All measures agree on the fact that the CODA model is better than the MCI model and the MCODA model for our application.

Table 2.9: Cross-validated quality measures

| | MCI | MCODA | CODA |
|---|---|---|---|
| Adj. $R^2$ | 0.9250 | 0.9274 | **0.9310** |
| $R^2_T$ | 0.3183 | 0.4002 | **0.4513** |
| RMSE | 0.0326 | 0.0913 | **0.0322** |

## 2.5 Conclusion

The focus of this chapter is to combine the best part of the MCI model and of the CODA model presented in Chapter 1, in order to improve the interpretability of CODA models, and to improve the estimation of MCI models: we stress the positive aspects of using an isometric log-ratio (ILR) transformation to estimate the MCI model, as in the CODA model, instead of the usual centered log-ratio (CLR) transformation used in marketing. We also develop an intermediate specification, the MCODA model, allowing to have a simple specification for some explanatory variables and a complex specification with cross effects between components for others. A model selection procedure is proposed using an adapted Fisher test, considering that the CODA model is the unconstrained model to be compared to the constrained models, the MCI model or the MCODA model.

This chapter presents a set of possible measures, mutually consistent, to interpret parameters of these models: marginal effects, elasticities and odds ratios. The elasticity of a component relative to an explanatory variable is the relative variation of this component to a relative variation of the explanatory variable, *ceteris paribus*. This type of measures is totally adapted to enhance the interpretability of these models. However, this measure is observation dependent and we have to make sure that it is stable across observations to use it. Marginal effects are not well adapted to interpret this kind of models because they depend a lot on the considered observation. The other types of measures presented have the advantage to be observation independent, but they are more difficult to interpret in practice because they involve ratios of shares.

The two models, and an intermediate specification, are applied to the B segment of the French automobile market, for the purpose of measuring the impact of brands' media investments on brands' market shares. The CODA model fits our data better than the MCI and the MCODA models according to cross-validated quality measures and to the Fisher tests. In the CODA model, Renault is the brand which has the largest direct elasticity to media investments. This model also shows interesting non-symmetric synergies between brands.

In the next chapter, we are going to show how to properly take into account the carryover effect of the media investments, in a better manner than taking different lags of media investments as we have done here. We are also going to consider separately the impacts of the different advertising channels.

# Chapter 3

# Impact of advertising on brand's market shares in the automobile market: a multi-channel attraction model with competition and carryover effects

This third chapter aims to present the final regression model chosen to answer the major question addressed in this thesis. In order to do that we address here all the issues highlighted before, concerning the application itself or the statistical aspects of the modeling. In particular, we propose here to align the media investments data on the registration data, and we construct a model able to take into account the carryover effects of the advertising in a multi-channel case, and the cross effects of competitors. The corresponding model, called CODAAd, is validated after a residuals diagnostic and the measurement of several accuracy indicators. We also explain how to construct confidence and prediction ellipsoids, and we interpret the advertising elasticities.

This chapter will be submitted to a marketing journal. Thus, this chapter is less theoretical than the previous ones, and focuses more on the practical interpretation of the impact of advertising on market shares.

## 3.1 Introduction

The effect of marketing mix variables (advertising, price, promotion, distribution) have been modeled since the 50's using the so-called market response models, where the response variable is usually the sales or the market shares of products or brands (see Hanssens et al. [27] for a review of existing models).

Three main categories of models are used in practice: linear, multiplicative and attraction models. Only the latter category complies with the constraints of positivity and summing up to one of market shares data, as emphasized by Cooper and Nakanishi [10] (p.28). Nevertheless, attraction models are not used systematically for market shares modeling, because of three main reasons.

1. The first one is that some authors have shown empirically that attraction models, as the multiplicative competitive interaction model (MCI) for example, do not give significantly better results than the others in terms of fitting and prediction accuracy (see for example Ghosh et al. [20] and Leeflang et al. [36]). Nevertheless, Naert et al. [51] have made the opposite claim a few years ago, suggesting that the conclusion can depend on the considered application.

2. The second reason is that the estimation of an attraction model is not straightforward: it is a non-linear model which can be linearized by a transformation, generally the log-centering transformation, also called centered log-ratio transformation (CLR) in the compositional data analysis literature (see Aitchison [1]). A simple estimation by ordinary least squares is generally run on the resulting coordinates, while it is obvious that the log-centered error terms cannot be independently distributed. Generalized least squares (GLS) and iterative generalized least squares (IGLS) have also been considered by several authors, but without concluding to a significant improvement of the estimation (see for example Ghosh et al. [20], Leeflang and Reuyl [36], and Cooper and Nakanishi [10], p.128).

3. The third reason is that they are often overparametrized. The classical MCI suggests that the impact of a marketing instrument is the same for all brands, which is often too restrictive. The differential MCI model (DMCI) includes brand specific parameters, leading to $D + DK$ parameters, where $D$ is the number of brands and $K$ the number of explanatory variables, but it ignores the potential cross effects between brands. The additional specification of cross effects, done in the so-called fully extended MCI model (FEMCI), leads to a huge number of parameters: $D(1 + DK)$. With the estimation on the CLR transformed model, only the centered version of these coefficients can be identified, although they are sufficient for interpreting the model, according to Cooper and Nakanishi [10] (p.145).

We argue in favor of the use of attraction models to model market shares. Our claims concerning the three previous points are the following:

1. From a mathematical point of view, market shares data are "compositions" (vectors of positive numbers where only the relative information is of interest) and they

belong to the simplex space. Their compositional nature must be considered to analyze them. Compositional data analysis (CODA) is a recent field in statistics which has developed a set of tools, including compositional regression models with the advantage of including brand specific (differential) parameters and flexible (non-symmetric and not necessarily negative) cross effects. We prove in Chapter 1, Section 1.3.5, that these models are very close to the fully extended MCI model.

2. We suggest in Chapter 2, Section 2.2.1, to use another transformation than CLR called isometric log-ratio (ILR) transformation, which is recommended in the CODA literature because it allows to obtain orthonormal transformed error terms, with non-constant variance between the obtained coordinates. The associated estimation method is easy to implement (e.g. with the R package **compositions**).

3. Concerning the DMCI model, it should not be used as we prove in Chapter 1, Section 1.3.5, that this model is not scale invariant[1]. Concerning the FEMCI model, we prove in the same section that if it is estimated using the ILR transformation, it is then identical to the CODA model presented in Chapter 1. Then, the $D(1+DK)$ parameters to be estimated can be recovered estimating only $(D-1)(1+(D-1)K)$ parameters. Moreover, in order to determine if cross effects for a given marketing instrument are really improving the model, we propose in Chapter 2, Section 2.2.2, a model selection based on an adapted Fisher test.

In addition to the MCI model, the Dirichlet (DIR) regression model can also be used to model market shares respecting their compositional nature. Although rarely used in marketing, the DIR model is a flexible model allowing the specification of differential effects for example, and it is easy to implement (R package **DirichletReg**).

Once the type of market response model is chosen, we then need to determine how to take into account the dynamic aspect of the relationship between market shares and advertising. Some authors have emphasized the existence of short term and long term effect of advertising on sales (see for example Assmus et al. [2] and Lodish et al. [39]). In the case of durable and expensive goods like automobile, we can expect the advertising impact to be spread over several periods, with diminishing returns effect on sales.

This is called the carryover effect of advertising and it is usually integrated in market response models using a stock variable, built using a retention rate which can be estimated econometrically. In advertising research, this notion is also called "adstock" variable and was initiated by Broadbent (1979). The most commonly used adstock model is the Koyck model, defined as $Q_t = \mu + \beta Adstock_t + \epsilon_t$ where the adstock function is equal to $Adstock_t = (1-\lambda)(M_t + \lambda M_{t-1} + \lambda^2 M_{t-2} + \dots)$, $Q_t$ is the demand at time $t$, $M_t$ is the media investment at time $t$, and $\lambda$ is the retention rate. Then, $\beta(1-\lambda)$ can be interpreted as the current (short term) effect of advertising and $\beta$ the carryover (long term) effect of advertising, and we can say that $\theta\%$ of the advertising impact occurs in the $\log(1-\theta)/\log(\lambda) - 1$ periods after advertising.

[1] It means that if marketing instruments are used in thousands of euro or in euro, the DMCI model will not lead to the same results in terms of fitted market shares.

Vakratsas and Ambler [66] report that "*Clarke (1976) and Assmus, Farley, and Lehmann (1984), in meta-analytic studies, suggest that 90% of the advertising effects dissipate after three to fifteen months. Leone (1995), in an empirical generalizations study, suggests that the range be narrowed to six to nine months*". However, Leone [37] also emphasizes the fact that the retention parameter $\lambda$ "*should increase as the level of aggregation increases*". Note that advertising campaigns are often analyzed at the week level and for FMCGs (fast moving consumer goods), whereas in our application we are observing the monthly advertising budgets for a durable good. Then we can expect to find larger carryover effects of advertising.

Usually carryover effects are estimated in the case of market response models for sales and for only one marketing instrument (advertising in most of the cases), not for market share models with multi-channel advertising.

In this chapter, we first present a descriptive compositional analysis of the competition situation in the French automobile market, of the marketing mix habits and pricing strategy.

In a second phase, we develop a fully extended multi-channel attraction model with adstock (what we call the CODAAd model), which considers the cross effects of television, outdoor, radio and press advertising budgets between brands and their carryover effects. This model allows to distinguish between short-term and long-term effects of the advertising. As market shares are compositions belonging to the simplex, we take benefit from the compositional data analysis literature to estimate properly this model. We also explain how to determine the decay parameters of advertising in this case.

Then, we present an application to the main segment of the French automobile market where different specifications of market share models, including Dirichlet models, are compared in terms of complexity, goodness of fit and prediction accuracy. A residuals diagnostic is done and we explain how to build confidence and prediction ellipsoids in the space of market shares.

We then interpret the chosen model in terms of short term advertising elasticities of market shares by channel, and we conclude on practical findings for car manufacturers concerning marketing mix strategies.

## 3.2 A compositional data analysis of the French automobile market

We are working on a data base coming from the French registration data base, which contains the sales of all brands. It is important to note that what we call "sales" at time $t$ actually correspond to registrations of new passenger vehicles at time $t$, which can correspond to purchases during the previous months, due to the delivery delay.

### 3.2.1 A market in 5 segments

The usual segmentation of the automobile market in Europe is done in five main segments for passenger cars, from A to E, according to the size of the chassis: small vehicles are in the A segment, and largest one in the E segment. Sport and luxury cars are grouped in another additional F segment.

Figure 3.1 represents the registrations (called sales below by abuse of language) in volume and the corresponding market shares of each segment from 2003 to 2015 in France. A strong seasonality exists in this market. The economic crisis of 2008 led to a decrease of sales in most of the segments, especially in the E segment made of the most expensive cars, except for the A segment which gained market shares at the expense of the others. The scrapping incentive put in place by the French government from December 2008 to December 2010, and delimited by black dotted lines in Figure 3.1, has clearly boosted sales of the first two segments.

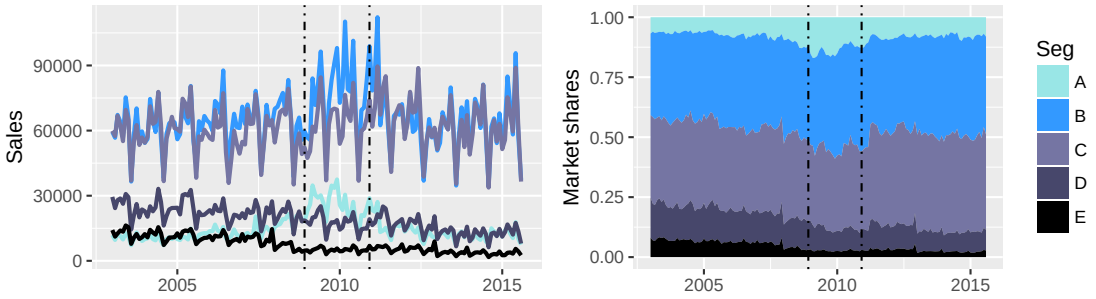


Figure 3.1: Segmentation of the French automobile market

For the rest of this chapter, we focus on the B segment which is the main segment in France (almost 40% of sales on average in the period) and includes for example the best-seller model Renault Clio.

### 3.2.2 Overview of competition in the B segment

From 2003 to 2015, 34 different brands sell vehicles in the B segment in France. However, three main brands take the lead of the market on the whole period: the three French manufacturers Renault, Peugeot and Citroën, as can be seen in Figure 3.2.

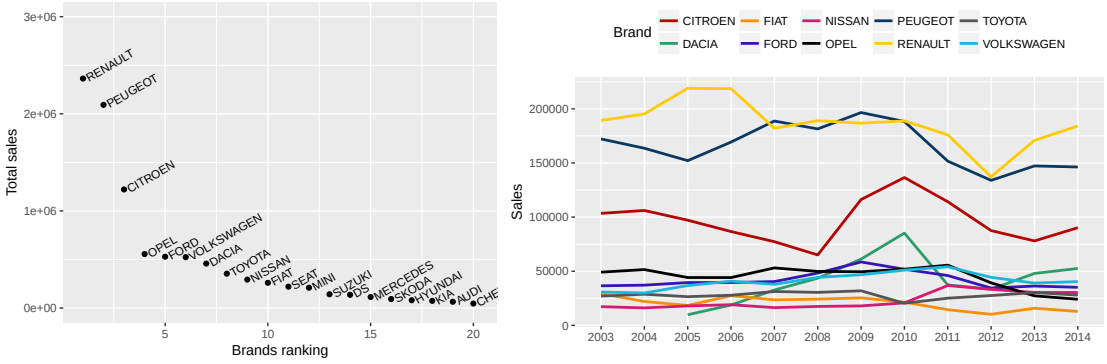


Figure 3.2: Brands ranking on total sales from 2003 to 2015 (left); Yearly sales of top 10 brands (right)

For the sake of simplicity and for confidentiality reasons, we focus on Renault, Peugeot and Citroën, and the remaining brands are grouped in an "Others" category (also denoted "ZZZ" below).

From Figure 3.3, we can easily see the order of magnitude of each brand in terms of sales across time. During the scrapping incentive period, the sales of the group of smaller brands Others increased quite a lot. Citroën also seems to take advantage from this incentive, contrary to Renault and Peugeot. We observe that the strong seasonality of sales is not visible on market shares, suggesting that this seasonality has the same impact on all brands.

In terms of market shares, it is easier to see the evolution of the competition with the 3D ternary diagram[2] presented in Figure 3.4, and we clearly observe a move in the direction of Others from 2003 to 2012. The market share of Others was around 35% in 2003 versus almost 50% in 2011-2012. We note that the supremacy of the three leaders was especially true at the beginning of the period (blue points) but things changed since the economic crisis (in green). The scrapping incentive (in anise green) benefits to Citroën and Others, as seen previously. At the end of the period we observe a slight backward step of the Others market share. Renault seems to be the brand the most affected by the competition of new or small brands (Others), according to 2D ternary diagrams in Figure 3.5.

### 3.2.3 Advertising budgets and channels

The question we want to answer is: "what is the impact of advertising on brands' market shares?". To do so, it is important to have an idea of who is spending the most relative to other brands, and in which channel.

[2]Example of interpretation of a 3D ternary diagram: the closer we are from the Others vertex, the higher the Others' market share is. If the point is on the face delimited by Citroën/Renault/Peugeot vertices, then the Others' market share is null. If the point is in the center of the tetrahedron, then all market shares are equal to 1/4.

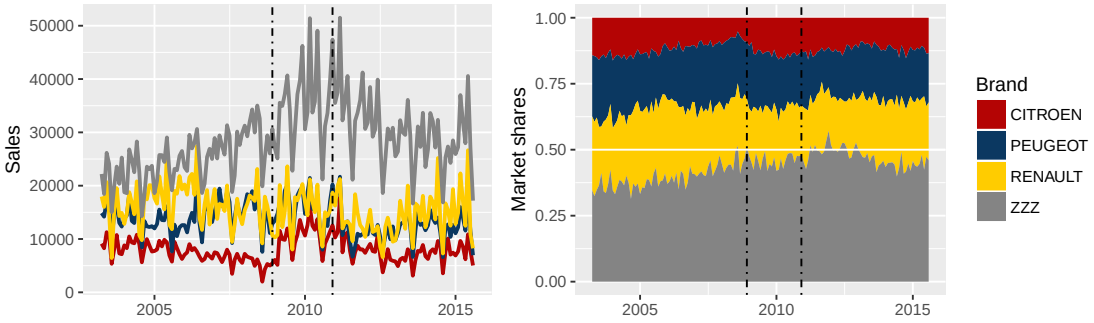


Figure 3.3: Sales and market shares - Citroën, Peugeot, Renault and Others

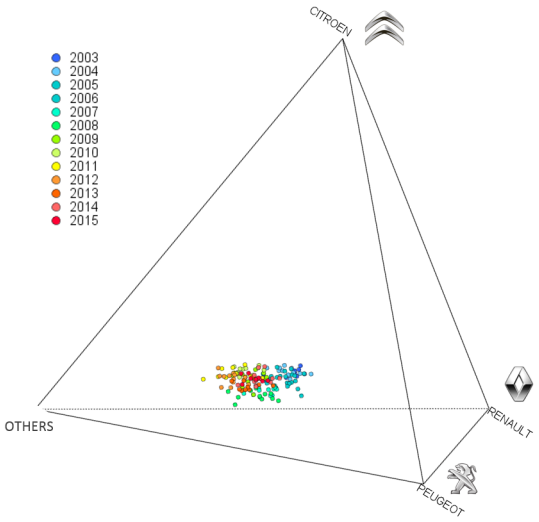


Figure 3.4: 3D ternary diagram of sales - Citroën, Peugeot, Renault and Others

### Aligning advertising with registrations

First of all, media investments, expressed in euro, are not coming from the registration data base but from a different data base of advertising tracking. Customers who register their vehicle at time $t$ have not necessarily been exposed to media investments spent at time $t$ before their purchase decision, because of the delay between the purchase act and the registration: the customer may have purchased his car at $t-2$. As we want to put in parallel the advertising potentially seen by customers and the purchase acts, we need to readjust media investment on registration time.

In order to do so, we use empirical knowledge about delivery times, which almost correspond to the difference of time between the purchase act and the registration. This delay evolves a little bit across years and across brands, but on average, the registrations of month $t$ are made of 31% of sales of month $t$, 34% of sales of month $t-1$, 21% of sales of month $t-2$, and 14% of sales of month $t-3$ or before. Then, we can consider that the

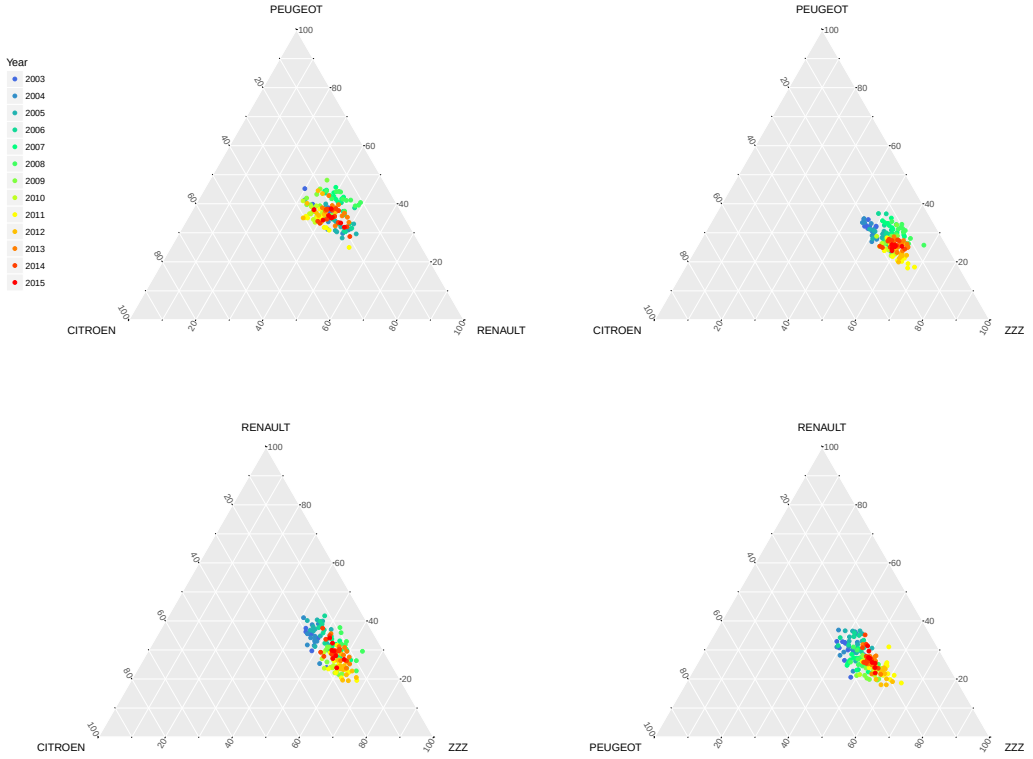


Figure 3.5: 2D ternary diagrams of sales - Citroën, Peugeot, Renault and Others

media investments corresponding to the vehicles registered at time $t$ are made of 31% of month $t$ media investments, 34% of month $t-1$ media investments, 21% of month $t-2$ media investments, and 14% of month $t-3$ media investments (see Table A.4 in the appendix). We use these "weighted" media investments throughout this chapter.

