



**UNIVERSITÉ
DE GENÈVE**

**FACULTÉ DES SCIENCES
ÉCONOMIQUES ET SOCIALES**

Département des sciences
économiques

Mémoire de Diplôme en Econométrie
Analyse économétrique de la demande d'électricité solaire:
le cas de la bourse solaire de Genève

Giancarlo Fiorito

Mai 2002

Introduction

Il existe en Suisse une sensibilité manifeste aux problèmes de l'environnement. Les instances fédérales et cantonales ont souligné telle volonté par plusieurs mesures :

- la Loi sur l'énergie du 26 juin 1998, qui considère les approvisionnement, l'utilisation rationnelle, la promotion des nouvelles énergies et explicite dans les principes la volonté populaire d'accroître l'emploi d'énergie renouvelable ;
- l'Ordonnance sur l'énergie du 7 décembre 1998, disposant sur les producteurs indépendants, les expertises et le financement des projets ;
- la Loi cantonale sur l'énergie du 18 septembre 1986, qui « vise à contribuer à un approvisionnement énergétique suffisant, diversifié, sûr, économique et compatible avec la protection de l'environnement »¹.

Parmi les mesures « fortes » d'impulsion aux énergies renouvelables se situe la bourse solaire, qui est, très schématiquement, une initiative d'encouragement aux projets de construction des centrales photovoltaïques. Les producteurs peuvent revendre au distributeur local d'électricité (SIG) l'énergie excédentaire et ce dernier met en vente cette énergie à ses abonnés. L'énergie solaire pouvant venir de centrales privés ou publiques elle vaut un prix plus haut que le prix du kWh courant².

Les premières expériences de bourse solaire ont été faites dans le canton de Zurich, où une enquête menée en 1995 sur 3500 abonnés du distributeur local EWZ concernant le projet d'une centrale, a montré que ces derniers sont favorables à l'énergie solaire. Les interviewés répondaient à la question s'ils étaient prêts à payer un surplus pour avoir telle énergie et, malgré l'incertitude concernant la consistance des résultats vis-à-vis de la réelle disposition à payer, l'étude a conduit ensuite à l'établissement de la *ewz-Solarstrombörse*.

Le concept est simple : EWZ achète l'énergie solaire sur le marché libre et la revend sans coûts additionnels aux clients au prix moyen d'achat. Ce prix, qui est supérieur à celui de la facture normale³, permet l'amortissement des centrales. EWZ constate le succès – en partie inattendu - de l'initiative, confirmant qu'une partie de la population suisse est prête à payer plus chère une partie de son électricité pour autant qu'elle soit « verte »⁴. La bourse a permis d'autre part une co-opération avec des associations environnementales. Il existe à Zurich 43 installations PV délivrant 1.65 MW, qui fournissent du courant à 5700 clients particuliers.

A Genève le mécanisme de la bourse solaire genevoise est le suivant : le distributeur propose des tranches d'énergie solaire de 100 kWh chacune à un prix de 100 CHF (1CHF/kWh). Les clients du canton peuvent dès lors souscrire une ou plusieurs tranches que le distributeur s'engage à fournir aux consommateurs – si possible - dans l'année de souscription du contrat⁵. Les Services Industriels de Genève (SIG) rachètent l'énergie solaire depuis les centrales PV présentes sur le canton (les installations PV peuvent être de propriété du distributeur ou privées) pour l'injecter dans le réseau. Les SIG payent aux producteurs le même prix qu'ils

¹ Sans formuler de manière explicite des dispositions sur l'énergie solaire, la loi dit (Art. 16): « Les projets de constructions de bâtiments des collectivités et établissements de droit public doivent comporter une étude complète d'une possibilité d'approvisionnement énergétique fondée, tout ou en partie sur des énergies renouvelables ».

² *Strictu sensu* il ne s'agit pas d'une bourse puisque le prix est fixe.

³ 95 ct/kWh en moyenne

⁴ Voir document non publié « *ewz-Solarstrombörse* », contact : annina.jaggi@ewz-stzh.ch.

⁵ En pratique, le distributeur injecte à n'importe quel moment les kWh PV souscrits dans le réseau, l'on ne pourrait évidemment pas « véhiculer » les électrons provenant de la station PV vers un client particulier.

demandent aux consommateurs, il n'y a donc aucun bénéfice matériel de la part du distributeur d'électricité qui met pratiquement le réseau « à disposition ».

Le mécanisme est vertueux dans deux sens. Du côté de l'offre l'installation de panneaux PV par des particuliers devient rentable puisque ces derniers ont la garantie du rachat de tout surplus par le distributeur local⁶. Pour ce qui touche à la demande, la bourse en détermine quantité et prix et, dans un sens, elle « l'emmené au marché » par le réseaux conventionnel de l'énergie.

Ce mécanisme doit pouvoir engendrer un effet sur la production des panneaux PV en permettant à l'industrie du solaire l'avancement sur la courbe d'apprentissage (*learning curve*) sans passer par des subventions, qui sont souvent source d'inefficacité.

La bourse solaire se révèle donc une idée de succès pour stimuler l'entrée des renouvelables dans le marché de l'électricité par un effet de « traction par la demande » (*demand pull*) qui rend viable (et visible) la technologie PV.

La Bourse nous fournit donc des indications quant à la disponibilité à payer des consommateurs suisses pour l'énergie renouvelable ou, dans le cas du PV, à émissions zéro⁷. Cela est d'autant plus d'actualité car à présent d'autres produits énergétiques environnementalement « qualifiés » entrent sur le marché de l'énergie suisse. Ces électricités sont caractérisées par des modes de production « doux » envers l'environnement : il s'agit du *courant vert*.⁸

⁶ Les producteurs ayant la garantie du rachat du surplus électrique de la part du distributeur à 1 CHF/kWh, il devient intéressant d'investir en une installation PV lorsqu'il y a de l'espace à disposition pour les panneaux, tels les propriétaires des villas.

⁷ En excluant l'énergie grise consommée dans la production des panneaux, qui est importante.

⁸ Un tableau avec des exemples de courant vert est présenté en annexe

1 – L'intérêt d'une étude quantitative pour la bourse solaire

L'objectif principal de notre étude est l'estimation de modèles de variables qualitatives et de dénombrement avec des données longitudinales, mais aussi de décrire d'une manière plus générale le mécanisme d'adhésion à un marché nouveau et de proposer des lignes de recherche pour des études ultérieures, notamment pour la prise en compte de la dimension temporelle.

Cette étude de la bourse solaire genevoise est axée sur une approche quantitative, qui vise en particulier à caractériser le comportement du consommateur qui a déjà choisit d'adhérer à la bourse. Les données à disposition concernent 654 clients de la bourse¹.

1.1 Déficit d'offre

Les premiers chiffres fournis par les Services Industriels de Genève (SIG) nous indiquent qu'en fin d'été 2001 il y a 220.000 kWh souscrits (2.200 tranches), cette demande est confrontée à une production réalisée de 90.000 kWh. Les prévisions en début d'année 2001 estimaient une production attendue de 180.000, qui n'a pas été atteinte à cause du faible ensoleillement de l'été ainsi qu'au retard de la centrale du Parking de l'étoile.

Les SIG ont ainsi pu livrer en 2001 environs 21 % de l'électricité solaire souscrite. Nous sommes donc confrontés à une demande excédentaire d'électricité solaire, qui peut toutefois engendrer un effet de mécontentement si les adhérents à la bourse devaient juger trop faible l'effort du distributeur à « garder ses promesses ».

Nous devons également remarquer que l'excès de demande est du au *nombre* d'adhérent plus qu'aux *quantités* souscrites, car la plupart des clients ont souscrit seulement une tranche (68.5%).

Mais, est-ce que la bourse marche trop bien ? Au-delà de la simple description de la clientèle solaire, il existe, nous croyons, un intérêt à analyser le comportement des adhérents par une approche économétrique, notamment pour ce qui concerne leurs caractéristiques de tarification et de consommation, conscients du fait que les données disponibles ne seront en tout cas pas suffisant pour expliquer parfaitement le processus du choix d'adhésion quant à la décision de souscrire et combien.

1.2 La campagne d'information

Au printemps 2000 les SIG ont lancé une campagne publicitaire pour la promotion de la Bourse Solaire en envoyant à la majorité des usagers du Canton de Genève une brochure explicative de la Bourse incluant un coupon à retourner où le consommateur manifeste sa volonté d'adhésion et indique le nombre de tranches qu'il souhaite souscrire. En cas de réponse positive du client le contrat d'adhésion est ensuite envoyé.

La campagne d'information, démarré en mars-avril 2000 avec une brochure et des stands (foire de Genève, centres commerciaux), pour continuer avec l'affichage et le mailing en été.

¹ Données au 6 septembre 2001.