### Marketing mix between channels

Media investments are allocated among six channels: television, outdoor, press, radio, internet and cinema. Cinema's budget is marginal or null for most of the brands and the nature of advertising on internet has changed a lot between 2003 and 2015. Then, we choose not to consider these two channels.

The (weighted) media investments by channel are compared across brands in Figure 3.6, in euro and in share of voice. The media investments of Renault, Peugeot and Citroën are of the same order of magnitude for outdoor, radio and press channels, but in general Citroën spends less in television than the two leaders. Globally, advertising expenses vary a lot from one month to another for every channel and for all brands, much more than market shares. Moreover, we remark a net increase of outdoor, television and

press shares-of-voice for the group Others between 2010 and 2013, which seems to be concomitant with the increase in Others market shares. Similarly, the TV share of voice of Renault is quite large before 2007 and after 2013, which are the periods where Renault has larger market shares.

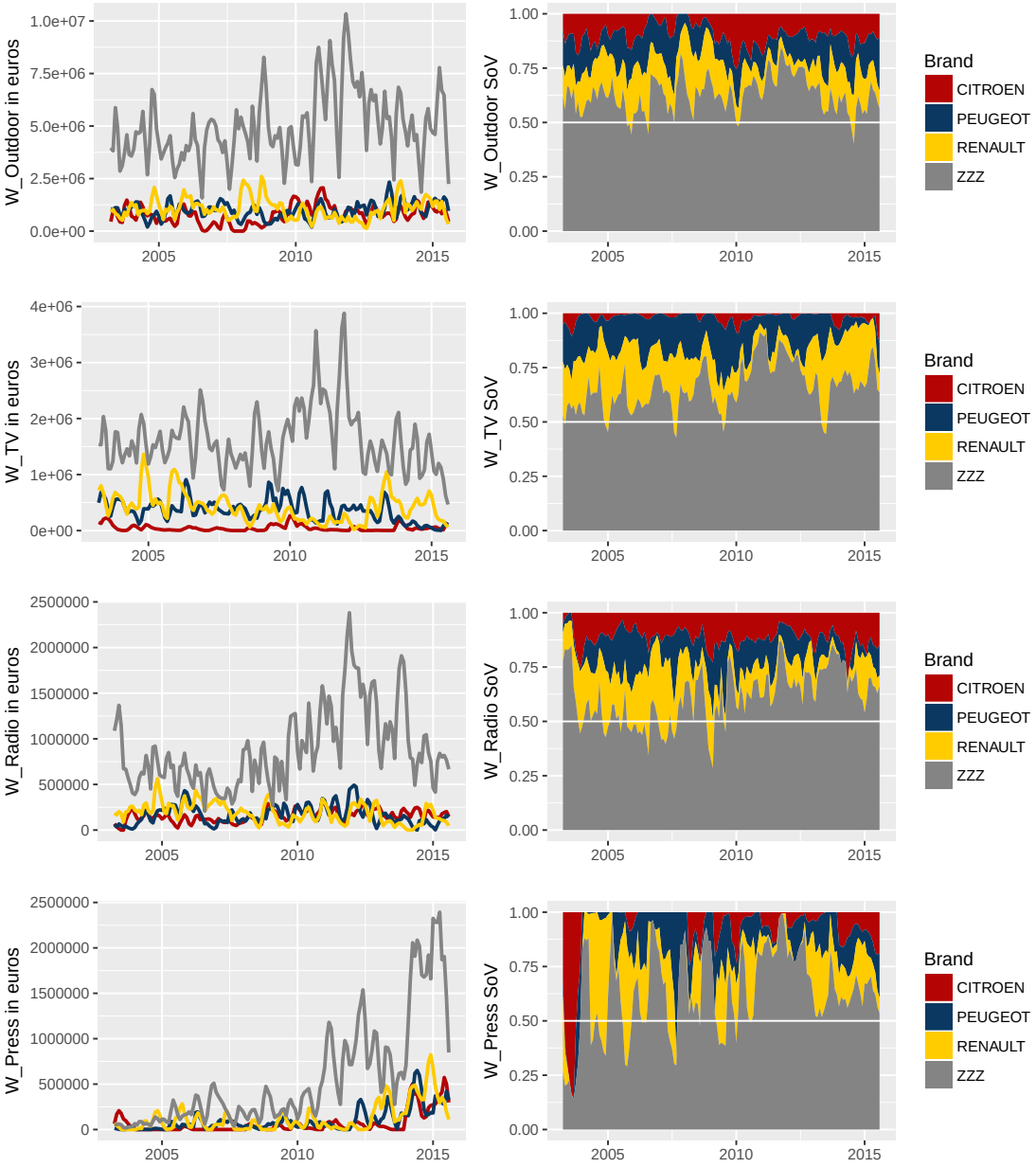


Figure 3.6: Advertising budgets and Share of Voice by channel - Citroën, Peugeot, Renault and Others

Figure A.1 in the appendix compares for each brand the advertising expenses by channel. Outdoor advertising (in orange) represents the main item of expenditure for all brands. Television is generally the second one (except for Citroën). We can remark an increase in the press budget for all brands at the end of the period. Note that here we are not interested in the impact of the composition across channels of the advertising budgets, but in the impact of the share of voice across brands for every channel.

Let us look at the relationship between sales and advertising by channel and by brand. In Figure A.2 in the appendix, the graphs in the left column represent the volume of registrations of a brand as a function of the advertising budget in euro of this brand, while in the right column the graphs represent the market shares of brands as a function of shares of voice (relative media investment) for each channel. We observe a positive relationship between sales and advertising, in volume and in share. However, this positive relationship is even stronger, according to correlation coefficients, in shares than in volumes, which confirms our assumption that competition cannot be omitted, and that a compositional approach should be undertaken. Note that correlation coefficients indicated in Figure A.2 are the correlation coefficients between the y and the x axes variables, all brands combined.

### 3.2.4 Pricing strategy

Registrations and prices are based on the same source: the registration data base. Then, prices at time $t$ do correspond to vehicles registered at time $t$. We computed average prices in euro by brand for the B segment using the catalogue prices of each vehicle model of the brand and weighting by the corresponding sales volume of each vehicle model. The smallest brands grouped in Others tend to have a lower price on average, but it is less true at the end of the period. Renault had a price decline in 2012 due to the liquidation of the Clio's stock before the new version (Clio IV). Prices are increasing across time for all brands almost in the same way. Then, if we look at the relative prices, they are quite stable, as can be seen in Figure 3.7. They indicate the position of the brand inside the segment (low cost versus high quality). Figure A.3 in the appendix shows that, as for media investments, the relationship between sales and price is stronger (negatively) in shares than in volumes.

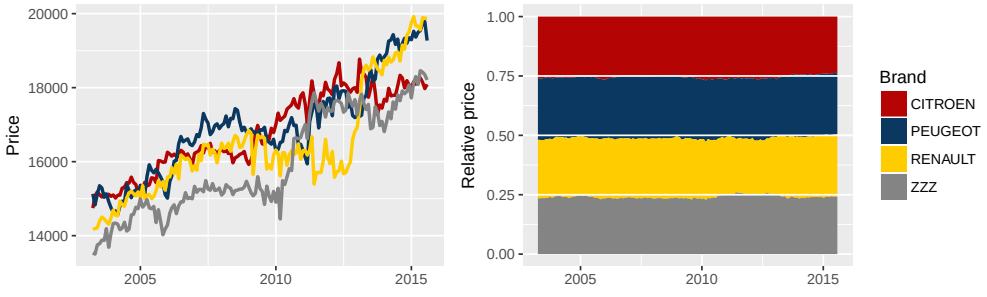


Figure 3.7: Catalogue prices and relative prices - Citroën, Peugeot, Renault and Others

## 3.3 Multi-channel attraction model with carryover effects

### 3.3.1 Extending the Koyck model to the multi-channel attraction case

In order to properly measure the link between advertising budgets by channel and market shares for the three leaders of the French B segment and the group of others brands, we develop a multi-channel attraction model specifying the carryover effect of advertising in a similar manner as in the Koyck model. The Koyck model (or geometric distributed lag (GL) model in Hanssens et al. [27]), the most famous adstock model, is defined by:

$$Q_t = \mu + \beta Adstock_t + \epsilon_t \tag{3.1}$$

where $Adstock_t = \beta(1-\lambda)\sum_{\tau=0}^{\infty}\lambda^\tau M_{t-\tau}$ is the adstock at time $t$, $M_t$ represents the media investment at time $t$, and $0 \leq \lambda < 1$ is the decay parameter.

Let us now adapt this model with adstock for the multi-channel advertising case and an attraction model formulation of MCI type. We call this model MCIAd:

$$S_{jt} = \frac{a_j \prod_{c=1}^{C} Adstock_{cjt}^{b_c}\epsilon_{jt}}{\sum_{l=1}^{D} a_l \prod_{c=1}^{C} Adstock_{clt}^{b_c}\epsilon_{lt}} \tag{3.2}$$

where $Adstock_{cjt} = \prod_{\tau=0}^{\infty} M_{cj,t-\tau}^{\lambda_c^\tau(1-\lambda_c)}$ is the adstock at time $t$ of brand $j$ for the advertising channel $c$, and where each channel $c = 1, \ldots, C$ has its proper decay parameter $\lambda_c$ (the same for all brands $j$). This model can be equivalently written with the operators of the simplex, as in the compositional data analysis literature (see Chapter 1 for brief definitions and Pawlowsky-Glahn and Buccianti [54] for more detail):

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{c=1}^{C} b_c(1-\lambda_c) \odot \bigoplus_{\tau=0}^{\infty} \lambda_c^\tau \mathbf{M}_{c,t-\tau} \ \oplus \boldsymbol{\epsilon}_t,$$

where $\mathbf{S}_t, \mathbf{a}, \mathbf{M}_{c,t-\tau}, \boldsymbol{\epsilon}_t$ are respectively the compositions (the whole vectors) of market shares at time $t$, intercept terms, advertising of channel $c$ at time $t-\tau$, and error terms at time $t$. The advantage of this presentation is that it looks like a linear model but with simplicial notations, and then it is easier to see the link with the transformed model. Indeed, the linearization of the MCI model in equation (3.2) is done by a transformation. The ILR (isometric log-ratio transformation) is the best choice because contrary to the log-centered (CLR) transformation usually used in marketing, the error terms of the coordinates are orthonormal and the OLS estimation can be applied without any issue, as explained in Chapter 2, Section 2.2.1. The ILR transformed MCIAd model can be written as follows:

$$S_{j't}^* = a_{j't}^* + \sum_{c=1}^{C}\sum_{\tau=0}^{\infty} \lambda_c^\tau(1-\lambda_c) b_c M_{cj',t-\tau}^* + \epsilon_{j't}^* \qquad \text{for } j' = 1, \ldots, D-1, \tag{3.3}$$

where $S_{j't}^*, M_{cj't}^*, \epsilon_{j't}^*$ are respectively the $j'^{\text{th}}$ ILR coordinate of market shares, advertising of channel $c$ and error terms.

Now, if we think that brands may have different advertising impacts and that cross effects may exist between brands, we can consider the CODA model with a multi-channel carryover effects specification, which we call CODAAd:

$$S_{jt} = \frac{a_j \prod_{c=1}^{C} \prod_{m=1}^{D} Adstock_{cm,t}^{b_{cjm}} \epsilon_{jt}}{\sum_{l=1}^{D} a_l \prod_{c=1}^{C} \prod_{m=1}^{D} Adstock_{cm,t}^{b_{clm}} \epsilon_{lt}} \tag{3.4}$$

where $Adstock_{cjt} = \prod_{\tau=0}^{\infty} M_{cj,t-\tau}^{\lambda_c^\tau (1-\lambda_c)}$ is the adstock at time $t$ of brand $j$ for the advertising channel $c$, as in Equation (3.2), but here in equation (3.4) the adstocks of all brands are directly impacting the market share $S_{jt}$. The simplicial formulation of this model is

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{c=1}^{C} \mathbf{B}_c (1-\lambda_c) \boxdot \bigoplus_{\tau=0}^{\infty} \lambda_c^\tau \mathbf{M}_{c,t-\tau} \oplus \boldsymbol{\epsilon}_t$$

The ILR transformed version of the CODAAd model is a system of $D-1$ equations, one by ILR coordinate $j' = 1, \ldots, D-1$, such as:

$$S_{j't}^* = a_{j't}^* + \sum_{c=1}^{C} \sum_{m'=1}^{D-1} \sum_{\tau=0}^{\infty} \lambda_c^\tau (1-\lambda_c) b_{cj'm'}^* M_{cm't}^* + \epsilon_{j't}^* \qquad \text{for } j' = 1, \ldots, D-1 \tag{3.5}$$

Each equation of this system can be estimated separately by OLS in order to get different variances for error terms. If we fit this model in a unique equation using dummy variables in order to estimate the $b_{cj'm'}^*$ parameters, it implies that one assumes that the variance of error terms are constant across coordinates.

The matrix of parameters $\mathbf{B}_{D\times D}$ can be recovered after inverse ILR transformation of the matrix of parameters $\mathbf{B}_{(D-1)\times(D-1)}^*$, using the following equation: $\mathbf{B} = \mathbf{V}\mathbf{B}^*\mathbf{V}'$, where $\mathbf{V}$ is the balance matrix used for the ILR transformation (see Chapter 1, Section 1.2.5).

### 3.3.2 Optimal advertising carryover parameters

In order to find the optimal adstock parameters, we have tested every combination of $\lambda_c$ for each channel $c$, taking values from 0 to 0.9 with a step of 0.1 (10000 possible combinations). We also imposed a maximum $\tau$ lag of 24 months because after this delay, even with a strong decay parameter, the residual impact becomes negligible ($(1-0.9) \times 0.9^{24} \leq 0.008$). For each combination, we run the MCIAd and the CODAAd models with all the explanatory variables (the adstock functions of outdoor, press, radio and television, the price and the scrapping incentive), and we report the corresponding $R^2$, computed on the ILR transformed models.

The best MCI model is obtained for the decay parameters ($\lambda_{Outdoor} = 0.9; \lambda_{Press} = 0.9; \lambda_{Radio} = 0.8; \lambda_{TV} = 0.8$), and the best CODA model is obtained for ($\lambda_{Outdoor} = 0.9; \lambda_{Press} = 0.9; \lambda_{Radio} = 0; \lambda_{TV} = 0.9$), as shown in Figures A.4 and A.5 in the appendix. As the MCI model is a particular case of the CODA model, and because

no particular $\lambda_{Radio}$ gives better results, we choose to consider the decay parameters obtained in the CODA model (they also give very good results in the MCI model).

These decay parameters suggest that the half life of advertising in outdoor, television and press communications is about 5.6 months, while the advertising through the radio only has a contemporaneous effect within the month of diffusion. We demonstrate in the appendix A.3.5 how the half life can be computed for attraction models with carryover effects. The half life of advertising in outdoor, television and press may seem to be quite strong, but it is not surprising in a durable good market.

## 3.4 Final model specification and results

### 3.4.1 Comparison of model specifications

We want to explain brands' market shares by the advertising budgets[3] in television, outdoor, radio and press, the average prices and the scrapping incentive. Several models can be considered, with more or less complexity. In this section, we compare the MCI model, the Dirichlet model (DIR, see Chapter 1, Section 1.2.4) and the CODA model. The last two models specify brand-specific parameters for the marketing explanatory variables, while the first one does not. The CODA model is the only model specifying additionally cross effects between brands. All these models can be expressed in attraction formulation: the market share of brand $j$ is defined as the relative attraction of brand $j$, that is the attraction of brand $j$ divided by the sum of attractions of all brands of the market (see Chapter 1, Section 1.3.2 for details).

MCI, CODA and DIR models are fitted with and without adstock variables in order to assess the relevance of the carryover effects. Note that in the case of the DIRAd model, adstock variables are computed additively whereas they are computed multiplicatively in the MCIAd and CODAAd models[4].

In order to enhance the importance of explanatory variables, these models are compared to naive models, called Constant MCI and Constant DIR models, where only brand-specific intercepts are used as explanatory variables. In the case of the MCI model, it is equivalent to estimate market shares to their closed geometric means (called "center" in compositional data analysis).

Models are adjusted on the period from 2005/01 to 2014/12 (in sample, $T = 120$) and are validated on the period from 2015/01 to 2015/08 (out of sample, $T' = 8$). The period 2003-2004 has been sacrificed for the computation of weighted media which uses 3 lags (from 2003/01 to 2003/03) and adstock variables which uses 21 lags (from 2003/04 to 2004/12).

We compare the considered models according to their goodness of fit (in sample) and their prediction accuracy (out of sample), reported in Table 3.1. The adjusted $R^2$ and the non adjusted $R^2$ are computed respectively on the in-sample and out-of-sample ILR coordinates of market shares, in the case of the MCI and CODA models. Then, they reflect the quality of the model on log ratios of shares, giving more importance to the relative error than to the absolute error. The compositional $R^2_T$ (R-squared based on the total variance, see details in the appendix A.1.4) is computed for all models on the market shares directly, as the RMSE. Nevertheless, the $R^2_T$ is a measure of quality on the log ratios of shares, whereas the RMSE gives the same importance to an error of 1 percentage point made on a share of 1% or on a share of 50%. We conclude from Table 3.1 that the CODAAd model is the best model in terms of goodness of fit according to

[3]When the advertising budgets are null, they are replaced by one euro, as explained in A.1.3 in the appendix.

[4]Multiplicative adstock functions in equations (3.2) and (3.4) correspond to additive adstock functions in the ILR transformed models (3.3) and (3.5).

all quality measures. This suggests that the carryover effect of the advertising does exist and that the advertising of each brand has potentially a different impact on the other brands.

Concerning the prediction accuracy, we can notice in Figure 3.8 that the market shares from January to August 2015 are not varying a lot and are very close to the center of the the data from 2005 to 2014. Thus, the result is that, by chance, the Constant MCI and the Constant DIR models give very good results in terms of RMSE. The CODAAd model still is the best model of the MCI family[5] according to the $R^2$ on ILR coordinates. Note that the maximum $R^2_T$ for prediction is higher than 1 for the CODA model, meaning that the variability of the predicted log ratios is larger than the variability of the real log ratios.

The complexity of each model can be appreciated through the number of parameters to be estimated. In the case of the models of the MCI family, the number of parameters to be estimated is lower than the number of parameters in the attraction form of the model, thanks to the use of the ILR transformation, and the number of observations to consider is $T(D-1)$. Then, in Table 3.1, we see that CODA and CODAAd require the estimation of 51 parameters for 360 in-sample ILR observations, which is feasible. In the case of the Dirichlet family models, the number of parameters to estimate is equal to the final number of parameters.