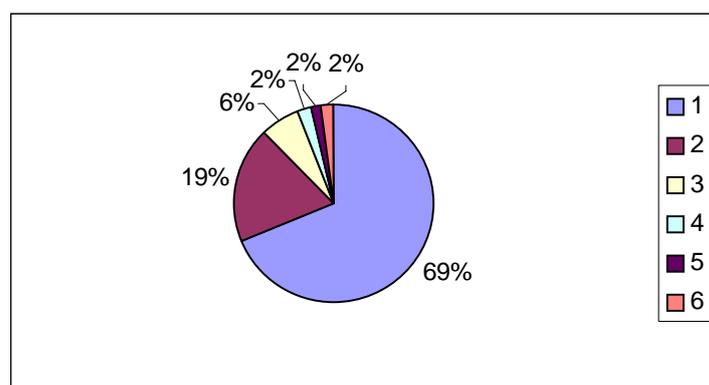
En été 2001 il y a eu l'inauguration de la centrale PV de M. Dupraz (10 kW) avec conférence de presse et en décembre 2001 inauguration de la centrale du Parking de l'étoile qui emmène 140 kW supplémentaires (depuis : 9 clients en plus pour 15 tranches).

Notre approche peut intégrer dans le futur d'autres *panachés* énergétiques, qui vont être proposés par les producteurs et distributeurs dans le marché libéralisé de l'électricité. « Courant-vert » est le nom pour indiquer l'électricité d'origine contrôlée et, nous croyons, dans une perspective dynamique, nous pourrions voir dans le futur la bourse solaire comme un « moteur » pour d'autres sources renouvelables. En effet, l'électricité d'origine photovoltaïque peut être vue comme un courant vert « extrême » : cette énergie est à zéro émissions et coût maximal.

2 Analyse exploratoire des données

2.1 Analyse descriptive

Le bilan des SIG pour l'année 2001 nous révèle que seulement 21% de l'énergie PV souscrite a été livrée, à cause du retard d'un projet important (voir Ch. 1). En septembre 2001 les adhérents à la bourse solaire sont 659 ; ils sont distribués comme suit :



Graphique 2.1 – Répartition de la clientèle solaire selon les tranches achetées

En regardant le genre des adhérents, nous remarquons que les individus ont en grande majorité ($\approx 2/3$) souscrit une seule tranche d'énergie solaire. Parmi tous les adhérents 243 sont des femmes (« Madame » comme signataire du contrat) ce qui correspond à un pourcentage de 36.8 % ; 31 effectifs (4.5%) signent « Madame & Monsieur » ce qui indique une décision conjointe ; 15 souscriptions sont au nom d'une entreprise ou collectivité publique (Communes etc.) et, peut-être remarquablement, cela n'indique pas forcément des quantités plus grandes : 9 individus ont 1 tranche seulement – si on fait exception de l'acheteur des 700 tranches (collectivité publique).

Pour plus de détails :

| TRANCHES | Madame | Mad. & Mon. | Monsieur | Soc | Total |
|----------|--------|-------------|----------|-----|-------|
| 1 | 176 | 20 | 246 | 9 | 451 |
| 2 | 46 | 8 | 67 | | 121 |
| 3 | 11 | 4 | 23 | 4 | 42 |
| 4 | 7 | 1 | 7 | | 15 |
| 5 | | 1 | 8 | 1 | 10 |
| 6 | 1 | | 2 | | 3 |
| 7 | 1 | | 1 | | 2 |
| 8 | 1 | | 1 | | 2 |
| 10 | | 1 | 1 | | 2 |
| 12 | 1 | | | | 1 |
| 15 | | 1 | | | 1 |
| 20 | | | 1 | | 1 |
| 50 | | 1 | | | 1 |
| 700 | | | | 1 | 1 |
| Total | 244 | 37 | 357 | 15 | 653 |

Tableau 2.1 – Répartition de la clientèle solaire selon les genre

Pour ce qui concerne la localité, 346 clients de la bourse sont résidents dans la commune de Genève (52.5%) ; pour les autres adhérents nous avons opéré un regroupement logique car

certains gens ont marqué le code postal au lieu de la commune ou encore le lieu de résidence¹.
 Nous avons obtenu le regroupement suivant² :