Table 3.1: Market share models accuracy (in sample: 2005-2014, out of sample: 2015)

| | | In sample | | | | | Out of sample | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Param | Adj. $R^2$ | $R^2$ | Adj. $R^2_T$ | $R^2_T$ | $RMSE$ | $R^2$ | $R^2_T$ | $RMSE$ |
| Constant MCI | 3 | 0.589 | 0.591 | 0 | 0 | 0.035 | 0.853 | 0 | **0.018** |
| Constant DIR | 4 | - | - | - | 0 | 0.035 | - | 0 | 0.019 |
| MCI | 11 | 0.727 | 0.734 | 0.315 | 0.350 | 0.029 | 0.835 | 0.385 | 0.022 |
| MCIAd | 11 | 0.786 | 0.792 | 0.464 | 0.491 | 0.025 | 0.722 | 0.139 | 0.026 |
| DIR | 28 | - | - | - | 0.422 | 0.024 | - | 0.478 | 0.029 |
| DIRAd | 28 | - | - | - | 0.561 | 0.021 | - | 0.640 | 0.025 |
| CODA | 51 | 0.789 | 0.819 | 0.488 | 0.557 | 0.024 | 0.623 | **1.913** | 0.029 |
| CODAAd | 51 | **0.859** | **0.879** | **0.657** | **0.703** | **0.019** | **0.858** | 0.585 | 0.026 |

Adj. $R^2$ and $R^2$ computed on the ILR coordinates (Dirichlet models not concerned).
Adj. $R^2_T$ and $R^2_T$ computed on the compositions of market shares.
$RMSE$ computed on the market shares.

The observed, fitted and predicted market shares of the different models are represented in Figure 3.8. We can notice that taking into account the carryover effect of advertising in the MCI model allows to give a better fit of the lowest point of Citroën in 2008. The consideration of the cross effects between brands (CODA model) results in a better adjustment of Renault's market shares in 2005-2006. Finally, the combination of cross effects and adstock specification of advertising (CODAAd) gives very satisfying fitted market shares during the whole period 2005-2015.

[5]We call "MCI family" the MCI and CODA type models, in opposition to Dirichlet models.

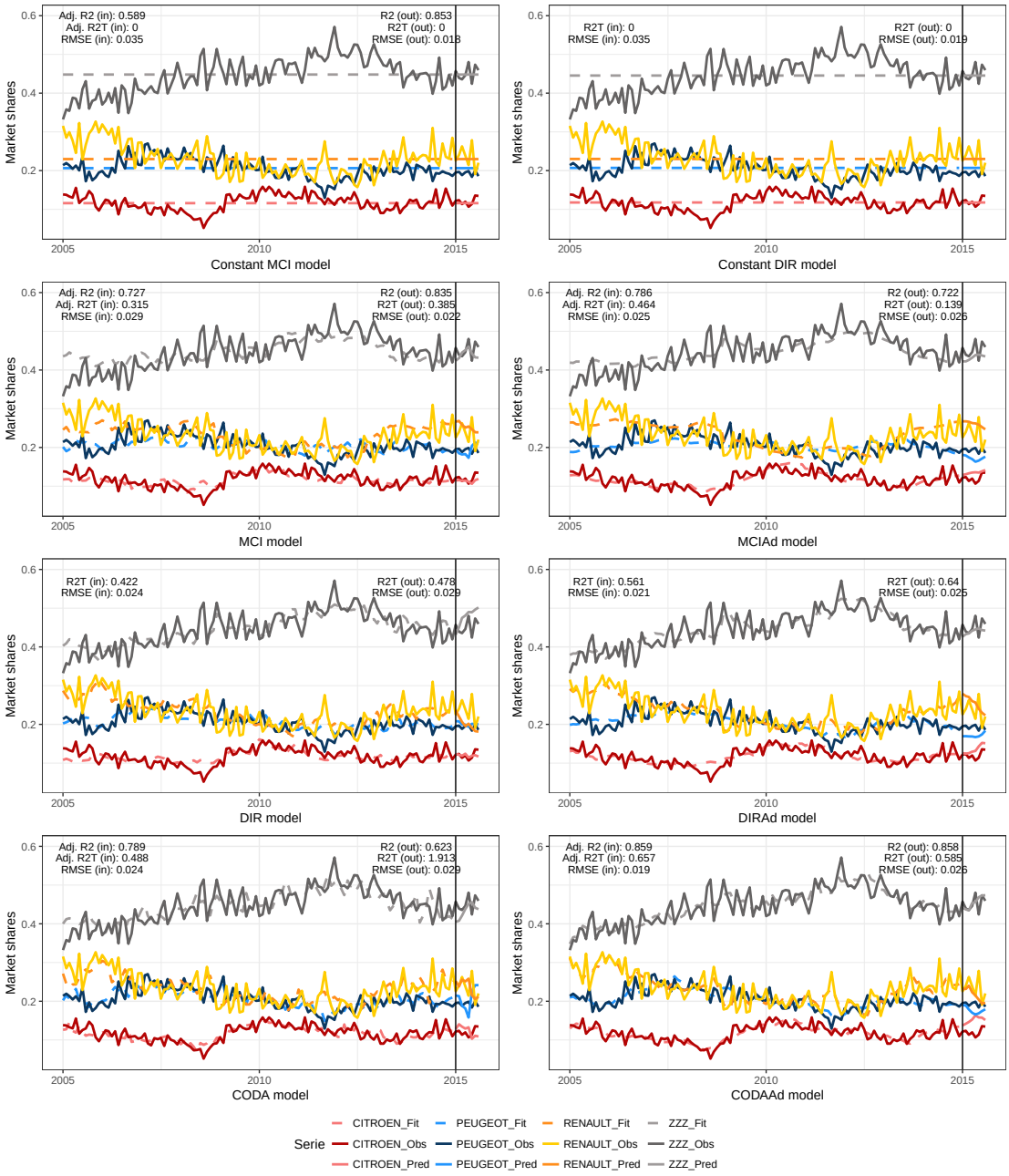


Figure 3.8: Observed, fitted and predicted market shares and accuracy measures

Moreover, using an adapted Fisher test (see Chapter 2, Section 2.2.2), we have tested whether the complexity of the "unconstrained" CODAAd gives significantly better results that the "constrained" model MCIAd where the explanatory variables impacts are the same for all brands, without cross effect. The Fisher statistic is equal to 5.51

while the 95% quantile is equal to 0.44, then we can largely reject H0 and conclude that CODAAd is better than MCIAd. We have also compared the CODAAd model and the constrained model where brand specific and cross effects parameters are defined for advertising channel but not for price, but the conclusion is the same (the Fisher statistics equals 1.76 and the 95% quantile equals 0.52).

The CODAAd model is then considered to be the best specification for the modeling of advertising impact on market shares.

### 3.4.2 Residual diagnostic

Let us check that the CODAAd model residuals have good features. The residual diagnostic is done on the ILR residuals, for the CODAAd model and for the MCIAd model in order to have a benchmark. The first graph of Figure 3.9 suggests that there is no heteroscedasticity problem in the CODAAd model, while it is less clear for the MCIAd model. ILR residuals look normal according to the QQ-plot even if the tails are a little bit heavier. Finally, the graphs of the last column allow to conclude that the CODAAd model residuals are not autocorrelated, as confirmed by the Breusch-Godfrey test at order 1, contrary to those of the MCIAd model.

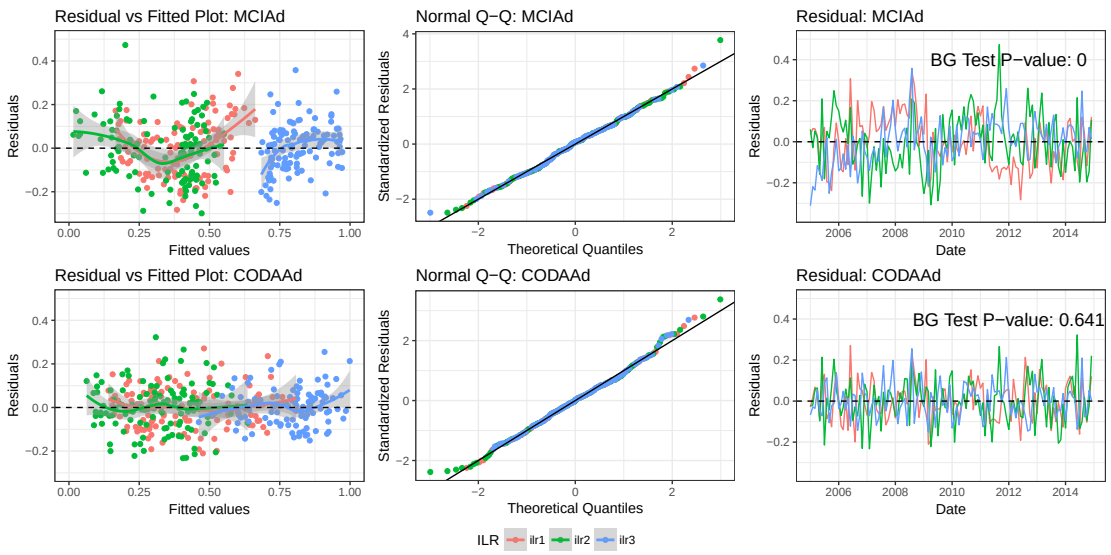


Figure 3.9: ILR residuals diagnostic of models for MCIAd and CODAAd models

### 3.4.3 Confidence and prediction ellipsoids

Now, let us look more precisely at the prediction power of the CODAAd model. It is possible to construct confidence and prediction intervals for market shares using the ellipsoid of confidence and the ellipsoid of prediction of the ILR coordinates, as we demonstrate in Appendix A.3.7. Here it is important to estimate the CODAAd model as proposed in

compositional data analysis with the ILR transformation, instead of the estimation in a unique equation with dummy variables proposed by Nakanishi and Cooper [52]. Indeed, while it does not change the predicted values, it does change the prediction intervals because the estimated values of standard deviations of parameters are different. We have run a test of equality of variance on residuals across ILR coordinates, and we conclude that the variance of ILR error terms should not be considered as equal, which reinforces the choice of this estimation method.

Figure 3.10 represents the confidence and prediction ellipsoids for the Citroën, Peugeot, Renault and Others market shares in January 2015, in the simplex $\mathcal{S}^4$.

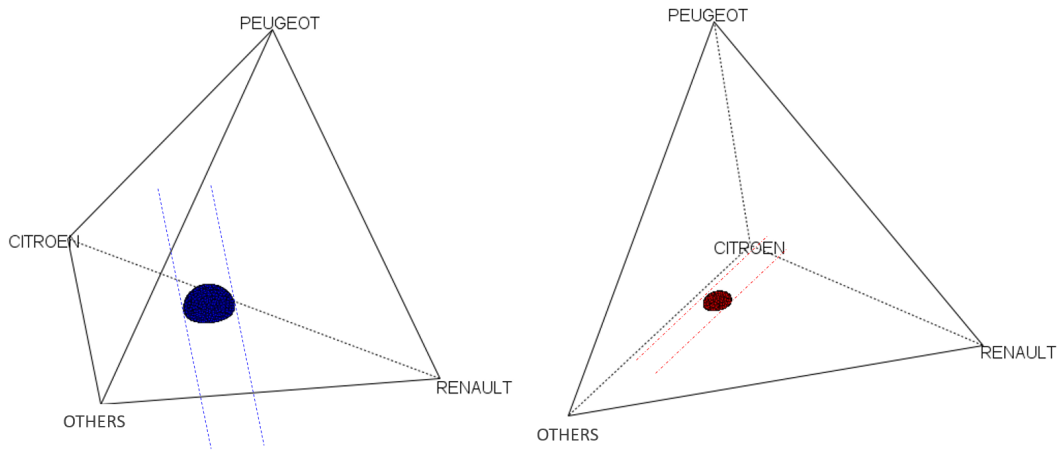


Figure 3.10: Prediction (left) and confidence (right) ellipsoids at 95% for market shares in January 2015

A 95% marginal prediction interval for each share is obtained taking the minimum and the maximum values of the projection of the ellipsoid on the corresponding simplex edge in $\mathcal{S}^4$ (see for example the minimum and maximum for Renault delimited by dotted lines in Figure 3.10). These maximum and minimum points are computed for each prediction time and are represented in Figure 3.11. We observe that the true market shares (in grey) are always in the 95% prediction intervals and very close to the predictions (black). However, the CODAAd model tends to overestimate the Citroën market shares and underestimate the Peugeot market shares in 2015. The Renault's market share in June is surprisingly high at the expense of the Others market share, but the model does not succeed in reflecting this temporary change, while the predictions of the rest of the period are very accurate.

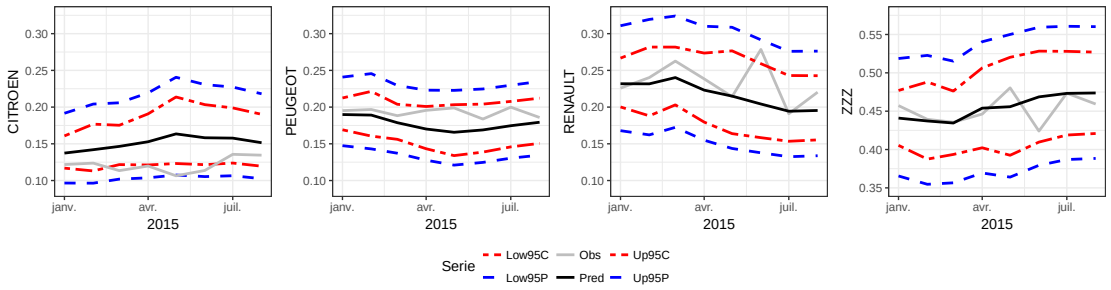


Figure 3.11: Observed (grey) and predicted (black) market shares with 95% confidence (red) and prediction (blue) intervals

### 3.4.4 Advertising elasticity of market shares

We now focus on the explanatory power of the CODAAd model. According to the analysis of variance, all explanatory variables are significant in this model, except the scrapping incentive (see ANOVA results in Table A.5 in the appendix). Note that the scrapping incentive was generally significant in other models, suggesting that the specification of adstock and cross-effects are sufficient to take into account the perturbation of market shares during this special period of governmental incentive. Moreover, if $SI$ is not included as explanatory variable in CODAAd, this model stays the best according to its accuracy measures which remain almost the same.

Then, for the purpose of interpreting the model, we fit the CODAAd model without the scrapping incentive, on the total period from 2005 to 2015, and we call the resulting model CODAAd_SI. All explanatory variables are strongly significant in this model according to the ANOVA (see Table 3.2) and the accuracy measures are fully satisfactory (see Table 3.3).

Table 3.2: Analysis of variance table for CODAAd_SI model

| | Df | Pillai | approx F | num Df | den Df | Pr(>F) |
|---|---|---|---|---|---|---|
| (Intercept) | 1 | 0.99 | 3902.2 | 3 | 102 | $< 2.2e^{-16***}$ |
| ilr(OutdoorAd) | 3 | 1.34 | 28.1 | 9 | 312 | $< 2.2e^{-16***}$ |
| ilr(PressAd) | 3 | 0.68 | 10.2 | 9 | 312 | $9.384e^{-14***}$ |
| ilr(RadioAd) | 3 | 0.35 | 4.5 | 9 | 312 | $1.374e^{-05***}$ |
| ilr(TVAd) | 3 | 0.58 | 8.3 | 9 | 312 | $4.206e^{-11***}$ |
| ilr(Price) | 3 | 0.23 | 2.9 | 9 | 312 | $0.002926^{**}$ |
| Residuals | 104 | | | | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ilr(X) denotes the vector of ILR coordinates of the variable X.

The short-term elasticities of market share $S_j$ at time $t$ relative to the brand $l$ media investments in channel $c$ at time $t$, $\check{X}_{clt}$, can be computed according to the following

Table 3.3: Accuracy of the CODAAd_SI model (from 01-2005 to 08-2015)

| | In sample | | | | |
|---|---|---|---|---|---|
| Model | Adj. $R^2$ | $R^2$ | Adj. $R^2_T$ | $R^2_T$ | $RMSE$ |
| CODAAd_SI | 0.856 | 0.874 | 0.639 | 0.682 | 0.019 |

formula (see Chapter 2, Section 2.3.2 for details):

$$elast(S_{jt}, \check{X}_{clt}) = \frac{\partial \log S_{jt}}{\partial \log \check{X}_{clt}} = \frac{\partial S_{jt}/S_{jt}}{\partial \check{X}_{clt}/\check{X}_{clt}} = (1-\lambda_c)(b_{cjl} - \sum_{m=1}^{D} S_{mt} b_{cml})$$

We talk about short-term elasticity because it only measures the relative impact of $\check{X}_{clt}$ on $S_{jt}$ at time $t$, ignoring the future impacts on $S_{jt+1}, S_{jt+2}, \ldots$. As we are working on market shares, it is not possible to summarize the overall long term effect of an evolution of $\check{X}_{clt}$ on the $S_{jt'}$ for $t' \geq t$. However, it is of course possible to compute the following elasticities:

$$elast(S_{jt+\tau}, \check{X}_{clt}) = \lambda_c^{\tau}(1-\lambda_c)(b_{cjl} - \sum_{m=1}^{D} S_{mt+\tau} b_{cml})$$

Table 3.4 presents the average short-term elasticities of market shares relative to outdoor, press, radio and television. Direct elasticities are on the diagonal and cross elasticities are extra diagonal. We remark that the largest elasticities (in bold) are often off-diagonal, which highlights the importance of cross effects between brands. For example, if Citroën increases by 1% its TV advertising budget, its market share increases on average by 0.0192%, but if Others increases by 1% its TV advertising budget, the Citroën market share decreases on average by 0.0372%.

For outdoor, press and television, we can compute "long-term elasticities", that is the elasticities of market shares relative to these channels' adstocks, but care must be taken for the interpretation: they correspond to the relative impact of market shares for a relative change in the adstock variable, which can come from changes in one or several lags of the media investments. Long-term elasticities for these three channels are equal to ten times the short-term elasticities (short-term elasticities divided by $1-\lambda$, with $\lambda = 0.9$). Note that in the case of radio, as the estimated retention rate is equal to zero, and in the case of price for which we assumed contemporaneous effect only, there is no long-term elasticities.

We can see in Figure 3.12 that the direct elasticities of the four channels are quite constant across time, which means that the average elasticities presented in Table 3.4 are good indicators of the real elasticity at time $t$. Moreover, the Renault's elasticity of market share to its own outdoor and television advertising are very close (around 0.018) but an increase of 1% of TV represents less money than an increase of 1% in outdoor, as on average the budget of outdoor is much larger than the TV's budget. Note that Citroën has a particular profile regarding advertising elasticities: it has a very positive direct elasticity for the radio while other brands have a negative direct elasticity, and negative

outdoor and press direct elasticities while other brands have positive direct elasticities. Peugeot has surprisingly quite low outdoor and television elasticities compared to its direct competitor Renault.

Table 3.4: Advertising short term elasticities of market shares (CODAAd_SI)

| | Outdoor_Citroën | Outdoor_Peugeot | Outdoor_Renault | Outdoor_Others |
|---|---|---|---|---|
| S_Citroën | -0.0042 | -0.0372 | -0.0014 | **0.0428** |
| S_Peugeot | -0.0133 | 0.0076 | -0.0064 | 0.0121 |
| S_Renault | 0.0088 | 0.0354 | 0.0183 | *-0.0624* |
| S_Others | 0.0027 | -0.0058 | -0.0047 | 0.0078 |
| | Press_Citroën | Press_Peugeot | Press_Renault | Press_Others |
| S_Citroën | -0.0092 | -0.0008 | -0.0039 | **0.0139** |
| S_Peugeot | 0.0053 | 0.0110 | 0.0027 | -0.0190 |
| S_Renault | -0.0007 | *-0.0212* | 0.0128 | 0.0091 |
| S_Others | 0.0012 | 0.0049 | -0.0051 | -0.0010 |
| | Radio_Citroën | Radio_Peugeot | Radio_Renault | Radio_Others |
| S_Citroën | 0.0639 | 0.0011 | -0.0140 | -0.0510 |
| S_Peugeot | 0.0085 | -0.0031 | 0.0308 | -0.0362 |
| S_Renault | *-0.0768* | 0.0099 | -0.0124 | **0.0793** |
| S_Others | 0.0083 | -0.0033 | -0.0021 | -0.0030 |
| | TV_Citroën | TV_Peugeot | TV_Renault | TV_Others |
| S_Citroën | **0.0192** | 0.0024 | 0.0157 | *-0.0372* |
| S_Peugeot | -0.0014 | 0.0056 | 0.0014 | -0.0056 |
| S_Renault | -0.0091 | 0.0016 | 0.0184 | -0.0109 |
| S_Others | -0.0018 | -0.0036 | -0.0132 | 0.0185 |

The most positive (resp. negative) relative impact due to an increase in advertising is in bold (resp. italic).

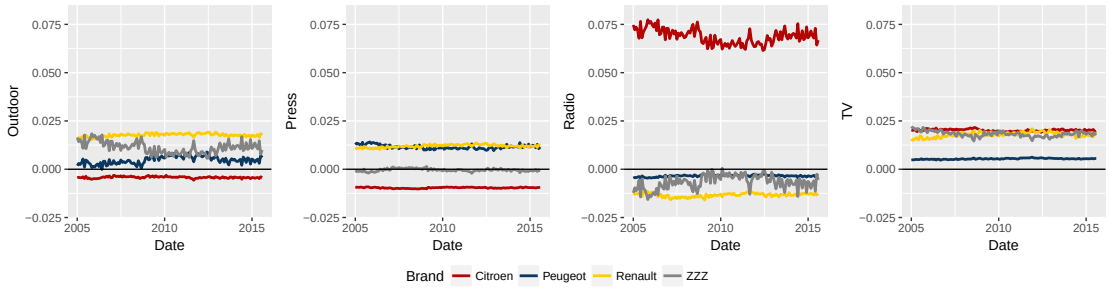


Figure 3.12: Short term direct elasticities of channels by brand

## 3.5 Conclusion

The aim of this chapter is to measure the impact of advertising investments through different channels (outdoor, press, radio and television), on brands' market shares in the French automobile market, from 2005 to 2015. We focus on the main segment of this market, namely the B segment, and on the three leaders of this segment: Citroën, Peugeot and Renault, aggregating the other brands in a group.

After analyzing this market in terms of absolute and relative sales and media investments for the different brands, we emphasize the importance of taking into account the competition on the one hand, and the advertising carryover effect on the other hand. For this purpose, we build a multi-channel attraction model with carryover effects, called CODAAd model (for CODA model with adstock), combining the classical Koyck model and the CODA model presented in previous chapters. We stress the positive aspects of using an isometric log-ratio (ILR) transformation, coming from the compositional data analysis (CODA) literature, to estimate this type of models, instead of the usual centered log-ratio (CLR) transformation used in marketing. We explain how to determine the carryover parameters $\lambda$ for several channels in a simultaneous way, and we conclude that outdoor, press and television advertisements have a large retention rate implying an advertising half life of 5.6 months, which seems realistic for a durable and expensive good such as automobile. On the contrary, the radio advertising appears to have only contemporaneous effect on market shares in the B segment of the automobile market.

Several model specifications are compared: with or without explanatory variables (constant models), with or without cross effects, with or without adstock variables, using a model belonging to the Dirichlet family or to the MCI family. According to goodness-of-fit (on the 2005-2014 period) and prediction accuracy (on 2015) measures, the CODAAd model, is considered to be the best model for our purpose. An adapted Fisher test confirms that the inclusion of cross effects improves significantly the model. The residual diagnostic suggests that the CODAAd model has good properties. Moreover, we draw the 95% confidence and prediction ellipsoids in the space of market shares (i.e. in the simplex), and we derive from that the 95% confidence and prediction intervals for the prediction period, in 2015.

In order to interpret the impact of each channel on brands' market shares, we compute direct and cross advertising elasticities. While television direct elasticities are positive for all brands, this is not the case for the other channels. Citroën has a different advertising impact profile from the other considered brands: the model suggests that Citroën can increase its market share diminishing its outdoor and press advertising budget in favor of the radio and television advertisements, whereas for Peugeot, Renault and the group of others, it suggests to increase the investments in outdoor, press and television and to reduce the radio communication.

Further research should be done in order to evaluate the elasticities significance, in particular for cross elasticities. A standardized interpretation could be proposed in order to be able to compare the marginal impact of channels whose total expenses are not of the same order of magnitude. Threshold effects in the advertising elasticities may also be considered. Moreover, the geometric distributed lagged adstock function of the Koyck model can be seen as too restrictive, and more flexible functions could be investigated. Finally, these elasticities can be useful for advertising budgeting optimization, like in the Dorfman-Steiner theorem [12], but the optimization problem has to be adapted for a multi-channel attraction model.

# Chapter 4

# Further directions

This last chapter of the thesis aims to list further directions to explore in order to answer the question "What is the impact of media investments on the brands' market shares in the automobile market?". For example, additional models could be considered, the distributional assumption of the proposed market share models could be questioned, and the interpretation of market share models could be improved. Finally, the ultimate goal of measuring the impact of each advertising channel is to optimize the advertising budget and the marketing mix. We show that the famous Dorfman-Steiner theorem can be generalized for the multi-channel and competitive case.

## 4.1 Additional models

During the thesis, we have estimated various market share models only on the B segment of the French automobile market, focusing on the main brands on this market. Of course, it would be very interesting to apply these models to other data. Below, we give a non exhaustive list of additional modelings which can be considered in order to get a better understanding of the impact of media investments on market shares:

- We could apply the CODA model to other segments (A, C, D and E), in order to compare the significance of the impact of each advertising channel, their carryover effects, and their cross effects between brands. Advertising may have a higher impact on larger and more expensive vehicles (D and E segments).
- We could choose to focus on other brands than Citroën, Dacia, Nissan, Peugeot and Renault, or to consider more brands simultaneously, but it would imply to have more observations because the number of parameters to be estimated is rising dramatically with the number of brands ($D$) considered.
- We could imagine that market share models are useful to analyze the impact of advertising at a micro level, on each model of vehicle for example (instead of at a macro level, on each brand within a segment), but it requires again a large number of observations, at a weekly frequency if possible, to avoid zero problems.

We also think that a nested compositional model could be built, where segments are the nests, and vehicle (or brands) within each segment are the alternatives. Nested multinomial logit models (NMNL) have been developed in econometrics for individual data. In a working paper, Fry and Chong [19] start to explore the use of NMNL for (aggregated) compositional data. Compositional contingency tables may be a fruitful avenue for considering this aspect, where rows of the contingency table can be segments and columns can be brands (see Greenacre and Lewi [24] and Egozcue et al. [14]).

Other models could also be interesting, considering sales in volumes instead of market shares:

- The sales of a particular vehicle or brand can be modeled as a function of its advertising composition in terms of channels: the explanatory variable is then the vector of proportions of each channel (outdoor, press, radio, television, ...) in the total advertising budget (or in the the total adstock function) at time $t$. This type of model would allow to focus on the marketing mix impact and on the "winning combinations" of media channels.

- We could also consider a panel data model, where individuals are vehicles or brands, observed across time, with potential interactions between advertising budgets of each vehicle or brand.

It would also be valuable to observe the actual purchase prices instead of the catalogue prices, or to observe promotions, and to measure properly the media investments in internet and the impact of social media.

Finally, one of our starting assumptions is that the economic context is playing a role for customers in the choice of the vehicle segment. We are currently analyzing this impact in an article: Morais, J., and Thomas-Agnan, C. Impact of economic conditions on automobile market segment shares: a compositional approach. *CSBIGS* (2018). In progress.

## 4.2 Assumptions of the models

Concerning the assumptions of the CODA model we use, further research could be done on the following aspects:

- The CODA and the MCI models assume that the transformed error terms (in CLR or ILR) are normally distributed. Even if this condition seems to be satisfied in our example, it may not be the case. When appropriate, a Box-Cox transformation of the initial market shares can be used until reaching the normality of error terms. Other distributions can also be investigated: for instance using more flexible versions of the Dirichlet distribution (see for example Monti et al. [44]) or a Student distribution for transformed error terms (see for example Katz and King [32]).

- In Chapter 3, we introduce the carryover effect of advertising through an adstock function, inspired from the classical Koyck model. This geometric lag specification is quite simple and suggests that the impact of advertising across time is constantly decreasing. Other types of adstock functions could be tested (see Joy [31] for a brief review) in order to allow for threshold effects for example (increasing impact until a certain threshold, and then decreasing impact after this threshold). However, more complex functions go hand in hand with more complicated estimation of their own parameters.

- Concerning again the chosen adstock function, the way to determine the decay parameters $\lambda$ could be more precise with a finer grid of possible values (with a step of 0.01 instead of 0.1 for example), and possibly more lags. We could also make the choice to estimate the decay parameters separately for each channel and without the rest of the explanatory variables in order to isolate the pure correlation between one advertising channel and the market shares.

## 4.3 Interpretation

In Chapter 2, we emphasize the difficulty to interpret properly market share models taking into account the constraints applicable to share data. We highlight the benefits of elasticities which are consistent with the relativeness of market shares. However, two main points need to be deeply examined:

- The significance of elasticities should be analysed, especially for cross elasticities which seem to vary a lot across time. Standard errors for elasticities can be computed by bootstrap (see Green et al. [23] for example) or by the Delta method. Presently, we have only checked the distribution of direct elasticities to see if they are sufficiently stable in order to use the average elasticities to summarize the impact of media investments on market shares.

- Elasticities measure the relative impact on market shares of a relative variation in an advertising budget. For example, it allows to compare the relative impact of an increase of 1% in the Renault's television budget on the Renault's market share, and the relative impact of an increase of 1% in the Peugeot's television budget on the Peugeot's market share. However, as the level of advertising budget is not equivalent in all channels, the increase of 1% in the Renault's television budget and the increase of 1% in the Renault's outdoor budget do not represent the same amount of money. But in the end, manufacturers need to optimize the allocation of their total advertising budget between the different channels. Thus, we can imagine a standardized version of elasticities, based on standardized explanatory variables or standardized coefficients.

## 4.4 Advertising budgeting optimization

The final aim of this research is to optimize the advertising budget of car manufacturers, considering the competitors' actions and the multi-channel case. Advertising budgeting optimization has been studied in marketing, firstly by Dorfman and Steiner [12]. They proved that the solution to the budgeting optimization problem is linked to the advertising elasticities and the price elasticities by a simple formula. The so-called "Dorfman-Steiner theorem" has been developped in a simple case, focusing on one brand (or product) and one global advertising budget. Then, Levy and Simon [38] generalize it to the dynamic case where advertising has a multi-period impact. Let us now start to generalize the Dorfman-Steiner theorem for the multi-channel and competitive case. The notations used are in Table 4.1.

Table 4.1: Notations used for advertising budgeting optimization

| | |
|---|---|
| Profit | $\Pi = G - M$ |
| Gross revenue | $G = Q(P - C)$ |
| Quantity sold | $Q$ |
| Unit price | $P$ |
| Unit cost | $C$ |
| Margin rate | $\tau = (P - C)/P$ |
| Media investment (advertising) | $M$ |
| Advertising elasticity | $e_M$ |
| Price elasticity | $e_P$ |
| Index of the $j^{\text{th}}$ brand | $j$ |

### 4.4.1 The Dorfman-Steiner theorem

Dorfman and Steiner [12] demonstrate how to simultaneously optimize advertising budget and price, or to optimize only advertising budget considering fixed prices, in a very simple demand function framework.

#### Joint optimization of advertising budget and price

The quantity sold $Q$ of a given product (or brand) is assumed to be a function of price and advertising, such that $Q = f(P, M)$. Then, the change in $Q$ when $P$ and $M$ change respectively by $dP$ and $dM$ is

$$dQ = \frac{\partial Q}{\partial P} dP + \frac{\partial Q}{\partial M} dM$$

The changes $dP$ and $dM$ can compensate each other such that the final quantity $Q$ is not impacted:

$$dQ = 0 \quad \Leftrightarrow \quad dP = -\frac{\partial Q/\partial M}{\partial Q/\partial P} dM$$

Then, the profit function being equal to $\Pi = Q(P - C) - M$, the net effect on profit of the changes $dP$ and $dM$ such that $Q$, and $C$, are stable is

$$d\Pi = QdQ - dM = -Q\frac{\partial Q/\partial M}{\partial Q/\partial P}dM - dM,$$

and the profit is maximized in the following conditions:

$$\begin{aligned} Max\Pi \quad &\Leftrightarrow \quad \left(Q\frac{\frac{\partial Q}{\partial M}}{\frac{\partial Q}{\partial P}} + 1\right) = 0 \text{ if M>0} \quad \text{or} \quad \left(Q\frac{\frac{\partial Q}{\partial M}}{\frac{\partial Q}{\partial P}} + 1\right) \geq 0 \text{ if M=0} \\ &\Leftrightarrow \quad P\frac{\partial Q}{\partial M} = -\frac{\partial Q}{\partial P}\frac{P}{Q} \\ &\Leftrightarrow \quad e_M\frac{PQ}{M} = -e_P \\ &\Leftrightarrow \quad \frac{M}{PQ} = \frac{e_M}{-e_P} \end{aligned}$$

where $e_M, e_P$ are the advertising and price elasticities of demand. In this framework, the ratio of optimal advertising and price is a function of the quantity multiplied by the ratio of advertising and price elasticities.

**Optimal advertising with fixed price**

In the case of fixed price, the optimal advertising is simply a function of advertising elasticity and gross product:

$$\begin{aligned} Max\Pi = Q(P - C) - M \Leftrightarrow \quad & dQP - dQC - dM = 0 \\ \Leftrightarrow \quad & P - C = \frac{dM}{dQ} \quad \Leftrightarrow \quad \frac{Q}{M}(P - C) = \frac{dM}{dQ}\frac{Q}{M} \\ \Leftrightarrow \quad & e_M = \frac{M}{Q(P - C)} \quad \Leftrightarrow \quad M = Ge_M \end{aligned}$$

Wright, in an article called "A new theorem for optimizing the advertising budget" [71], actually considers this case, specifying the demand function as $Q = kM^{e_M}$. Then,

$$\begin{aligned} Max\Pi = (P - C)Q - M = (P - C)kM^{e_M} - M \quad &\Leftrightarrow \quad \frac{\partial \Pi}{\partial M} = 0 \\ &\Leftrightarrow \quad e_M = \frac{M}{G} \quad \Leftrightarrow \quad M = Ge_M \end{aligned}$$

The result is the same than in Dorfman and Steiner [12].

### 4.4.2 A dynamic version of the Dorfman-Steiner theorem

Levy and Simon [38] propose "A Generalization That Makes Useful the Dorfman-Steiner Theorem with Respect to Advertising". Indeed, they include the fact that advertising

has an impact on several periods after the advertising diffusion, according to a customer-retention factor and a cost-of-capital discount factor.

They consider the following function for the gross revenue:

$$G_t = (P_t - C_t)Q_t = bG_{t-1} + f(M_t),$$

where the gross revenue at time $t$ is a function of the accumulated carryover sales due to previous advertising and of the sales due to the advertising at time $t$. The retention rate is denoted by $b = 1 -$ decay rate. Then, we get:

$$f(M_t) = G_t - bG_{t-1} = \Delta G$$

Assuming stationarity, we have $b(G_t - bG_{t-1}) = b\Delta G$.

They define the net present value of the present and future effects of $M_t$ as:

$$\begin{aligned} NPV(M_t) &= \Delta G + \Delta G bd + \Delta G b^2 d^2 + \cdots - M_t \\ &= f(M_t) + bdf(M_t) + b^2 d^2 f(M_t) + \cdots - M_t \\ &= f(M_t)[1 + bd + b^2 d^2 + \ldots] - M_t, \end{aligned}$$

where $d$ is the money discount factor. If $b$ is the same at all periods, then:

$$\begin{aligned} NPV(M_t) &= f(M_t)\left(\frac{1}{1-bd}\right) - M_t \\ &= G_t - bG_{t-1}\left(\frac{1}{1-bd}\right) - M_t \\ MaxNPV(M_t) \quad &\Leftrightarrow \quad \frac{\partial NPV(M_t)}{\partial M_t} = \left(\frac{1}{1-bd}\right)\frac{\partial f(M_t)}{\partial M_t} - 1 = 0 \\ &\Leftrightarrow \quad \frac{\partial f(M_t)}{\partial M_t} = 1 - bd \end{aligned}$$

They advise to invest in advertising until reaching the point such that $\frac{\partial f(M_t)}{\partial M_t} = 1 - bd$. Note that in Dorfman and Steiner, $b = 0$, and then $\frac{\partial f(M_t)}{\partial M_t} = 1$ (it is a particular case).

### 4.4.3 A multi-channel and competitive version of the Dorfman-Steiner theorem

In the classical Dorfman-Steiner theorem presented above, the profit function of a brand $j$ (or product) is defined as:

$$\Pi_j = G_j - M_j = Q_j(P_j - C_j) - M_j$$

where $Q_j = k_j M_j^{e_{Mj}}$ in Wright [71].

Let us define a specification for $Q_j$ corresponding to the models we use in the previous chapters. $Q_j$ is a function of the total quantity of the market $Q$ and of the market share

$S_j$, the first one being a linear function of media investments, and the second one being modeled with a CODA model:

$$\mathbb{E}Q_j = \mathbb{E}Q \times \mathbb{E}^{\oplus}S_j = (\alpha + \sum_{l=1}^{D}\sum_{k=1}^{K}\beta_{kl}M_{kl}) \times \frac{a_j \prod_{l=1}^{D} M_l^{b_{jl}}}{\sum_{m=1}^{D} a_m \prod_{l=1}^{D} M_l^{b_{ml}}}$$

To simplify, $\mathbb{E}Q_j, \mathbb{E}Q, \mathbb{E}^{\oplus}S_j$ are replaced below by they estimations $\hat{Q}_j, \hat{Q}, \hat{S}_j$, the number of brands considered is $D = 2$, and the number of advertising channel is $K = 1$. The profit of brand $j$ is then:

$$\begin{aligned}\Pi_j &= \tau_j P_j \hat{Q}_j - M_j = \tau_j P_j \hat{Q}\hat{S}_j - M_j \\ &= \tau_j P_j (\hat{\alpha} + \hat{\beta}_1 M_1 + \hat{\beta}_2 M_2) \frac{\hat{a}_j M_1^{\hat{b}_{j1}} M_2^{\hat{b}_{j2}}}{\hat{a}_1 M_1^{\hat{b}_{11}} M_2^{\hat{b}_{12}} + \hat{a}_2 M_1^{\hat{b}_{21}} M_2^{\hat{b}_{22}}} - M_j\end{aligned}$$

Let us maximize the profit of brand 1 with respect to its own media investment $M_1$ for example:

$$\begin{aligned}\frac{\partial \Pi_1}{\partial M_1} &= \tau_1 P_1 (\frac{\partial Q}{\partial M_1} S_1 + \frac{\partial S_1}{\partial M_1} Q) - 1 \\ &= \tau_1 P_1 (\hat{\beta}_1 S_1 + \frac{1}{M_1} Q(b_{11}S_1 - b_{11}S_1^2 - b_{21}S_1 S_2)) - 1 \\ &= G_1 (\frac{1}{Q}\hat{\beta}_1 + \frac{1}{M_1} e_{11}) - 1 \\ &= G_1 \frac{1}{M_1}(\epsilon_1 + e_{11}) - 1 \\ \frac{\partial \Pi_1}{\partial M_1} = 0 \quad &\Leftrightarrow \quad \frac{M_1}{G_1} = \epsilon_1 + e_{11} \quad \Leftrightarrow \quad M_1 = G_1(\epsilon_1 + e_{11})\end{aligned}$$

where $G_j = \tau_j P_j Q S_j$ is the gross revenue of brand $j$, $\epsilon_j = \frac{\partial Q}{\partial M_j}\frac{M_j}{Q} = \hat{\beta}_j \frac{M_j}{Q}$ is the advertising elasticity of the total market $Q$ relative to the media investment of brand $j$, and $e_{jl} = \frac{\partial S_j}{\partial M_l}\frac{M_l}{S_j} = b_{jl} - \sum_{m=1}^{D} b_{ml} S_m$ is the advertising elasticity of the market share $S_j$ relative to the media investment of brand $l$.

We thus find a similar equation to the Dorfman-Steiner theorem. However, here we need to take into account the direct advertising elasticity of the $j^{\text{th}}$ market share and the advertising elasticity of the total market relative to the media investment of brand $j$. The former takes into account the cross effects of $M_j$ on other brands' market shares, and we expect the latter to be very small in the case of a non-oligopolistic competitive market which is the case of the automobile market.

This case can be generalized for any number of brands $D$ and for any number of channel $K$. Further research should also be done to incorporate the dynamic aspect of advertising impact, combining our adstock specification and the proposition of Levy and Simon [38].