| <i>Commune</i> | <i>Pop</i> | <i>Pop %</i> | <i>Adhérent</i> | <i>Tranches*</i> | <i>Tranches%</i> |
|---------------------------|------------|--------------|-----------------|------------------|------------------|
| <i>Aire-la-Ville</i> | 771 | 0.19 | 0 | 0 | 0.00 |
| <i>Anières</i> | 1991 | 0.48 | 5 | 7 | 0.39 |
| <i>Avully</i> | 1737 | 0.42 | 3 | 54 | 3.00 |
| <i>Avusy</i> | 1176 | 0.28 | 0 | 0 | 0.00 |
| <i>Bardonnex</i> | 2097 | 0.51 | 11 | 16 | 0.89 |
| <i>Bellevue</i> | 1770 | 0.43 | 2 | 3 | 0.17 |
| <i>Bernex</i> | 8709 | 2.11 | 18 | 29 | 1.61 |
| <i>Carouge</i> | 17425 | 4.21 | 16 | 27 | 1.50 |
| <i>Cartigny</i> | 762 | 0.18 | 2 | 4 | 0.22 |
| <i>Céligny</i> | 624 | 0.15 | 0 | 0 | 0.00 |
| <i>Chancy</i> | 915 | 0.22 | 2 | 5 | 0.28 |
| <i>Chêne-Bougeries</i> | 9739 | 2.35 | 20 | 32 | 1.78 |
| <i>Chêne-Bourg</i> | 7232 | 1.75 | 5 | 8 | 0.44 |
| <i>Choulex</i> | 948 | 0.23 | 2 | 4 | 0.22 |
| <i>Collex-Bossy</i> | 1295 | 0.31 | 1 | 1 | 0.06 |
| <i>Collonge-Bellerive</i> | 6336 | 1.53 | 4 | 7 | 0.39 |
| <i>Cologny</i> | 4748 | 1.15 | 12 | 19 | 1.06 |
| <i>Confignon</i> | 3008 | 0.73 | 7 | 8 | 0.44 |
| <i>Corsier</i> | 1699 | 0.41 | 2 | 3 | 0.17 |
| <i>Dardagny</i> | 1296 | 0.31 | 0 | 0 | 0.00 |
| <i>Genève</i> | 177692 | 42.96 | 355 | 583 | 32.42 |
| <i>Genthod</i> | 2211 | 0.53 | 3 | 3 | 0.17 |
| <i>Grand-Saconnex</i> | 8348 | 2.02 | 3 | 3 | 0.17 |
| <i>Gy</i> | 367 | 0.09 | 0 | 0 | 0.00 |
| <i>Hermance</i> | 818 | 0.20 | 2 | 3 | 0.17 |
| <i>Jussy</i> | 1114 | 0.27 | 2 | 3 | 0.17 |
| <i>Laconnex</i> | 534 | 0.13 | 0 | 0 | 0.00 |
| <i>Lancy</i> | 25620 | 6.19 | 26 | 41 | 2.28 |
| <i>Meinier</i> | 1693 | 0.41 | 7 | 9 | 0.50 |
| <i>Meyrin</i> | 19655 | 4.75 | 17 | 24 | 1.33 |
| <i>Onex</i> | 16479 | 3.98 | 22 | 34 | 1.89 |
| <i>Perly-Certoux</i> | 2775 | 0.67 | 3 | 3 | 0.17 |
| <i>Plan-les-Ouates</i> | 6803 | 1.64 | 10 | 26 | 1.45 |
| <i>Pregny-Chambésy</i> | 3100 | 0.75 | 4 | 5 | 0.28 |
| <i>Presinge</i> | 546 | 0.13 | 1 | 1 | 0.06 |
| <i>Puplinge</i> | 2054 | 0.50 | 1 | 2 | 0.11 |
| <i>Russin</i> | 404 | 0.10 | 1 | 4 | 0.22 |
| <i>Satigny</i> | 2773 | 0.67 | 6 | 8 | 0.44 |
| <i>Soral</i> | 597 | 0.14 | 2 | 2 | 0.11 |
| <i>Thônex</i> | 12452 | 3.01 | 13 | 21 | 1.17 |
| <i>Troinex</i> | 1988 | 0.48 | 4 | 6 | 0.33 |
| <i>Vandoeuvres</i> | 2334 | 0.56 | 3 | 3 | 0.17 |
| <i>Vernier</i> | 29241 | 7.07 | 66 | 48 | 2.67 |

¹ Par ex. « Les Acacias » ou « Genève 11 ».