## 4.5 Conclusion

In order to answer the question "What is the impact of media investments on the brands' market shares in the automobile market?", additional models can be considered in order to cover the whole French automobile market in a finer way, and to take into account its segmentation. Moreover, the distributional assumption of the proposed market share models can be questioned and alternative distributions should be considered if the Gaussian distribution seems to be inadequate. Concerning the interpretation of market share models, elasticities also have drawbacks which may be circumvented. Finally, the ultimate goal of measuring the impact of each advertising channel is to optimize the advertising budget and the marketing mix. We show that the famous Dorfman-Steiner theorem can be generalized for the multi-channel and competitive case, but work still has to be done to integrate the carryover effect of advertising.

# Conclusion (English version)

The objective of this thesis is to answer from a mathematical point of view the following question: "What is the impact of media investments on the brands' market shares in the automobile market?".

Because of the constraints of shares data, classical regression models cannot be used directly to model market shares. Market share models have been developed in the marketing literature, but other statistical models can be adapted to this type of applications. In the first chapter of this thesis, we present four types of models suitable when modeling market shares, or share data in general: the multinomial logit model (MNL), the generalized multiplicative competitive interaction model (GMCI), the Dirichlet model (DIR) and the linear compositional model (CODA).

We express all of them in an attraction formulation to ease their comparison, and we highlight the similarities and the differences of these models from a theoretical point of view. We prove that GMCI can be written as a particular compositional model, and that it can be considered as a particular case of the CODA model. The CODA model comes out to be similar to the fully extended attraction model used in marketing, but with several advantages: for example, it manages to capture all cross effects with a relative parsimony, thanks to the isometric log-ratio (ILR) transformation involved in the estimation. These four types of models are all easy to implement, with the R software for example.

The focus of the second chapter is to combine the best part of the MCI model and of the CODA model presented in Chapter 1, in order to improve the interpretability of CODA models, and to improve the estimation of MCI models. We stress the positive aspects of using an isometric log-ratio (ILR) transformation to estimate the MCI model, as in the CODA model, instead of the usual centered log-ratio (CLR) transformation used in marketing. We also develop an intermediate specification between the MCI and the CODA models, which we call the MCODA model. A model selection procedure is proposed using an adapted Fisher test, considering that the CODA model is the unconstrained model to be compared to the constrained models, the MCI model or the MCODA model.

We present a set of possible measures, mutually consistent, to interpret the coefficients of these models: marginal effects, elasticities and odds ratios. For example, we are

able to compute the direct and cross elasticities of brands' market shares relative to the advertising investment of a given brand in a given communication channel. This type of interpretation is totally suitable to answer our initial question, and is consistent with the derivatives in the simplex. However, this measure is observation dependent (it varies across time in our case) and we have to make sure that it is stable across observations to use it.

The third chapter presents the final application and the final model. We measure the impact of advertising investments through different channels (outdoor, press, radio and television), on brands' market shares in the French automobile market, from 2005 to 2015. We focus on the main segment of this market, namely the B segment, and on the three leaders of this segment: Citroën, Peugeot and Renault, aggregating the other brands in a group.

After analyzing in a descriptive manner this market in terms of absolute and relative sales and media investments for the different brands, we emphasize the importance of taking into account the competition on the one hand, and the advertising carryover effect on the other hand. For this purpose, we build a multi-channel attraction model with carryover effects, called CODAAd model for "CODA model with adstock", where the adstock is a cumulative variable made of actual advertising expenses and of a decreasing function of past advertising expenses. This CODAAd model is a combination of the Koyck model (the most common model used to introduce advertising carryover effects) and of the CODA model. We explain how to determine the carryover parameters $\lambda$ for several channels in a simultaneous way, and we conclude that outdoor, press and television advertisements have a large retention rate implying an advertising half life of 5.6 months, which seems realistic for a durable and expensive good such as automobile. On the contrary, the radio advertising appears to have only contemporaneous effect on market shares in the B segment of the automobile market.

Several model specifications are compared: with or without explanatory variables (constant models), with or without cross effects, with or without adstock variables, using a model belonging to the Dirichlet family or to the MCI family. According to goodness-of-fit (on the 2005-2014 period) and prediction accuracy (on 2015) measures, the CODAAd model, is considered to be the best model for our purpose. An adapted Fisher test confirms that the inclusion of cross effects improves significantly the model. The residual diagnostic confirms that the CODAAd model has good properties. We explain how to draw the 95% confidence and prediction ellipsoids in the space of market shares (i.e. in the simplex) and how to derive the 95% confidence and prediction intervals for each market share.

In order to interpret the impact of each channel on brands' market shares, we compute direct and cross advertising elasticities. While television direct elasticities are positive for all considered brands, this is not the case for the other channels. Citroën has a different advertising impact profile from the other considered brands: the model suggests that Citroën can increase its market share diminishing its outdoor and press advertising budget in favor of the radio and television advertisements, whereas for Peugeot, Renault

and the group of others, it suggests to increase the investments in outdoor, press and television and to reduce the radio communication.

Finally, the fourth and last chapter of this thesis addresses further directions to be investigated in order to answer our initial question. Additional models can be considered in order to cover the whole French automobile market in a finer way, and to take into account its segmentation. Moreover, the distributional assumptions of the proposed market share models can be questioned and alternative distributions should be considered if the Gaussian distribution and the Dirichlet distribution seem to be inadequate. Concerning the interpretation of market share models, elasticities also have drawbacks which may be circumvented. Finally, the ultimate goal of measuring the impact of each advertising channel is to optimize the advertising budget and the marketing mix. We show that the Dorfman-Steiner theorem can be generalized for the multi-channel and competitive case, but work still has to be done to integrate the carryover effect of advertising.

# Conclusion (version française)

L'objectif de cette thèse a été répondre d'un point de vue mathématique à la question suivante : "Quel est l'impact des investissements publicitaires sur les parts de marché des marques dans le marché automobile ?".

A cause des contraintes qui sont propres aux données de parts, les modèles de régression classiques ne peuvent pas être utilisés directement pour modéliser des parts de marché. Des modèles de parts de marché ont été développés dans la littérature marketing, mais d'autres modèles statistiques peuvent également convenir pour ce type d'applications. Dans le premier chapitre de cette thèse, quatre types de modèles adaptés pour modéliser des parts de marché, ou des données de parts en général, sont présentés : le modèle multinomial logit (MNL), le modèle d'interaction concurrentielle multiplicative généralisé (GMCI), le modèle de Dirichlet (DIR) et le modèle de composition linéaire (CODA). Tous ces modèles ont été réexprimés sous forme de modèles d'attraction pour faciliter leur comparaison, et nous avons mis en évidence les points communs et les différences d'un point de vue théorique. Nous avons prouvé que le modèle GMCI peut s'écrire de manière compositionnelle et qu'il peut être considéré comme un cas particulier du modèle CODA. Ce dernier se révèle être similaire au modèle d'attraction étendu ("fully extended attraction model") utilisé en marketing, mais avec plusieurs avantages : par exemple, il permet de capturer la complexité de tous les effets croisés avec une relative parcimonie, grâce à la transformation log ratio isométrique (ILR) utilisée pour l'estimation du modèle. Ces quatre types de modèles sont faciles à implémenter, notamment via le logiciel R.

L'objectif du deuxième chapitre a été de combiner les atouts des modèles MCI et CODA présentés dans le premier chapitre, de manière à améliorer l'interprétabilité du modèle CODA et de perfectionner la méthode d'estimation du modèle MCI. Nous avons montré qu'il est préférable d'estimer le modèle MCI en utilisant une transformation ILR des données de parts, comme dans le modèle CODA, plutôt qu'une transformation log ratio centrée (CLR) comme habituellement en marketing. Nous avons également développé une spécification intermédiaire entre le modèle MCI et le modèle CODA, que nous appelons modèle MCODA. Une procédure de sélection de modèle est proposée, basée sur un test de Fisher adapté, où le modèle CODA est considéré comme le modèle non contraint, à comparer aux modèles contraints que sont les modèles MCI et MCODA.

Nous avons présenté un ensemble de mesures possibles, cohérentes entre elles, pour l'interprétation de ces modèles : des effets marginaux, des élasticités et des rapports de cotes. Nous sommes par exemple capables de calculer les élasticités directes et croisées des parts de marché des marques à l'investissement publicitaire d'une certaine marque dans un certain canal de communication. Ce genre d'interprétation est parfaitement approprié pour répondre à la problématique qui est la nôtre, et il est cohérent avec les dérivées dans le simplexe. Cependant, cette mesure dépend des observations (elle varie au cours du temps dans notre cas) et nous devons nous assurer de sa stabilité pour pouvoir l'utiliser convenablement.

Le troisième chapitre présente l'application finale et le modèle final. Il s'agit de mesurer l'impact des investissements publicitaires de différents canaux de communication (affichage, presse, radio et télévision) sur les parts de marché des marques du marché automobile français, de 2005 à 2015. Nous nous sommes concentrés sur le segment principal de ce marché, le segment B, et sur les trois leaders de ce segment : Citroën, Peugeot et Renault, les autres marques étant agrégées ensemble.

Après avoir analysé de manière descriptive ce marché en termes de ventes et de dépenses publicitaires pour les différentes marques, nous avons insisté sur l'importance de prendre en compte la concurrence d'une part, et l'effet retard de la publicité d'autre part. Ainsi, nous avons construit un modèle d'attraction multicanal avec effets retard, appelé CODAAd pour "modèle CODA avec adstock", l'adstock désignant la variable cumulative de la publicité constituée des investissements publicitaires courants et d'une fonction décroissante des investissements publicitaires passés. Ce modèle CODAAd est une combinaison du modèle de Koyck (le modèle le plus usuel pour introduire des effets retard de la publicité) et du modèle CODA. Nous détaillons comment estimer les paramètres de rétention de la publicité de cette fonction d'adstock pour les différents canaux simultanément. Dans notre cas, les communications en affichage, presse et télévision ont un fort taux de rétention estimé correspondant à une demi-vie de la publicité de 5,6 mois, ce qui semble réaliste pour un bien durable et coûteux comme l'automobile. Au contraire, on estime que la publicité radio n'a qu'un impact ponctuel lors du mois de sa diffusion pour le segment B du marché automobile français.

Plusieurs spécifications de modèles ont été comparées : avec ou sans variables explicatives (modèles constants), avec ou sans effets croisés, avec ou sans variables d'adstock, en utilisant un modèle de la famille Dirichlet ou de la famille MCI. D'après les mesures de qualité d'ajustement (sur la période 2005-2014) et de précision de la prédiction (sur 2015), le modèle CODAAd est considéré comme le meilleur modèle dans notre cas. Un test de Fisher adapté a permis de valider que l'inclusion d'effets croisés améliore significativement le modèle. Le diagnostic des résidus a confirmé que le modèle CODAAd a de bonnes propriétés. Nous avons montré comment construire des ellipsoïdes de confiance et de prédiction à un niveau de 95% dans l'espace des parts de marché (i.e. dans le simplexe) et comment en déduire des intervalles de confiance et de prédiction pour chaque part de marché séparément.

Pour interpréter l'impact de chaque canal de communication sur les parts de marché

des marques, nous avons calculé les élasticités directes et croisées. Alors que les élasticités directes relatives à la télévision sont positives pour toutes les marques considérées, ce n'est pas le cas pour les autres canaux. Citroën a un profil différent des autres marques considérées concernant l'impact publicitaire : le modèle suggère que Citroën peut augmenter sa part de marché en diminuant son budget publicitaire en affichage et en presse, en faveur de la publicité radio et télévisée, alors que pour Peugeot, Renault et le groupe des autres marques, il suggère d'augmenter le budget affichage, presse et télévision et de réduire les dépenses en radio.

Enfin, le quatrième chapitre de cette thèse aborde les pistes de recherche qui permettraient d'améliorer la réponse à notre question initiale. Des modèles complémentaires peuvent être considérés de manière à couvrir plus finement tout le marché automobile français, et à prendre en compte sa segmentation. De plus, les hypothèses de distribution des modèles de parts de marché proposés peuvent être remises en question et des distributions alternatives peuvent être proposées si les distributions gaussienne ou de Dirichlet semblent inadéquates. Concernant l'interprétation des modèles de parts de marché, les élasticités ont quelques défauts qui pourraient être abordés. Pour finir, le but ultime de la mesure de l'impact publicitaire de chaque canal est d'optimiser le budget publicitaire et le mix marketing. Nous avons montré que le théorème de Dorfman et Steiner peut être généralisé au cas compétitif et multicanal, mais il reste à intégrer les effets retard de la publicité.

# Bibliography


[1] Aitchison, J. *The statistical analysis of compositional data.* Monographs on statistics and applied probability. Chapman and Hall, 1986.

[2] Assmus, G., Farley, J. U., and Lehmann, D. R. How advertising affects sales: Meta-analysis of econometric results. *Journal of Marketing Research* (1984), 65–74.

[3] Bechtel, G. G. Share-ratio estimation of the nested multinomial logit model. *Journal of Marketing Research 27*, 2 (1990), pp. 232–237.

[4] Brehm, J., Gates, S., and Gomez, B. A Monte Carlo comparison of methods for compositional data analysis. In *1998 annual meeting of the Society for Political Methodology* (1998).

[5] Broadbent, S. One way tv advertisements work. *Journal of the Market Research Society 21*, 3 (1979), 139–166.

[6] Campbell, G., and Mosimann, J. E. Multivariate analysis of size and shape: modelling with the Dirichlet distribution. In *ASA Proceedings of Section on Statistical Graphics* (1987), pp. 93–101.

[7] Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., and Vignes, C. Spatial scale in land use models: application to the Teruti-Lucas survey. *Spatial Statistics* (2016).

[8] Chen, J., Zhang, X., and Li, S. Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics* (2016), 1–16.

[9] Clarke, D. G. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* (1976), 345–357.

[10] Cooper, L., and Nakanishi, M. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness.* International Series in Quantitative Marketing. Springer, 1988.

[11] Danaher, P. J., Bonfrer, A., and Dhar, S. The effect of competitive advertising interference on sales for packaged goods. *Journal of Marketing Research 45*, 2 (2008), pp. 211–225.

[12] Dorfman, R., and Steiner, P. O. Optimal advertising and optimal quality. *The American Economic Review 44*, 5 (1954), 826–836.

[13] Egozcue, J. J., Daunis-I-Estadella, J., Pawlowsky-Glahn, V., Hron, K., and Filzmoser, P. Simplicial regression. the normal model. *Journal of applied probability and statistics* (2012).

[14] Egozcue, J. J., Pawlowsky-Glahn, V., Templ, M., and Hron, K. Independence in contingency tables using simplicial geometry. *Communications in Statistics-Theory and methods 44*, 18 (2015), 3978–3996.

[15] Elff, M. Social divisions, party positions, and electoral behaviour. *Electoral Studies 28*, 2 (2009), 297–308.

[16] Elff, M. *mclogit: Mixed Conditional Logit*, 2014. R package version 0.3-1.

[17] Friendly, M., Monette, G., Fox, J., et al. Elliptical insights: understanding statistical methods through elliptical geometry. *Statistical Science 28*, 1 (2013), 1–39.

[18] Fry, J. M., Fry, T., and McLaren, K. Compositional data analysis and zeros in micro data. Centre of policy studies/impact centre working papers, Victoria University, Centre of Policy Studies/IMPACT Centre, 1996.

[19] Fry, T. R. L., and Chong, D. A tale of two logit, compositional data analysis and zero observations. April 2006.

[20] Ghosh, A., Neslin, S., and Shoemaker, R. A comparison of market share models and estimation procedures. *Journal of Marketing Research 21*, 2 (1984), pp. 202–210.

[21] Glerant, A. *Evolution de l'élasticité de la demande de biens durables aux éléments de marketing-mix selon les phases du cycle de vie : une application au marché automobile*. PhD thesis, Paris Dauphine, 1993.

[22] Green, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), 149–192.

[23] Green, R., Hahn, W., and Rocke, D. Standard errors for elasticities: a comparison of bootstrap and asymptotic standard errors. *Journal of Business & Economic Statistics 5*, 1 (1987), 145–149.

[24] Greenacre, M., and Lewi, P. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification 26*, 1 (Apr 2009), 29–54.

[25] Gruca, T., and Sudharshan, D. Equilibrium characteristics of multinomial logit market share models. 480.

[26] Haaf, C. G., Michalek, J., Morrow, W. R., and Liu, Y. Sensitivity of vehicle market share predictions to discrete choice model specification. *Journal of Mechanical Design 136* (12 2014).

[27] Hanssens, D. M., Parsons, L. J., and Schultz, R. L. *Market response models: Econometric and time series analysis*, vol. 12. Springer Science & Business Media, 2003.

[28] Hijazi, R. H. Residuals and diagnostics in Dirichlet regression. *ASA Proceedings of the General Methodology Section* (2006), 1190–1196.

[29] Hijazi, R. H., and Jernigan, R. W. Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics 4*, 1 (2009), 77–91.

[30] Hron, K., Filzmoser, P., and Thompson, K. Linear regression with compositional explanatory variables. *Journal of Applied Statistics 39*, 5 (2012), 1115–1128.

[31] Joy, J. Understanding advertising adstock transformations. *Munich Personal RePEc Archive*, 7683 (MAY 2006).

[32] Katz, J. N., and King, G. A statistical model for multiparty electoral data. *American Political Science Review 93*, 1 (1999), 15–32.

[33] Koppelman, F. S., and Bhat, C. *A self instructing course in mode choice modeling: multinomial and nested logit models*. FTA, 2006.

[34] Kumar, V. Forecasting with market response models forecasting performance of market share models: an assessment, additional insights, and guidelines. *International Journal of Forecasting 10*, 2 (1994), 295 – 312.

[35] Kynclova, P., Filzmoser, P., and Hron, K. Modeling compositional time series with vector autoregressive models. *Journal of Forecasting 34*, 4 (2015), 303–314.

[36] Leeflang, P. S. H., and Reuyl, J. C. On the predictive power of market share attraction models. *Journal of Marketing Research 21*, 2 (1984), pp. 211–215.

[37] Leone, R. P. Generalizing what is known about temporal aggregation and advertising carryover. *Marketing Science 14*, 3_supplement (1995), G141–G150.

[38] Levy, H., and Simon, J. L. A generalization that makes useful the Dorfman–Steiner theorem with respect to advertising. *Managerial and Decision Economics 10*, 1 (1989), 85–87.

[39] Lodish, L. M., Abraham, M. M., Livelsberger, J., Lubetkin, B., Richardson, B., and Stevens, M. E. A summary of fifty-five in-market experimental estimates of the long-term effect of TV advertising. *Marketing Science 14*, 3 (1995), pp. G133–G140.

[40] Maier, M. J. DirichletReg: Dirichlet regression for compositional data in R. Research Report Series/Department of Statistics and Mathematics 125, WU Vienna University of Economics and Business, Vienna, January 2014.

[41] Martin-Fernandez, J., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. Zero replacement in compositional data sets. In *Data Analysis, Classification, and Related Methods*. Springer, 2000, pp. 155–160.

[42] Martin-Fernandez, J. A., Bren, M., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. A measure of difference for compositional data based on measures of divergence. In *Proceedings of IAMG* (1999), vol. 99, pp. 211–216.

[43] McFadden, D. L. Econometric analysis of qualitative response models. *Handbook of econometrics 2* (1984), 1395–1457.

[44] Monti, G., Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. Shifted-Dirichlet regression vs simplicial regression: a comparison. *Welcome to CoDawork 2015* (2015).

[45] Morais, J., and Thomas-Agnan, C. Impact of economic conditions on automobile market segment shares: a compositional approach. *CSBIGS* (2018). In progress.

[46] Morais, J., Thomas-Agnan, C., and Simioni, M. A tour of regression models for explaining shares. *TSE Working Paper*, 16-742 (2016).

[47] Morais, J., Thomas-Agnan, C., and Simioni, M. Interpretation of explanatory variables impacts in compositional regression models. Working paper, July 2017.

[48] Morais, J., Thomas-Agnan, C., and Simioni, M. Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper 17-805* (2017).

[49] Morais, J., Thomas-Agnan, C., and Simioni, M. Using compositional and Dirichlet models for market-share regression. Working paper, July 2017.

[50] Morais, J., Thomas-Agnan, C., and Simioni, M. Using compositional and Dirichlet models for market-share regression. *TSE Working Paper 17-804* (2017).