² En excluant une collectivité publique de la Commune de Bernex qui a souscrit 700 tranches.

| | | | | | |
|----------------|---------------|------|------------|-------------|------|
| <i>Versoix</i> | 10826 | 2.62 | 10 | 15 | 0.83 |
| <i>Veyrier</i> | 8943 | 2.16 | 13 | 24 | 1.33 |
| Totaux | 413645 | | 687 | 1798 | |

Tableau 2.2 – Répartition de la clientèle solaire selon la commune

Nous devons remarquer que des particuliers ont souscrit 50 et 20 tranches d'énergie solaire et que sans enlever le plus gros consommateur de la commune de Bernex cette commune afficherait plus que 40% de l'électricité solaire du canton.

Pour vérifier si la distribution géographique de la population jouait un rôle, nous avons calculé le coefficient de corrélation entre le poids démographique de la commune et le nombre d'adhérents de la commune-même³. Le résultat, après avoir enlevé les deux plus grands adhérents (700 tranches à Bernex et 50 à Avully), est très significatif : 0.989 (corrélation quasi parfaite), qui nous indique une répartition uniforme de la clientèle solaire dans l'espace et par conséquent une absence de pouvoir explicatif de ce critère en matière de clientèle solaire⁴.

2.1.1 Données des clients solaires

Nous avons eu l'accès au fichier des clients qui ont répondu à l'appel à la souscription des SIG. Le fichier inclut les variables suivantes :

- numéro de rapport d'usage (RUS) : identifiant du client
- journées techniques : donnée interne pour le relevé
- genre
- adresse
- quantité prévue et souscrite : kWh solaires en centaines (tranches)
- date de la souscription
- type de tarif⁵

Les autres variables présentes dans le fichier « clients solaires.xls » sont :

- date d'envoi du coupon de la part du client,
- date de la signature du contrat,
- la période du contrat
- autres données administratives (date de retour du coupon, remarques, etc.).

Cette information étant néanmoins insuffisante pour nos besoins d'analyse et modélisation, nous avons demandé aux SIG les données de consommation des clients de la bourse. Ces données font partie de la base de données de la clientèle⁶. Nous avons opéré un premier traitement en utilisant une nouvelle interface ayant l'avantage de calculer la consommation du

³ Pas la quantité souscrite.

⁴ La motivation de cette analyse étant de chercher si – par exemple – les communes « riches » souscrivent plus que les « pauvres ». Nous envisageons ensuite d'attribuer une variable proxy de la richesse de la commune en codant la commune de résidence par la valeur des centimes additionnels de la commune.

⁵ Annexe 2.

⁶ Il s'agit d'une base *Oracle*[®].

client sur l'arc temporel d'intérêt. La base fournit également la liste de tous les autres compteurs attribués au client⁷.

L'information sur la consommation à un intérêt évident pour l'estimation du niveau de vie, ainsi que pour une caractérisation générale du client, en nous disant, par exemple, s'il a une activité commerciale, un logement secondaire, le code NOGA, le type de construction et - s'il y en a - les compteurs pour gaz et eau.

En attendant l'extraction des données de facturation, nous avons commencé manuellement le traitement des données pour les clients de la bourse de 2001. Cette démarche « lente » a permis des premières estimations pour environ 80 individus⁸. Nous nous sommes toutefois vite rendus compte qu'il était impossible de continuer avec la procédure manuelle de transcription de la consommation depuis *Data Warehouse*[®] vers le fichier clients solaire en *Excel*[®] pour tous les clients de la bourse.

2.1.2 Données de facturation (clients SIG)

Suite aux démarches formelles entreprises auprès de la direction des SIG afin que le service informatique nous donne l'information sur la consommation, nous avons réalisé un appariement des deux bases de données. Le premier résultat a été l'obtention de 3 fichiers *Excel*[®] avec les variables suivantes :

Fichier 1 « consommation.xls »

| No rapport usage | Identifiant de l'individu |
|-----------------------------------|--|
| No install. (Point de Fourniture) | Info technique (souvent égale au RUS) |
| Code service | E, F, X, G, O |
| Code fourniture | Codification numérique du code de service |
| Libellé fourniture | Descriptif |
| Type minuterie | Prime de puiss., fact.de puiss, heures pleines/creuses |
| Date début relevé | Début de la période de facturation |
| Date dernier relevé | Fin de la période de facturation |
| Quantité en kWh | Unité physique tels kWh ou m ³ d'eau |
| Montant hors taxes en CHF | Montant de la facture |

Fichier 2 « client_boursesolaireetcodenoga.xls »

| Code service | No rapport usage | Code étage | Description courte | NOGA | | | | |
|--------------|------------------|------------|-----------------------|--------|--------|-------|------------------------|-----------------------------|
| | | | Division | Groupe | Classe | Genre | % activité ménagère | % activité professionnel |
| | | | | | | | | |

Cette information nous donne une répartition des individus selon l'activité professionnelle et ménagère.