[51] Naert, P., and Weverbergh, M. On the prediction power of market share attraction models. *Journal of Marketing Research 18*, 2 (1981), pp. 146–153.

[52] Nakanishi, M., and Cooper, L. G. Simplified estimation procedures for MCI models. *Marketing Science 1*, 3 (1982), pp. 314–322.

[53] Palarea-Albaladejo, J., Martín-Fernández, J. A., and Soto, J. A. Dealing with distances and transformations for fuzzy c-means clustering of compositional data. *Journal of classification 29*, 2 (2012), 144–169.

[54] Pawlowsky-Glahn, V., and Buccianti, A. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

[55] Pawlowsky-Glahn, V., and Egozcue, J. J. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*, 5 (2001), 384–398.

[56] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[57] Peyhardi, J., Trottier, C., and Guédon, Y. A new specification of generalized linear models for categorical responses. *Biometrika 102*, 4 (2015), 889–906.

[58] Quagrainie, K. K. Analysis of U.S. catfish fillet market share using a flexible logistic model. *Marine Resource Economics 21*, 1 (2006), pp. 33–45.

[59] Scealy, J., and Welsh, A. Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*, 3 (2011), 351–375.

[60] Smithson, M., and Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods 11*, 1 (2006), 54.

[61] So, Y., and Kuhfeld, W. F. Multinomial logit models. In *SUGI 20 Conference Proceedings, Cary, NC: SAS Institute Inc* (1995).

[62] Solano-Acosta, W., and Dutta, P. K. Unexpected trend in the compositional maturity of second-cycle sand. *Sedimentary Geology 178*, 3 (2005), 275–283.

[63] Templ, M., Hron, K., and Filzmoser, P. *robCompositions: an R-package for robust statistical analysis of compositional data*, 2011.

[64] Train, K. *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*, 1 ed., vol. 1. The MIT Press, 1985.

[65] Trinh, H. T., and Morais, J. Impact of socioeconomic factors on nutritional diet in Vietnam from 2004 to 2014: new insights using compositional data analysis. *TSE Working Papers*, 17-825 (2017).

[66] Vakratsas, D., and Ambler, T. How advertising works: what do we really know? *The Journal of Marketing* (1999), 26–43.

[67] van den Boogaart, K. G., Tolosana, R., and Bren, M. *compositions: Compositional Data Analysis*, 2014. R package version 1.40-1.

[68] Van Den Boogaart, K. G., and Tolosana-Delgado, R. *Analysing Compositional Data with R*. Springer, 2013.

[69] Wang, H., Liu, Q., Mok, H. M., Fu, L., and Tse, W. M. A hyperspherical transformation forecasting model for compositional data. *European journal of operational research 179*, 2 (2007), 459–468.

[70] Wang, H., Shangguan, L., Wu, J., and Guan, R. Multiple linear regression modeling for compositional data. *Neurocomputing 122* (2013), 490–500.

[71] Wright, M. A new theorem for optimizing the advertising budget. *Journal of advertising research 49*, 2 (2009), 164–169.

[72] Zantedeschi, D., Feit, E. M., and Bradlow, E. T. Measuring multichannel advertising response. *Management Science* (2016).

# Appendix A

# Appendix

## A.1 Appendix (Chapter 1)

### A.1.1 MCI model: a particular case of the CODA model

Let us consider a CODA model where the dimensions of the dependent composition and of the explanatory composition are such as $D_S = D_X = 3$, where the matrix of coefficients in the ILR transformed space is equal to

$$\mathbf{B}^* = \begin{bmatrix} b^* & 0 \\ 0 & b^* \end{bmatrix},$$

and where the balance matrix used for the ILR transformation is for example

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ -\frac{2}{\sqrt{6}} & 0 \end{bmatrix}$$

Then, the matrix of the coefficients in the simplex space is

$$\mathbf{B} = \mathbf{V}\mathbf{B}^*\mathbf{V}' = \frac{1}{3}b^* \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix},$$

such that the matrix $\mathbf{B}$ does verify the rows sum and columns sum equal to 0 requirement.

We can check that in this case we have $\mathbf{B} \boxdot \mathbf{X} = b \odot \mathbf{X}$:

$$\begin{aligned} \mathbf{B} \boxdot \mathbf{X} &= \mathcal{C}(X_1^{\frac{2}{3}b} X_2^{-\frac{1}{3}b} X_3^{-\frac{1}{3}b} \,,\, X_1^{-\frac{1}{3}b} X_2^{\frac{2}{3}b} X_3^{-\frac{1}{3}b} \,,\, X_1^{-\frac{1}{3}b} X_2^{-\frac{1}{3}b} X_3^{\frac{2}{3}b})' \\ &= \mathcal{C}(X_1^b (X_1X_2X_3)^{-\frac{1}{3}b} \,,\, X_2^b (X_1X_2X_3)^{-\frac{1}{3}b} \,,\, X_3^b (X_1X_2X_3)^{-\frac{1}{3}b})' \\ &= \mathcal{C}(X_1^b \,,\, X_2^b \,,\, X_3^b)' = b \odot \mathbf{X} \end{aligned}$$

Then, in this particular case, the CODA specification is equivalent to the MCI specification, meaning that the MCI model is a particular case of the CODA model. This relationship holds for any ILR transformation.

### A.1.2 Non scale invariance of the DMCI model

The differential MCI model (DMCI) used in the marketing literature (see for example Cooper and Nakanishi [10], p.58) is not a particular case of the CODA model, contrary to the MCI model, because it is not scale invariant. The DMCI is defined by:

$$S_{jt} = \frac{a_j X_{jt}^{b_j} c_j^{Z_t} \epsilon_{jt}}{\sum_{m=1}^{D} a_m X_{mt}^{b_m} c_m^{Z_t} \epsilon_{mt}}$$

with brand specific parameters $b_j$. Thus, it corresponds to the following equation in the simplex, for $D = 3$:

$$\mathbf{S}_t = \mathcal{C}(a_1 X_{1t}^{b_1} c_1^{Z_t}, a_2 X_{2t}^{b_2} c_2^{Z_t}, a_3 X_{3t}^{b_3} c_3^{Z_t})$$

but this function is not scale invariant. For example, let us consider that $X_j, \forall j = 1, \ldots, D$ is multiplied by 100 (e.g. thousands euro instead of euro). Then,

$$\mathcal{C}(a_1 X_{1t}^{b_1} c_1^{Z_t}, a_2 X_{2t}^{b_2} c_2^{Z_t}, a_3 X_{3t}^{b_3} c_3^{Z_t}) \neq \mathcal{C}(a_1 (100X_{1t})^{b_1} c_1^{Z_t}, a_2 (100X_{2t})^{b_2} c_2^{Z_t}, a_3 (100X_{3t})^{b_3} c_3^{Z_t})$$

because of the different $b_j$ parameters, whereas in the case of MCI and CODA models, scale invariance holds respectively thanks to the closure and thanks to the zero sums of columns and rows in the $\mathbf{B}$ matrix of parameters.

### A.1.3 Treatment of zero observations

Zeros are often an issue with share data. For GMCI, Dirichlet and CODA models, zeros cannot be tolerated because of the presence of the log transformation of shares in the likelihood. Many solutions to this problem have been considered, depending on the nature of zeros. Among the main ones, let us mention amalgamation of components (Pawlowsky-Glahn et al. [56]), ratio-preserving zero replacement (Martin-Fernadez et al. [41]) and conditional modeling for the CODA literature. Several transformations have been proposed for this problem, see for example Smithson and Verkuilen [60] in the case of the Dirichlet model. Wang et al. [69] and Scealy and Welsh [59] use a square root transformation together with models on the hypersphere. Fry et al. [18] compare their performance for the case of economic micro-data.

In order to fit the different models studied in this thesis, we have made the following substitutions: when during a given period, for a given brand, the media expense (total media in Chapter 1 and Chapter 2, or by channel in Chapter 3) is null, it is replaced by 1 euro (very small value in comparison to the usually huge advertising budgets). Indeed, most of the market share models use log ratios and then do not admit null values for compositional dependent and explanatory variables. Several imputation methods exist for zero values as explained above, but in the case of structural zeros, or “real zeros”, as in our case, the simplest way to replace zeros is to choose a small value. Note that thanks to the amalgamation of "Others" brands and the adstock computation, there is no zero problem in the final model of this thesis.

### A.1.4 Quality measures

We present here some goodness-of-fit measures adapted to the case where the dependent variable is a vector of shares. Two categories of measures are detailed: the $R^2$-type measures which are based on the notion of explained variability, and the distance-type measures which evaluate how far are the fitted values from the true values.

**$R^2$ based on total variability ($R^2_T$)** The compositional data analysis literature proposes a $R^2$ directly adapted to compositional data (see Hijazi [28], Monti et al. [44]). It uses the measure of the total variability of a set of compositions, based on the variance of log ratios. In terms of interpretation, it is similar to the classical $R^2$: it measures the proportion of the total variation explained by the model:

$$R^2_T = \frac{totvar(\widehat{\mathbf{S}})}{totvar(\mathbf{S})} \quad \text{where } totvar(\mathbf{S}) = \frac{1}{2D}\sum_{j=1}^{D}\sum_{l=1}^{D} var(\log\frac{S_j}{S_l})$$

This measure is always positive but is not guaranteed to be lower than 1. Note that for the constant model (where the only explanatory variables are component-specific intercepts), $R^2_T$ equals zero for all models because there is no variability in $\widehat{\mathbf{S}}$.

**$R^2$ based on Aitchison distance ($R^2_A$)** Another $R^2$ measure can be found in the CODA literature, based on the Aitchison distance between the observed compositions and the fitted compositions on one hand, and on the Aitchison distance between the observed compositions and the center of the data (closed geometric means of components) on the other hand (see Hijazi [28], Monti et al. [44]).

$$R^2_A = 1 - \frac{CSSE}{CSST}$$

with $CSST = \sum_{t=1}^{T} d^2_A(\mathbf{S_t}, \mathbf{g})$ ; $CSSE = \sum_{t=1}^{T} d^2_A(\mathbf{S_t}, \widehat{\mathbf{S}}_\mathbf{t})$. $\mathbf{g}$ is the closed vector of geometric means of each component over observations $t$, and

$$d_A(\mathbf{S_t}, \widehat{\mathbf{S}}_\mathbf{t}) = \sqrt{\sum_{j=1}^{D}\left(\log\frac{S_{jt}}{g(S_j)} - \log\frac{\widehat{S}_{jt}}{g(\widehat{S}_j)}\right)^2} = \sqrt{\frac{1}{D}\sum_{j=1}^{D}\sum_{l>j}^{D}\left(\log\frac{S_{jt}}{S_{lt}} - \log\frac{\widehat{S}_{jt}}{\widehat{S}_{lt}}\right)^2}$$

However, this $R^2$ measure can be misleading because it has a large variability and it can take negative values. Note that for the constant model, $R^2_A$ equals zero for CODA and GMCI models because $\widehat{\mathbf{S}} = g(\mathbf{S})$.

**Kullback-Leibler divergence (KL)** The Kullback-Leibler divergence is used as a goodness-of-fit measure or as a prediction accuracy measure (see Haaf et al. [26]). It is a sum of the log ratios between the observed values and the fitted values of the shares,

weighted by the observed value. The log ratio allows to take into account the relative error, and the weight emphasizes the importance of large errors in large shares.

$$KL(\mathbf{S}, \widehat{\mathbf{S}}) = \sum_{t=1}^{T} \sum_{j=1}^{D} \log \left( \frac{S_{jt}}{\widehat{S}_{jt}} \right) S_{jt}$$

A compositional version of this measure is defined as follows (see Martin-Fernandez et al. [42], and Palarea et al. [53]):

$$KLC(\mathbf{S}, \widehat{\mathbf{S}}) = \frac{D}{2} \left( KL(\mathbf{0_D}, \mathbf{S} \ominus \widehat{\mathbf{S}}) + KL(\mathbf{0_D}, \widehat{\mathbf{S}} \ominus \mathbf{S}) \right) = \frac{D}{2} \sum_{t=1}^{T} \log \left( \overline{(S_t/\widehat{S}_t)(\widehat{S}_t/S_t)} \right)$$

where $\mathbf{0_D} = (1/D, \ldots, 1/D)$ the compositional zero (center of the simplex $\mathcal{S}^D$), and $\overline{(S_t/\widehat{S}_t)}$ the arithmetic mean of shares ratios $\left( \frac{S_{1t}}{\widehat{S}_{1t}}, \ldots, \frac{S_{Dt}}{\widehat{S}_{Dt}} \right)$ for observation $t$.

The KLC measure is indeed well adapted to shares data because for the constant model, this measure of divergence is lower for models which predict the geometric means of the shares (CODA and GMCI models) than for models which predict the arithmetic means (MNL and DIR models), and it is well known that the geometric mean is more adapted to summarize compositional data than the arithmetic mean.

Other quality measures can be used for share data. See for example Kumar [34], Quagrainie [58], Leeflang and Reuyl [36], Naert and Weverbergh [51], Ghosh et al. [20].

## A.2 Appendix (Chapter 2)

### A.2.1 Marginal effect and elasticity calculus on ILR

We are going to demonstrate how to compute marginal effects of the volume $\check{X}_{lt}$ on the dependent shares $S_{jt}$, and elasticities of $S_{jt}$ relative to $\check{X}_{lt}$, using the transformed and the non-transformed compositional models. The demonstration is made for the CODA model, with $D = 3$ components and an ILR transformation defined by the transformation matrix

$$\mathbf{V} = \begin{bmatrix} \sqrt{\frac{2}{3}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Let us remind that $\mathbf{X}^* = ilr(\mathbf{X}) = \mathbf{V}' \log(\mathbf{X})$, and $\mathbf{X} = ilr^{-1}(\mathbf{X}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{X}^*))$. We define the following transformations:

$$\begin{aligned} T &: (\check{X}_1,\ \check{X}_2,\ \check{X}_3)' \to (\check{X}_1^*,\ \check{X}_2^*)' \\ F &: (\check{X}_1^*,\ \check{X}_2^*)' \to (\mathbb{E}S_1^*,\ \mathbb{E}S_2^*)' = (a_1^* + b_{11}^*\check{X}_1^* + b_{12}^*\check{X}_2^*,\ a_2^* + b_{21}^*\check{X}_1^* + b_{22}^*\check{X}_2^*)' \\ T^{-1} &: (\mathbb{E}S_1^*,\ \mathbb{E}S_2^*)' \to (\mathbb{E}^{\oplus}S_1,\ \mathbb{E}^{\oplus}S_2,\ \mathbb{E}^{\oplus}S_3)' \end{aligned}$$

We are going to use the following property of Jacobian matrices: $J = J_{T^{-1}} J_F J_T$, implying that:

$$\begin{aligned} ME(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t) &= \left[\frac{\partial \mathbb{E}^{\oplus} S_{it}}{\partial \check{X}_{jt}}\right]_{D,D} \\ &= \left[\frac{\partial \mathbb{E}^{\oplus} S_{it}}{\partial \mathbb{E}S_{jt}^*}\right]_{D,D-1} \left[\frac{\partial \mathbb{E}S_{it}^*}{\partial \check{X}_{jt}^*}\right]_{D-1,D-1} \left[\frac{\partial \check{X}_{it}^*}{\partial \check{X}_{jt}}\right]_{D-1,D} \end{aligned}$$

and

$$\begin{aligned} E(\mathbb{E}^{\oplus}\mathbf{S}_t, \check{\mathbf{X}}_t) &= \left[\frac{\partial \log \mathbb{E}^{\oplus} S_{it}}{\partial \log \check{X}_{jt}}\right]_{D,D} \\ &= \left[\frac{\mathbf{1}}{\mathbf{S_{it}}}\right] \odot \left[\frac{\partial \mathbb{E}^{\oplus} S_{it}}{\partial \mathbb{E}S_{jt}^*}\right]_{D,D-1} \left[\frac{\partial \mathbb{E}S_{it}^*}{\partial \check{X}_{jt}^*}\right]_{D-1,D-1} \left[\frac{\partial \check{X}_{it}^*}{\partial \check{X}_{jt}}\right]_{D-1,D} \odot [\mathbf{X_{jt}}] \end{aligned} \quad \text{(A.1)}$$

where $\odot$ denotes here the Hadamard product (term by term product)[1], $\left[\frac{\mathbf{1}}{\mathbf{S_{it}}}\right]$ is a $D \times D-1$ matrix with $1/S_{it}$ on the $i^{\text{th}}$ row and $[\mathbf{X_{jt}}]$ is a $D-1 \times D$ matrix with $X_{jt}$ on the $j^{\text{th}}$ column.

**The Jacobian of the model in coordinates $J_F$**

$$J_F = \begin{bmatrix} \frac{\partial \mathbb{E}S_1^*}{\partial \check{X}_1^*} & \frac{\partial \mathbb{E}S_1^*}{\partial \check{X}_2^*} \\ \frac{\partial \mathbb{E}S_2^*}{\partial \check{X}_1^*} & \frac{\partial \mathbb{E}S_2^*}{\partial \check{X}_2^*} \end{bmatrix} = \begin{bmatrix} b_{11}^* & b_{12}^* \\ b_{21}^* & b_{22}^* \end{bmatrix} = \mathbf{B}^*$$

[1]Note that $\odot$ in bold denote the Hadamard product whereas $\odot$ denote the power transformation.

**The Jacobian of the transformation** $J_T$ The ILR transformation associated to the $\mathbf{V}$ balance matrix is defined as

$$\begin{aligned}(\check{X}_1^*,\ \check{X}_2^*)' &= T(\check{X}_1,\ \check{X}_2,\ \check{X}_3)' \\ &= \left(\sqrt{\frac{2}{3}}\log\check{X}_1 - \frac{1}{\sqrt{6}}\log\check{X}_2 - \frac{1}{\sqrt{6}}\log\check{X}_3,\ \frac{1}{\sqrt{2}}\log\check{X}_2 - \frac{1}{\sqrt{2}}\log\check{X}_3\right)'\end{aligned}$$

Then, $J_T = \begin{bmatrix} \frac{\partial\check{X}_1^*}{\partial\check{X}_1} & \frac{\partial\check{X}_1^*}{\partial\check{X}_2} & \frac{\partial\check{X}_1^*}{\partial\check{X}_3} \\ \frac{\partial\check{X}_2^*}{\partial\check{X}_1} & \frac{\partial\check{X}_2^*}{\partial\check{X}_2} & \frac{\partial\check{X}_2^*}{\partial\check{X}_3} \end{bmatrix} = \mathbf{V}' \odot \left[\frac{\mathbf{1}}{\mathbf{X_j}}\right] = \begin{bmatrix} \sqrt{\frac{2}{3}}\frac{1}{X_1} & -\frac{1}{\sqrt{6}}\frac{1}{X_2} & -\frac{1}{\sqrt{6}}\frac{1}{X_3} \\ 0 & \frac{1}{\sqrt{2}}\frac{1}{X_2} & -\frac{1}{\sqrt{2}}\frac{1}{X_3} \end{bmatrix}$,

where $\left[\frac{\mathbf{1}}{\mathbf{X_j}}\right]$ is a $D-1\times D$ matrix with $1/X_j$ on the $j^{\text{th}}$ column.

**The Jacobian of the inverse transformation** $J_{T^{-1}}$ The inverse ILR transformation is such as

$$\begin{aligned}(\mathbb{E}^{\oplus}S_1,\ \mathbb{E}^{\oplus}S_2,\ \mathbb{E}^{\oplus}S_3)' &= T^{-1}(\mathbb{E}S_1^*,\ \mathbb{E}S_2^*)' = \mathcal{C}(\exp(\mathbf{V}\mathbb{E}\mathbf{S}^*)') \\ &= \mathcal{C}\left(\exp(\mathbb{E}S_1^*)^{\sqrt{\frac{2}{3}}},\ \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}}\exp(\mathbb{E}S_2^*)^{\frac{1}{\sqrt{2}}},\ \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}}\exp(\mathbb{E}S_2^*)^{-\frac{1}{\sqrt{2}}}\right)' \\ &= \left(\frac{u_1}{DEN},\ \frac{u_2}{DEN},\ \frac{u_3}{DEN}\right)',\end{aligned}$$

where

$$\begin{aligned}u_1 &= \exp(\mathbb{E}S_1^*)^{\sqrt{\frac{2}{3}}} \\ u_2 &= \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}}\exp(\mathbb{E}S_2^*)^{\frac{1}{\sqrt{2}}} \\ u_3 &= \exp(\mathbb{E}S_1^*)^{-\frac{1}{\sqrt{6}}}\exp(\mathbb{E}S_2^*)^{-\frac{1}{\sqrt{2}}} \\ DEN &= u_1+u_2+u_3\end{aligned}$$

In order to compute the matrix $J_{T^{-1}} = \begin{bmatrix} \frac{\partial\mathbb{E}^{\oplus}S_1}{\partial\mathbb{E}S_1^*} & \frac{\partial\mathbb{E}^{\oplus}S_1}{\partial\mathbb{E}S_2^*} \\ \frac{\partial\mathbb{E}^{\oplus}S_2}{\partial\mathbb{E}S_1^*} & \frac{\partial\mathbb{E}^{\oplus}S_2}{\partial\mathbb{E}S_2^*} \\ \frac{\partial\mathbb{E}^{\oplus}S_3}{\partial\mathbb{E}S_1^*} & \frac{\partial\mathbb{E}^{\oplus}S_3}{\partial\mathbb{E}S_2^*} \end{bmatrix}$, we need to compute the derivatives of the numerators of $\mathbb{E}^{\oplus}\mathbf{S}$, $\mathbf{u} = (u_1, u_2, u_3)'$, with respect to $\mathbb{E}\mathbf{S}^*$:

$$\left(\frac{\partial\mathbf{u}}{\partial\mathbb{E}\mathbf{S}^*}\right) = \mathbf{V}\odot\mathbf{u} = \begin{bmatrix} \frac{\partial u_1}{\partial\mathbb{E}S_1^*} = \sqrt{\frac{2}{3}}u_1 & \frac{\partial u_1}{\partial\mathbb{E}S_2^*} = 0 \\ \frac{\partial u_2}{\partial\mathbb{E}S_1^*} = -\frac{1}{\sqrt{6}}u_2 & \frac{\partial u_2}{\partial\mathbb{E}S_2^*} = \frac{1}{\sqrt{2}}u_2 \\ \frac{\partial u_3}{\partial\mathbb{E}S_1^*} = -\frac{1}{\sqrt{6}}u_3 & \frac{\partial u_3}{\partial\mathbb{E}S_2^*} = -\frac{1}{\sqrt{2}}u_3 \end{bmatrix}$$

Now we can compute the elements of $J_{T^{-1}}$. For example, the first element of this matrix is

$$\begin{aligned}\frac{\partial\mathbb{E}^{\oplus}S_1}{\partial\mathbb{E}S_1^*} &= \frac{DEN\sqrt{\frac{2}{3}}u_1 - u_1[\sqrt{\frac{2}{3}}u_1 - \frac{1}{\sqrt{6}}u_2 - \frac{1}{\sqrt{6}}u_3]}{DEN^2} = \frac{\frac{3}{\sqrt{6}}u_1(u_2+u_3)}{DEN^2} \\ &= \frac{3}{\sqrt{6}}\mathbb{E}^{\oplus}S_1(1-\mathbb{E}^{\oplus}S_1),\end{aligned}$$

using the fact that $u_1/DEN = \mathbb{E}^{\oplus} S_1$ and $u_2 + u_3 = DEN - u_1$. Similar computations give the results for the whole matrix:

$$J_{T^{-1}} = \begin{bmatrix} \frac{\partial \mathbb{E}^{\oplus} S_1}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^{\oplus} S_1}{\partial \mathbb{E} S_2^*} \\ \frac{\partial \mathbb{E}^{\oplus} S_2}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^{\oplus} S_2}{\partial \mathbb{E} S_2^*} \\ \frac{\partial \mathbb{E}^{\oplus} S_3}{\partial \mathbb{E} S_1^*} & \frac{\partial \mathbb{E}^{\oplus} S_3}{\partial \mathbb{E} S_2^*} \end{bmatrix} = \begin{bmatrix} \frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 (1 - \mathbb{E}^{\oplus} S_1) & \frac{1}{\sqrt{2}} \mathbb{E}^{\oplus} S_1 (\mathbb{E}^{\oplus} S_3 - \mathbb{E}^{\oplus} S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 \mathbb{E}^{\oplus} S_2 & \frac{1}{\sqrt{2}} \mathbb{E}^{\oplus} S_2 (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 \mathbb{E}^{\oplus} S_3 & -\frac{1}{\sqrt{2}} \mathbb{E}^{\oplus} S_3 (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_2) \end{bmatrix}$$

$$= [\mathbf{S_{it}}] \odot \begin{bmatrix} \frac{3}{\sqrt{6}} (1 - \mathbb{E}^{\oplus} S_1) & \frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_3 - \mathbb{E}^{\oplus} S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 & \frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 & -\frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_2) \end{bmatrix} = [\mathbf{S_{it}}] \odot \mathbf{W}^*$$

**The Jacobian of the model in the simplex** $J$ The Jacobian matrix of the model is the matrix of marginal effects of $X_l$s on $S_j$.

$$J = J_{T^{-1}} J_F J_T = \begin{bmatrix} \frac{\partial S_1}{\partial \check{X}_1} & \frac{\partial S_1}{\partial \check{X}_2} & \frac{\partial S_1}{\partial \check{X}_3} \\ \frac{\partial S_2}{\partial \check{X}_1} & \frac{\partial S_2}{\partial \check{X}_2} & \frac{\partial S_2}{\partial \check{X}_3} \\ \frac{\partial S_3}{\partial \check{X}_1} & \frac{\partial S_3}{\partial \check{X}_2} & \frac{\partial S_3}{\partial \check{X}_3} \end{bmatrix} ME(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t)$$

$$= [\mathbf{S_{it}}] \odot \mathbf{W}^* \mathbf{B}^* \mathbf{V}' \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right] = [\mathbf{S_{it}}] \odot \mathbf{W}^* \mathbf{V}' \mathbf{B} \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right] = [\mathbf{S_{it}}] \odot \mathbf{W}\mathbf{B} \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right]$$

$$= [\mathbf{S_{it}}] \odot \begin{bmatrix} \frac{3}{\sqrt{6}} (1 - \mathbb{E}^{\oplus} S_1) & \frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_3 - \mathbb{E}^{\oplus} S_2) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 & \frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_3) \\ -\frac{3}{\sqrt{6}} \mathbb{E}^{\oplus} S_1 & -\frac{1}{\sqrt{2}} (\mathbb{E}^{\oplus} S_1 + 2\mathbb{E}^{\oplus} S_2) \end{bmatrix} \begin{bmatrix} b_{11}^* & b_{12}^* \\ b_{21}^* & b_{22}^* \end{bmatrix} \begin{bmatrix} \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right]$$

$$= [\mathbf{S_{it}}] \odot \begin{bmatrix} 1 - S_1 & -S_2 & -S_3 \\ -S_1 & 1 - S_2 & -S_3 \\ -S_1 & -S_2 & 1 - S_3 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right],$$

where $\mathbf{W}^* \mathbf{V}' = \mathbf{W}$ is a $D, D$ matrix with $1 - S_i$ in the diagonal and $-S_i$ in the row $i$ otherwise. According to equation (A.1), we deduce that the matrix of elasticities is equal to

$$E(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t) = \left[\frac{\mathbf{1}}{\mathbf{S_{it}}}\right] \odot ME(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t) \odot \left[\check{\mathbf{X}}_\mathbf{j}\right] = \mathbf{W}\mathbf{B}$$

Then, marginal effects and elasticities matrices are easy to compute using coefficients in the simplex or coefficients in the transformed space, using the following equations:

$$ME(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t) = [\mathbf{S_{it}}] \odot \mathbf{W}\mathbf{B} \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right] = [\mathbf{S_{it}}] \odot \mathbf{W}\mathbf{V}\mathbf{B}^* \mathbf{V}' \odot \left[\mathbf{1}/\check{\mathbf{X}}_\mathbf{j}\right]$$
$$E(\mathbb{E}^{\oplus} \mathbf{S}_t, \check{\mathbf{X}}_t) = \mathbf{W}\mathbf{B} = \mathbf{W}\mathbf{V}\mathbf{B}^* \mathbf{V}'$$

### A.2.2 Derivatives in the simplex

We keep here the notations of chapter 13 in Pawlowsky-Glahn and Buccianti [54] except that we denote $\frac{\partial f}{\partial^{\oplus} x}$ the part-C derivatives. Let $f$ be a vector-valued scale-invariant function from $\mathbb{R}^{D_X}$ to $\mathbb{R}^k$. Let $\underline{f}$ be the corresponding vector-valued function on $\mathcal{S}^{D_X}$ induced by $\underline{f}(\mathbf{x}) = f(\mathbf{w})$, where $\mathbf{w}$ is the vector of volumes corresponding to the vector of shares $\mathbf{x}$. We have $f(\mathbf{w}) = \underline{f}(\mathcal{C}(\mathbf{w}))$. For the sake of simplicity, let us assume that $D_X = 3$. We denote by $w_+ = \sum_{i=1}^{D_X} w_i$ the total volume. Taking the derivative of the previous equation with respect to $w_j$ yields

$$\frac{\partial f(\mathbf{w})}{\partial w_j} = \sum_{i=1}^{3} \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \frac{\partial x_i}{\partial w_j}$$

Since $\frac{\partial x_i}{\partial w_i} = \frac{w_+ - w_i}{w_+^2}$ and $\frac{\partial x_i}{\partial w_j} = \frac{-w_i}{w_+^2}$ if $i \neq j$, we obtain

$$\begin{aligned} \frac{\partial f(\mathbf{w})}{\partial w_j} &= \frac{1}{w_+^2} \left[ w_+ \frac{\partial \underline{f}(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{3} w_i \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \right] \\ &= \frac{1}{w_+} \left[ \frac{\partial \underline{f}(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{3} x_i \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \right] \end{aligned} \tag{A.2}$$

Using equation (A.2) with $w_j$ replaced by $\log(w_j)$ yields

$$\begin{aligned} \frac{\partial f(\mathbf{w})}{\partial \log(w_j)} = w_j \frac{\partial f(\mathbf{w})}{\partial w_j} &= \frac{w_j}{w_+} \left[ \frac{\partial \underline{f}(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{3} x_i \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \right] \\ &= x_j \left[ \frac{\partial \underline{f}(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{3} x_i \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \right] \end{aligned} \tag{A.3}$$

Proposition 13.3.5 in Pawlowsky-Glahn and Buccianti [54] tells us that

$$\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus} x_j} = x_j \left[ \frac{\partial \underline{f}(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{3} x_i \frac{\partial \underline{f}(\mathbf{x})}{\partial x_i} \right] \tag{A.4}$$

Combining this equations (A.3) and (A.4) yields the following proposition, linking the semi-log derivatives of $f$ with the directional C-derivatives of $\underline{f}$:

$$\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus} x_j} = \frac{\partial f(\mathbf{w})}{\partial \log(w_j)} \tag{A.5}$$

Let us now consider the case of a function from the simplex $\mathcal{S}^{D_X}$ of $\mathbb{R}^{D_X}$ to the simplex $\mathcal{S}^{D_S}$ of $\mathbb{R}^{D_S}$. Rewriting equation (12.6) from chapter 12 (page 163) in [54] with our present notations we have $\frac{\partial^{\oplus} \mathbf{h}(t)}{\partial t} = \mathcal{C} \exp\left(\frac{\partial \log \mathbf{h}(t)}{\partial t}\right)$.

Combining this with equation (A.5), we can define the following simplicial derivatives of $\mathbf{h}$, denoted $\frac{\partial^{\oplus} \mathbf{h}(\mathbf{x})}{\partial^{\oplus} x_j}$ as

$$\frac{\partial^{\oplus} \mathbf{h}(\mathbf{x})}{\partial^{\oplus} x_j} = \mathcal{C}\left(\exp(\frac{\partial \log \mathbf{h}(\mathbf{x})}{\partial^{\oplus} x_j})\right) = \mathcal{C}\left(\exp(\frac{\partial \log \mathbf{h}(\mathbf{x})}{\partial \log w_j})\right)$$

### A.2.3 Nullity of the sum of elasticities weighted by shares

We have to prove that $\sum_{m=1}^{D} e_{mlt}\mathbb{E}^{\oplus}S_{mt} = 0$. This is the necessary condition for new shares $S'_{mt}$, resulting from a change in $X_{lt}$, to sum up to one: $\sum_{m=1}^{D} S'_{mt} = 1 \Leftrightarrow \sum_{m=1}^{D} e_{mlt}\mathbb{E}^{\oplus}S_{mt} = 0$.

Proof:

$$\sum_{m=1}^{D} \mathbb{E}^{\oplus}S_{mt} = 1 \Leftrightarrow \sum_{m=1}^{D} \frac{\partial \mathbb{E}^{\oplus}S_{mt}}{\partial \log X_{lt}} = 0 \Leftrightarrow \sum_{m=1}^{D} \frac{\partial \mathbb{E}^{\oplus}S_{mt}}{\partial \log X_{lt}} \frac{1}{\mathbb{E}^{\oplus}S_{mt}} \mathbb{E}^{\oplus}S_{mt} = 0 \Leftrightarrow \sum_{m=1}^{D} e_{mlt}\mathbb{E}^{\oplus}S_{mt} = 0$$

### A.2.4 Elasticities of shares ratios and odds ratios

Table A.1: Elasticity of ratios of market shares $\frac{S_{jt}}{S_{j't}}$ relative to media $\check{M}_{l,t-1}$

| MCI | | CODA | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\check{M}_{t-1}$ | | $\check{M}_{C,t-1}$ | | $\check{M}_{P,t-1}$ | | $\check{M}_{R,t-1}$ | | $\check{M}_{Z,t-1}$ |
| $e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j,t-1}\right)$ | 0.0267 | $S_{C/P}$ | 0.0258 | $S_{P/C}$ | 0.0127 | $S_{R/C}$ | 0.0424 | $S_{Z/C}$ | 0.0239 |
| $e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j',t-1}\right)$ | -0.0267 | $S_{C/R}$ | 0.0246 | $S_{P/R}$ | 0.0272 | $S_{R/P}$ | 0.0208 | $S_{Z/P}$ | 0.0325 |
| $e\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{l,t-1}\right)^{*}$ | 0 | $S_{C/Z}$ | 0.0211 | $S_{P/Z}$ | 0.0044 | $S_{R/Z}$ | 0.0535 | $S_{Z/R}$ | 0.0273 |

*where $l \neq j, j'$ and $S_{C/Z}$ means $S_{Citroën,t}/S_{Others,t}$ for example.

Table A.2: Odds ratios of market shares for an increase of 10% in media $\check{M}_{l,t-1}$

| MCI | | CODA | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| For $\Delta = 10\%$ | $\check{M}_{t-1}$ | | $\check{M}_{C,t-1}$ | | $\check{M}_{P,t-1}$ | | $\check{M}_{R,t-1}$ | | $\check{M}_{Z,t-1}$ |
| $\mathrm{OR}\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j,t-1}, \Delta\right)$ | 1.0025 | $S_{C/P}$ | 1.0025 | $S_{P/C}$ | 1.0012 | $S_{R/C}$ | 1.0045 | $S_{Z/C}$ | 1.0022 |
| $\mathrm{OR}\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{j',t-1}, \Delta\right)$ | 0.9975 | $S_{C/R}$ | 1.0024 | $S_{P/R}$ | 1.0030 | $S_{R/P}$ | 1.0026 | $S_{Z/P}$ | 1.0031 |
| $\mathrm{OR}\left(\frac{S_{jt}}{S_{j't}}, \check{M}_{l,t-1}, \Delta\right)^{*}$ | 0 | $S_{C/Z}$ | 1.0020 | $S_{P/Z}$ | 1.0007 | $S_{R/Z}$ | 1.0054 | $S_{Z/R}$ | 1.0028 |

*where $l \neq j, j'$ and $S_{C/Z}$ means $S_{Citroën,t}/S_{Others,t}$ for example.

Table A.3: Elasticity of ratios $\frac{S_{jt}}{g(S_{-jt})}$ relative to $\check{M}_{l,t-1}$

| MCI | | CODA | | | | |
|---|---|---|---|---|---|---|
| | | | $\check{M}_{C/g(-C)}$ | $\check{M}_{P/g(-P)}$ | $\check{M}_{R/g(-R)}$ | $\check{M}_{Z/g(-Z)}$ |
| $e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{M}_{j,t-1}\right)$ | 0.0267 | $S_{C/g(-C)}$ | **0.0239** | -0.0022 | -0.0176 | -0.0040 |
| | | $S_{P/g(-P)}$ | -0.0106 | **0.0148** | 0.0112 | -0.0154 |
| $e\left(\frac{S_{jt}}{g(S_{-jt})}, \check{M}_{l,t-1}\right)^{*}$ | 0 | $S_{R/g(-R)}$ | -0.0090 | -0.0215 | **0.0389** | -0.0085 |
| | | $S_{Z/g(-Z)}$ | -0.0043 | 0.0089 | -0.0324 | **0.0279** |

*where $l \neq j$.

$S_{C/g(-C)}$ means $\frac{S_{Ct}}{g(S_{-Ct})}$, where $g(S_{-Ct})$ is the geometric mean of others shares than Citroën.

## A.3 Appendix (Chapter 3)

### A.3.1 Aligning media investments on registrations

The following table explains how the media investments are weighted in order to be aligned on the registration time, such that the weighted media at time $t$ are those potentially seen by customers who purchase a new car for which the registration happens at time $t$.

Table A.4: Weighted media

| Registration time | Purchase time | % Sales registered in $t$ | Media | Weighted Media |
|---|---|---|---|---|
| $t$ | $t$ | 31% | $M_t$ | $WM_t = 0.31M_t$ |
| | $t-1$ | 34% | $M_{t-1}$ | $+0.34M_{t-1}$ |
| | $t-2$ | 21% | $M_{t-2}$ | $+0.21M_{t-2}$ |
| | $t-3$ | 14% | $M_{t-3}$ | $+0.14M_{t-3}$ |

### A.3.2 Media investments by brand

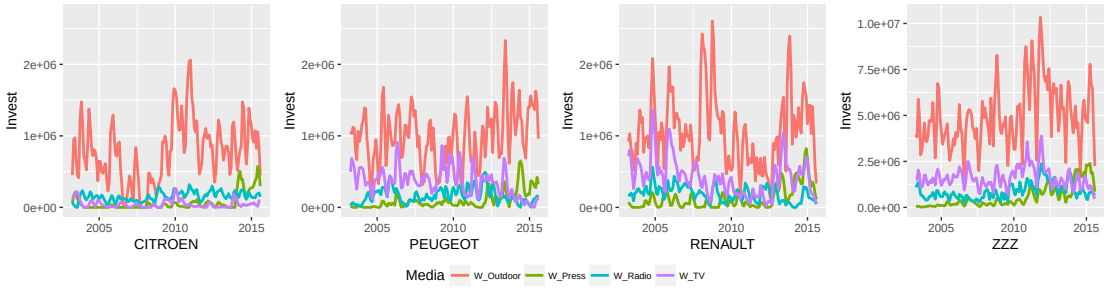


Figure A.1: Media investments by brand and by channel (in euro)

### A.3.3 Relationship between explanatory variables and dependent variable, in volume and in share

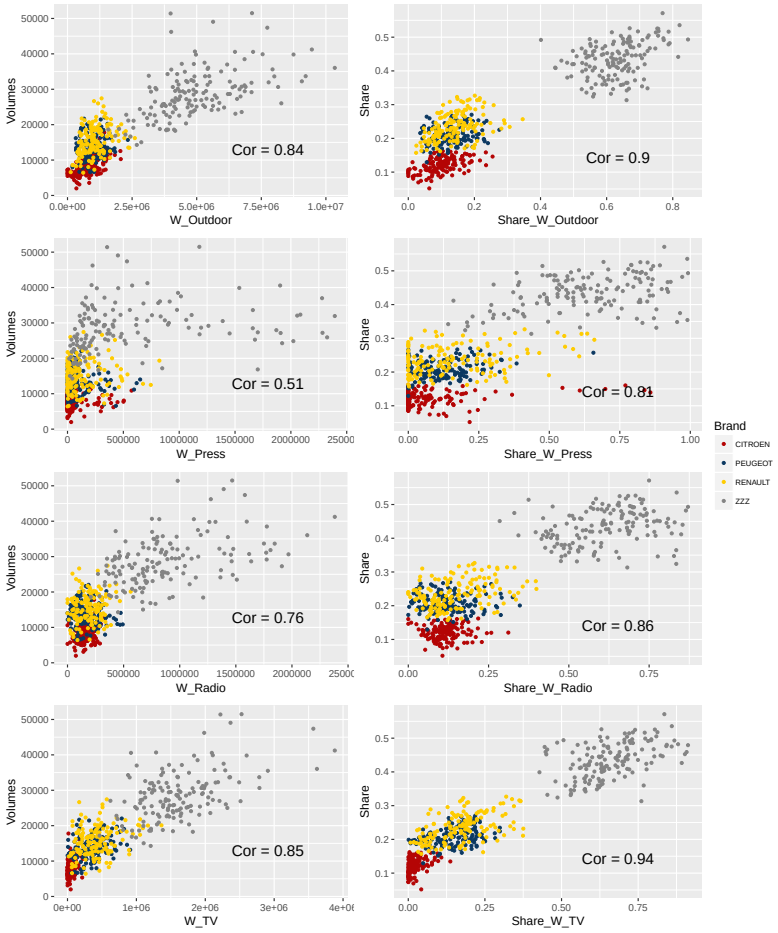


Figure A.2: Correlations between sales and media (in volume and in share)

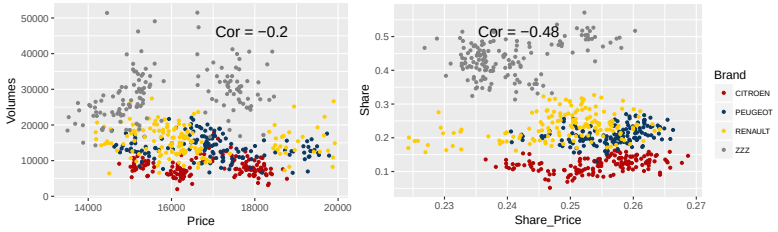


Figure A.3: Correlations between sales and price (in volume and in share)

### A.3.4 $R^2$ values for different adstock parameters

The following graphs show the $R^2$ values of MCI and CODA model, for the different combinaisons (10000) of decay parameters $\lambda_c$ of the different advertising channels. The black horizontal line is just here to help the reader to visualize the parameters values giving the higher $R^2$ values.