⁷ Il s'agit de *Data Warehouse*[®], une interface de consultation des données qui venait juste d'être mis en place.

⁸ Les résultats sont dans Annexe 1.

Fichier 3 « listeetagesetcodes.xls »

| Code étage | Etage |
|-------------|--|
| De 1 à 50 | Du 1 ^{er} au 50 ^{ème} étage |
| De 81 à 100 | Du 1 ^{er} au 20 ^{ème} sous-sol |
| 201 | Sous-sol |
| 202 | Rez-de-chaussée |
| 203 | Entre-sol |
| 204 | Mezzanine |
| 205 | Cave |
| 206 | Combles |
| 207 | Rez-inférieur |
| 208 | Rez-supérieur |
| 209 | Toit |
| 210 | Attique |
| 211 | Cour |
| 212 | Arcade |

Ces données révèlent le type d'habitation de l'individu.

2.2 La fusion des deux sources de données

Avec la nouvelle information acquise, nous avons procédé à l'agrégation de la consommation pour chaque client de la Bourse et pour les années 2000, 1999 et 1998.

Nous avons ainsi commencé la création d'une nouvelle base de données incluant notamment⁹:

- consommation de 2000, 1999, 1998 ;
- type de tarif ;
- commune de résidence du client ;
- genre.

La première tâche a impliqué la création d'un tableau croisé dynamique *EXCEL*[®] afin d'additionner les montants des kWh et CHF des clients pour la catégorie *E* (électricité) de la variable « Code de service » cela fournit une approximation de la consommation qui est un indicateur du niveau de vie de l'individu¹⁰. Pour estimer la dépense en électricité nous gardons l'option d'ajouter la catégorie *F* représentant les frais fixes dans la facture du client.

Nous avons ainsi agrégé les consommations (en CHF et kWh) des clients de la Bourse pour tous les rapports d'usage attribués. Ceci afin d'obtenir une estimation de la dépense du consommateur en énergie.

Cela nous a permis par la suite de vérifier l'existence d'une corrélation entre la facture et la quantité d'énergie solaire souscrite. Toutefois, le coefficient de corrélation est très faible (0.07) ce qui nous montre que ce n'est pas le cas (Ch. 4).

2.3 Les variables explicatives

2.3.1 Tarif

⁹ Cette démarche a dû se faire en grande partie manuellement à cause de la non-correspondance entre les RUS du fichier « clients_solaires.xls » et « consommation.xls ».

¹⁰ Il s'agit d'une approximation puisque la facture est basée sur une estimation de la consommation à partir des relevés des compteurs de l'année précédente.

Les différents tarifs des SIG sont présentés ci dessous¹¹ :

| Abréviation | Origine | Description | Exemples |
|----------------|-----------------|---|---|
| Ed (BT direct) | Direct | 400 V, sans TI (coupe-surintensité =<80A) | |
| Ed (U1 + U2) | Usagers | Usagers ménagers (+ activités professionnelle5s) | Appartements, petites villas, professions indépendantes, agriculteurs.. |
| (P1 + P2) (*1) | Professionnels | Utilisation générale (à but lucratif) | Professions indépendantes, bureaux, commerces, parkings, chaufferies, agriculteurs, ... |
| (TE) | Tout électrique | Chauffages électriques | Appartements, petites villas, professions indépendantes,.. |
| (forfait) | Forfait | Tarifs à forfait pour mobilier urbain | Appareils téléphoniques publics, distributeurs TPG,... |
| Ei (BT direct) | Indirect | 400 V, avec TI (coupe-surintensité =<1500A) | |
| Ei (B) | | | Magasins, banques, hôtels, agriculteurs, ... |

Tableau 2.3 – Types de tarif des clients solaires

Un premier critère consiste à séparer les tarifs des ménages des commerciales, les dernières caractérisant – en général – des plus grandes consommations. Pour cela nous avons codé les tarifs des clients par des entiers ; il y a essentiellement 6 types de tarifs dans notre sous-échantillon:

| Tarif | Code | Effectifs |
|--------|------|-----------|
| Ed(P1) | 1 | 18 |
| Ed(P2) | 2 | 3 |
| TE | 3 | 17 |
| Ed(U1) | 4 | 560 |
| Ed(U2) | 5 | 51 |
| Ei(B) | 6 | 4 |

Tableau 2.4 – Codage des tarifs et effectifs

Chaque tarif peut ensuite être reporté soit séparément comme une dummy (0, 1) soit sur une seule colonne pour des estimations selon différents critères. Dans nos premières estimations, par exemples, nous avons agrégé les tarifs de différentes manières, en discriminant – par exemple – le tarif privé « usagers » Ed(U1) de tous les autres types (voir Ch. 4).