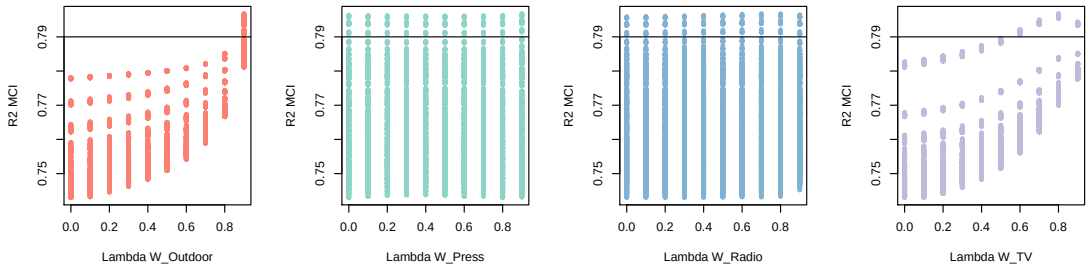


Figure A.4: $R^2$ values of MCI model for different values of the adstock parameters of channels

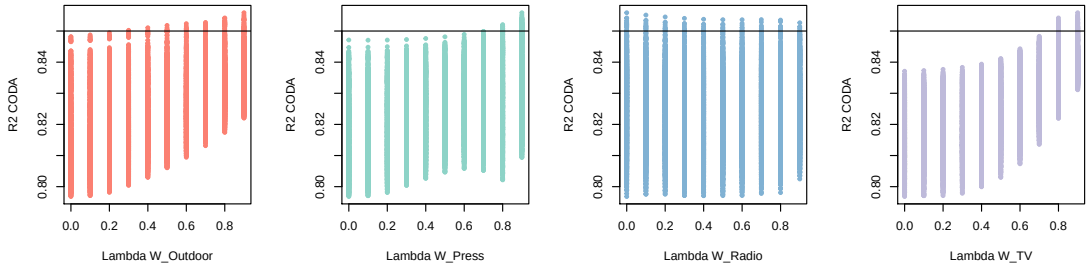


Figure A.5: $R^2$ values of CODA model for different values of the adstock parameters of channels

### A.3.5 Advertising half life in an attraction model with carryover effects

In a classical Koyck model as in equation (3.1), one can show easily that the half life of the advertising is equal to $\log(1-\theta)/\log(\lambda) - 1$ where $\theta = 1/2$. Indeed, if we consider the advertising investment $M_t$ at time $t$, ignoring all other investments, we are looking

for the period $n$ such that a fraction $\theta$ of the long run impact of $M_t$ occurs in $n$ periods:

$$\begin{aligned}
\theta M_t &= (1-\lambda)\sum_{\tau=0}^{n}\lambda^{\tau}M_t \\
\Leftrightarrow \theta &= (1-\lambda)\sum_{\tau=0}^{n}\lambda^{\tau} = (1-\lambda^{n+1}) \\
\Leftrightarrow \lambda^{n+1} &= 1-\theta \\
\Leftrightarrow (n+1)\log\lambda &= \log(1-\theta) \\
\Leftrightarrow n &= \frac{\log(1-\theta)}{\log\lambda} - 1
\end{aligned}$$

In the case of the attraction model with carryover effects, as in equations (3.2) and (3.4), we can make the same demonstration based on the ILR transformed models of equations (3.3) and (3.5) which are similar to equation (3.1). Thus, it gives exactly the same result.

### A.3.6 ANOVA of CODAAd

Table A.5: Analysis of variance table for CODAAd model

| | Df | Pillai | approx F | num Df | den Df | Pr(>F) |
|---|---|---|---|---|---|---|
| (Intercept) | 1 | 0.99 | 3883.01 | 3 | 101 | 0.0000*** |
| ilr(OutdoorAd) | 3 | 1.35 | 28.05 | 9 | 309 | 0.0000*** |
| ilr(PressAd) | 3 | 0.69 | 10.18 | 9 | 309 | 0.0000*** |
| ilr(RadioAd) | 3 | 0.35 | 4.55 | 9 | 309 | 0.0000*** |
| ilr(TVAd) | 3 | 0.58 | 8.30 | 9 | 309 | 0.0000*** |
| ilr(Price) | 3 | 0.23 | 2.91 | 9 | 309 | 0.0026*** |
| SI | 1 | 0.04 | 1.30 | 3 | 101 | 0.2799 |
| Residuals | 103 | | | | | |

### A.3.7 Prediction intervals for market shares

In order to compute the confidence and prediction intervals of market shares for each prediction time, we use the fact that the confidence and prediction regions of the ILR coordinates are ellipsoids in dimension $D-1=3$.

We denote $\mathbf{S}^* = ilr(\mathbf{S}) \sim \mathcal{N}_{D-1}(\mu^*, \mathbf{\Sigma}^*)$ where $\mathbf{\Sigma}^*$ is the covariance matrix of ILR error terms (estimated by the empirical covariance matrix of ILR residuals), and $\mathbf{S}^* = \mathbf{V}'\log\mathbf{S}$, where $\mathbf{V}$ is the balance matrix used for the ILR transformation.

The covariance matrix of the predicted ILR coordinates, $\hat{\mathbf{S}}^*$, is given by

$$Var_P\hat{\mathbf{S}}^* = Var\mathbf{X}^*\hat{\beta}^* + Var\epsilon^*,$$

where $Var\mathbf{X}^*\hat{\beta}^* = \mathbf{X}^*Var\hat{\beta}^*\mathbf{X}^{*\prime}$.

The points $\mathbf{S}^*$ in the confidence and prediction ellipsoids around the prediction $\hat{\mathbf{S}}^*$ are such that

$$H = (\mathbf{S}^* - \hat{\mathbf{S}}^*)' \widehat{\Sigma_P}^{*-1} (\mathbf{S}^* - \hat{\mathbf{S}}^*)$$

follows a Hotelling distribution of parameters $p = D - 1$ and $n - p$ with $n$ the number of observations (see Friendly et al. [17]), which is equivalent to follow a $\frac{p(n-1)}{n-p}$ Fisher distribution.

Thus, the ellipsoid of $\hat{\mathbf{S}}^*$ at a confidence level of $1 - \alpha$ can be written

$$\mathbb{P}((\mathbf{S}^* - \hat{\mathbf{S}}^*)' \widehat{\Sigma_P}^{*-1} (\mathbf{S}^* - \hat{\mathbf{S}}^*) \leq c_\alpha) = 1 - \alpha,$$

where $\widehat{\Sigma_P}^{*-1} = \widehat{Var_P}\hat{\mathbf{S}}^*$ for the prediction ellipsoid[2] and $c_\alpha$ verifies $\mathbb{P}(\mathcal{F}_{p,n-p} \leq \frac{n-p}{p(n-1)} c_\alpha) = 1 - \alpha$, meaning that $c_\alpha$ is the $1 - \alpha\%$ quantile of $\mathcal{F}_{p,n-p}$.

In order to visualize the prediction region in the simplex (which is the back transformed ellipsoid from the ILR coordinates space), we are going to simulate points on the ellipsoid, and back transform them into the simplex as follows.

We first rewrite the ellipsoid equation:

$$\begin{aligned} & (\mathbf{S}^* - \hat{\mathbf{S}}^*)' \widehat{\Sigma_P}^{*-1} (\mathbf{S}^* - \hat{\mathbf{S}}^*) = c_\alpha \\ \Leftrightarrow\ & (\widehat{\Sigma_P}^{*-1/2} (\mathbf{S}^* - \hat{\mathbf{S}}^*))' (\widehat{\Sigma_P}^{*-1/2} (\mathbf{S}^* - \hat{\mathbf{S}}^*)) = c_\alpha \\ & \Leftrightarrow ||\widehat{\Sigma_P}^{*-1/2} (\mathbf{S}^* - \hat{\mathbf{S}}^*)||_E^2 = c_\alpha \end{aligned} \tag{A.6}$$

It is equivalent to say that $\widehat{\Sigma_P}^{*-1/2}\mathbf{S}^*$ belongs to an hypersphere of center $\widehat{\Sigma_P}^{*-1/2}\hat{\mathbf{S}}^*$ and radius $\sqrt{c_\alpha}$ (where the "$E$" means Euclidean norm).

If $D - 1 = 3$, the points $\mathbf{U} = (sin(\Phi)cos(\Theta),\ sin(\Phi)sin(\Theta),\ cos(\Phi))'$, where $\Phi$ and $\Theta$ are independently uniformly distributed in $[0, 2\pi]$, are uniformly distributed on the sphere of center 0 and radius 1. Thus, the points $\mathbf{S}^*$ such that:

$$\widehat{\Sigma_P}^{*-1/2}\mathbf{S}^* = \widehat{\Sigma_P}^{*-1/2}\hat{\mathbf{S}}^* + \sqrt{c_\alpha}\mathbf{U} \ \Leftrightarrow\ \mathbf{S}^* = \hat{\mathbf{S}}^* + \sqrt{c_\alpha}\widehat{\Sigma_P}^{*1/2}\mathbf{U}$$

are distributed on the ellipsoid.

In order to come back to the simplex, we use the inverse ILR transformation. The points $ilr^{-1}(\mathbf{S}^*)$ are distributed on the prediction region in the simplex $\mathcal{S}^D$. Equation (A.6) is equivalent to the following equation:

$$||ilr^{-1}(\widehat{\Sigma_P}^{*-1/2}\mathbf{S}^*) \ominus ilr^{-1}(\widehat{\Sigma_P}^{*-1/2}\hat{\mathbf{S}}^*)||_A^2 = c_\alpha, \tag{A.7}$$

where the "$A$" stands for the Aitchison norm here. Then, we can say that the ellipsoid for ILR coordinates corresponds to an ellipsoid in the simplex, because equation (A.7) can also be written as:

$$||\widehat{\Sigma_P}^{-1/2} \boxdot (\mathbf{S} \ominus \hat{\mathbf{S}})||_A^2 = c_\alpha$$

[2]Note that to built the confidence ellipsoid for the mean share, we take $\widehat{\Sigma_P}^{*-1} = \widehat{Var}\epsilon^*$.

Indeed, this is true because

$$\begin{aligned} ilr^{-1}(\widehat{\Sigma_P}^{*-1/2}\mathbf{S}^*) &= \mathcal{C}(\exp \mathbf{V}\widehat{\Sigma_P}^{*-1/2}\mathbf{S}^*) = \mathcal{C}(\exp \mathbf{V}\widehat{\Sigma_P}^{*-1/2}\mathbf{V}' \log \mathbf{S}) \\ &= \mathcal{C}(\exp \widehat{\Sigma_P}^{-1/2} \log \mathbf{S}) = \widehat{\Sigma_P}^{-1/2} \boxdot \mathbf{S}, \end{aligned}$$

where $\widehat{\Sigma_P}^{-1/2} = \mathbf{V}\widehat{\Sigma_P}^{*-1/2}\mathbf{V}'$ is a $D \times D$ matrix.

In our case, we generate 10000 couples of angles $(\Phi, \Theta)$ and we compute the 10000 corresponding $\mathbf{S}^*$. Applying the ILR inverse transformation to these $\mathbf{S}^*$ allows us to represent the image of this ellipsoid in the simplex $\mathcal{S}^D$, for each prediction time $t$ (see Figure 3.10).

# Contents

# List of Figures

# Resume

## Resume (English version)

The aim of this CIFRE thesis, realized with the market research institute BVA in collaboration with the automobile manufacturer Renault, is to build a model in order to measure the impact of media investments of several channels (television, outdoor, etc.) on the brands' market shares, taking into account the competition et the potential cross effects and synergies between brands, as well as controlling for average price and regulatory context (scrapping incentive).

Market share models have been developed in the marketing literature, especially the GMCI model (generalized multiplicative competitive interaction model), inspired from the aggregated conditional MNL (multinomial logit) model. In the statistical literature, the compositional data analysis (CODA) allows to analyze share data respecting their nature (a vector of D shares subject to the unit sum constraint is a composition and belongs to the dimension D simplex space). Regression models for dependent and explanatory compositional variables exist but are rarely used in practice. Finally, the Dirichlet covariate model allows to model a simplex valued dependent variable.

In the first chapter, these different models avec compared from a theoretical and empirical point of view. It is shown that all of them can be expressed with a similar formulation using the notions of attraction and of simplicial expected value. The GMCI model appears to be a particular case of the CODA model, such that these two specifications can be combined into a unique model. The complexity of Dirichlet and CODA models turns out to be necessary in order to capture the diversity of competitive relationships.

In the second chapter, emphasis is given to the interpretation of models which is not very well developed in the CODA literature. Different types of interpretations are presented, but it is demonstrated that the calculation of the elasticities of market shares relative to media investments in particularly relevant from a mathematical point of view and from a practical perspective. Indeed, we prove that elasticities are consistent with C-derivatives of simplex valued functions of another simplex. Moreover, these elasticities can be easily interpreted by car manufacturers and can be used for advertising budgeting optimization (Dorfman-Steiner theorem).

In the third chapter, a practical application to the B segment of the French automobile market is presented for the purpose of measuring the impact of the different advertising channels on the market shares of the three leaders of this segment and of the group of other brands, taking into account the lagged effects of advertising (adstock function) and the competitive cross effects. The media investments elasticity of the brand market share varies from one brand to another and from one channel to another. Synergies between some brands can be highlighted.

The last chapter opens the discussion on different directions to be explored in order to improve the proposed model and to provide further answers to the considered issue.

# Résumé (version française)

L'objectif de cette thèse CIFRE, réalisée avec la société d'études et conseil BVA et en collaboration avec le constructeur automobile Renault, est de construire un modèle permettant de mesurer l'impact des investissements media à travers différents canaux (télévision, affichage, etc.) sur les parts de marché de différentes marques, en prenant en compte la concurrence et les potentiels effets croisés et synergies entre ces marques, ainsi qu'en contrôlant pour le prix et le contexte réglementaire (i.e. prime à la casse).

Des modèles de parts de marchés ont été développés dans la littérature marketing, notamment le modèle GMCI (generalized multiplicative competitive interaction model), inspiré du modèle MNL (multinomial logit) conditionnel agrégé. Dans la littérature statistique, l'analyse des données de composition (CODA) permet d'étudier des données de parts en respectant leur nature (un vecteur de D parts soumises à la contrainte de somme unitaire est une composition et appartient au simplexe de dimension D). Des modèles de régression pour variables dépendante et explicatives compositionnelles existent mais sont peu utilisés en pratique. Enfin, le modèle de régression de Dirichlet permet de modéliser une variable dépendante appartenant au simplexe.

Dans le premier chapitre, ces différents modèles sont comparés théoriquement et empiriquement. On montre notamment qu'ils peuvent tous être exprimés sous une forme similaire en utilisant les notions d'attraction et d'espérance dans le simplexe. Le modèle GMCI se trouve être un cas particulier du modèle CODA, de telle sorte que ces deux spécifications peuvent être combinées au sein d'un même modèle. La complexité des modèles Dirichlet et CODA se révèle être nécessaire pour capturer la diversité des relations de concurrence.

Dans le deuxième chapitre, l'accent est mis sur l'interprétation, peu développée dans la littérature CODA. Différents types d'interprétation sont présentés, mais nous montrons que le calcul d'élasticités des parts de marché aux investissements media est particulièrement pertinent d'un point de vue mathématique et applicatif. En effet, nous prouvons que les élasticités sont cohérentes avec les C-dérivées de fonctions d'un simplexe à valeurs dans un autre simplexe. De plus, ces élasticités peuvent être facilement interprétées par les constructeurs automobiles et peuvent être utilisées pour l'optimisation de la budgétisation de la publicité (théorème de Dorfman et Steiner).

Dans le troisième chapitre, une application concrète au segment B du marché automobile français est présentée pour mesurer l'impact des différents canaux publicitaires sur les parts de marché des trois leaders du segment et du groupe des autres marques, en tenant compte des effets retards de la publicité (fonction d'adstock) et des effets croisés de la concurrence.L'élasticité des parts de marché des différentes marques aux investissements media est variable d'une marque à l'autre et d'un canal à l'autre. Des relations de synergies entre certaines marques peuvent être mises en lumière.

Dans un dernier chapitre, nous ouvrons sur les pistes à explorer pour améliorer le modèle proposé et apporter des réponses complémentaires à notre problématique.

## Impact of media investments on brands' market shares: a compositional data analysis approach

The aim of this CIFRE thesis, realized with the market research institute BVA in collaboration with the automobile manufacturer Renault, is to build a model in order to measure the impact of media investments of several channels (television, outdoor, etc.) on the brands' market shares, taking into account the competition and the potential cross effects and synergies between brands, as well as accounting for the price, the regulatory context (scrapping incentive), and the lagged effects of advertising.

We have drawn from marketing and statistical literatures to develop, compare and interpret several models which respect the unit sum constraint of market shares. A practical application to the French automobile market is presented, for which it is shown that brands' market shares are more or less sensitive to advertising investments made in each channel, and that synergies between brands exist.

**Keywords: market shares, compositional data, automobile market, media investments, elasticities.**

## Impact des investissements media sur les parts de marché des marques : une approche par analyse des données de composition

L'objectif de cette thèse CIFRE, réalisée avec la société d'études de marché BVA en collaboration avec le constructeur automobile Renault, est de mesurer l'impact des investissements media pour différents canaux (télévision, affichage, etc.) sur les parts de marché de différentes marques, en prenant en compte la concurrence et les potentiels effets croisés et synergies entre ces marques, ainsi qu'en tenant compte du prix des véhicules, du contexte réglementaire (i.e. prime à la casse), et des effets retard de la publicité.

Nous avons puisé dans les littératures marketing et statistique pour développer, comparer et interpréter plusieurs modèles qui respectent la contrainte de somme unitaire des parts de marché. Une application concrète au marché automobile français est présentée, pour laquelle nous montrons que les parts de marché des marques sont plus ou moins sensibles aux investissements publicitaires consentis dans chaque canal, et qu'il existe de synergies entre certaines marques.

**Mots-clefs: part de marché, données de composition, marché automobile, investissements media, élasticités.**