Nous songeons à enlever les quatre observations du tarif Ei(B) car leurs consommations respectives sont trop grandes par rapport au reste de la base ; cela s’explique : il s’agit de collectivités publiques ou entreprises, donc ni le choix d’adhérer à la bourse peut être considéré comme individuel, ni leurs consommations vont être corrélées au quantités solaires souscrites. Le tarif nous permis de vérifier la représentativité de l’échantillon de la clientèle solaire par rapport à la population du canton ; les deux populations sont réparties selon les tarifs ci dessous:

| Population | Echantillon |
|--------------|--------------------|
| 8.6% | Ed(P1) 2.8% |
| 1.4% | Ed(P2) 0.5% |
| 1.4% | TE 2.6% |
| 74.8% | Ed(U1) 85.8% |
| 6.6% | Ed(U2) 7.8% |
| 0.9% | Ei(B) 0.6% |
| 6.3% | autres 0% |

Tableau 2.5 – Comparaison entre population et échantillon selon les tarifs

¹¹ Pour le description complète des tarifs, voir en annexe.

La première remarque concerne le tarif Ed(U1) qui est sur-représenté dans la clientèle solaire (+10%) ; la part de tarif TE présente une grandeur presque double ; une autre différence considérable concerne la clientèle professionnelle, $Ed(P1)$ et $Ed(P2)$, qui est nettement sous-représenté chez les clients solaire. Cette dernière information confirme que les entreprises n'adhèrent pas ou peu à la bourse solaire, qui reste attrayante pour les particuliers. Cela peut s'expliquer par deux raisons :

- les entreprises ne trouvent pas « rentable » d'adhérer à la bourse, même en incluant l'effet d'image
- les entrepreneurs adhèrent en tant qu'individus et substituent l'énergie consommée à la maison, mais pas celle sur le lieu de travail.

2.3.2 Le genre

Nous avons reporté le sexe du client reporté dans le fichier car il est ainsi possible se vérifier s'il est significatif dans l'explication de quantité d'énergie solaire achetée.

Nous avons codé le genre avec la variable muette :

Monsieur = 0
Madame ou Madame et Monsieur = 1

Lorsque l'indication « Monsieur et Madame » apparaissait nous avons choisi de considérer l'impulsion féminine comme présente dans le choix d'adhérer et nous lui avons donc attribué un « 1 ».

2.3.3 La commune de résidence

Nous voulions tester si une discrimination significative pouvait être admise selon l'appartenance ou non à une commune. Dans l'ensemble des indicateurs de la richesse nous avons trouvé que la valeur des centimes additionnels payés à la commune peut refléter la richesse de la commune.

L'appartenance à la commune a été donc jugée d'après les centimes additionnels. Les données proviennent du fichier envoyé par l'Office Cantonal de la Statistique. Nous employons d'abord ces données pour discriminer les consommateurs selon l'appartenance à une commune riche ou pauvre. La distribution des centimes additionnels est résumée ci-dessous :

| | |
|---------|------------|
| 51.87 | Moyenne |
| 44.70 | Médiane |
| 22.3824 | Ecart type |

Tableau 2.6 - Centimes additionnels des communes genevoises

Selon la moyenne les communes « riches » sont¹² :

¹² C'est à dire à droite de la moyenne.

Anières, Celigny, Cartigny, Chêne-Bougeries, Choulex, Collonge-Bellerive, Corsier, Genève, Genthod, Grand Saconnex, Hermance, Meinier, Plan-les-Ouates, Pregny-Chambésy, Presinge, Satigny, Troinex, Vandoeuvres.

Tandis que les communes « pauvres » sont : Aire-la-Ville, Avully, Avusy, Bardonnex, Bellevue, Bernex, Carouge, Chancy, Chêne-Bourg, Collex-Bossy, Confignon, Dardagny, Gy, Jussy, Laconnex, Lancy, Meyrin, Onex, Perly-Certoux, Puplinge, Russin, Soral, Thônex, Vernier, Versoix, Veyrier.

Si l'on prend la médiane comme barème, il y a Carouge, Dardagny et Jussy qui basculent vers les « riches », mais cela est sans grande influence puisqu'il n'y a que deux souscriptions à la bourse pour ces communes.

Nous avons ensuite codé la variable "commune" de la base de données avec la valeur des centimes additionnels de la commune¹³.

2.4 Les données de facturation principaux : kWh/an et CHF/an

Nous ne pouvons plus, à ce stade, dissocier l'explication de la procédure d'analyse de la donnée disponible : les fichiers de facturation présentent les montants en kWh de l'électricité consommée ainsi que du gaz et eau. Le gaz lorsque le consommateur en est desservi, l'eau s'il y a un compteur, comme c'est le cas pour les villas et commerces). Un exemple :

| No rapport usage | Code service | Libellé fourniture | Type minuterie | Date début relevé | Date dernier relevé | Quantité | Montant HT en CHF |
|------------------|--------------|-------------------------------|----------------|-------------------|---------------------|----------|-------------------|
| 70703 | E | Puissance mens. Appelée en kW | PP | 15.12.1999 | 13.12.2000 | 633 | 7841.60 |
| 70703 | E | Facteur de puissance | FP | 15.12.1999 | 13.12.2000 | 4848 | -1738.40 |
| 70703 | E | Cons. heures pleines en kWh | HP | 15.12.1999 | 13.12.2000 | 130431 | 23101.50 |
| 70703 | E | Cons. heures creuses en kWh | HC | 15.12.1999 | 13.12.2000 | 44292 | 4152.80 |
| 70703 | G | Prime de puissance | | 15.12.1999 | 13.12.2000 | 12 | 48.00 |
| 70703 | G | Consommation en kWh | | 18.01.2000 | 13.12.2000 | 874 | 87.40 |
| 134014 | F | Consommation estimée en fr. | | 04.02.2000 | 04.04.2000 | 78.75 | 78.75 |
| 134014 | X | ACOMPTE FACTURE A DEDUIRE | | 05.10.1999 | 08.05.2000 | 0 | -153.35 |
| 134014 | E | Consommation en kWh | HP | 05.10.1999 | 03.11.2000 | 1460 | 383.85 |
| 134014 | G | Prime de puissance | | 05.10.1999 | 03.11.2000 | 5 | 52.00 |
| 134014 | G | Consommation en kWh | | 05.10.1999 | 03.11.2000 | 776 | 61.75 |
| 136022 | F | Reprise de l'installation | | 17.09.1999 | 04.02.2000 | 0 | 23.50 |
| 136022 | F | Consommation estimée en fr. | | 04.02.2000 | 04.04.2000 | 79.35 | 79.35 |
| 136022 | X | ACOMPTE FACTURE A DEDUIRE | | 04.02.2000 | 08.05.2000 | 0 | -79.35 |
| 136022 | X | solde débiteur reporté | | 08.05.2000 | 06.07.2000 | 0 | 2.80 |
| 136022 | E | Consommation en kWh | HP | 17.09.1999 | 03.11.2000 | 1868 | 491.05 |
| 136022 | G | prime de puissance | | 17.09.1999 | 03.11.2000 | 5 | 56.00 |
| 136022 | G | Consommation en kWh | | 17.09.1999 | 03.11.2000 | 615 | 48.30 |

Tableau 2.7 – Exemple de données de facturation

Nous avons ici les données de facturation de trois clients SIG – identifié par leur numéro de rapport d'usage (à gauche) - qui ont adhéré à la bourse ; il se pose d'abord la question de quoi prendre comme information, car, en effet, les données présentes de l'information de type :

- Qualitatif : « il y a le gaz », « paie directement l'eau », « double facturation » etc. ;

¹³ Pour les résultats de ces estimations vois Ch. 6.

- Quantitatif : montant (en kWh et/ou CHF) pour le service *E* seulement, la montant total payé aux SIG ou toute autre combinaison de *montants*.

Il est du de remarquer que l'objectif dans le choix des variables et leurs transformations est d'avoir un indicateur du niveau de vie de l'individu, ou de cerner une autre caractéristique explicative de la quantité d'énergie solaire achetée.

Nous pouvons définir la variable « *consommation 2000* » comme fonction des données « *quantité* » ; ce qui nous laisse le choix d'inclure ou non :

- *G* (gaz) avec : prime de puissance, prime de puissance en kWh,
- *F* : frais financiers, tels : consommation estimée, reprises d'installation,
- *X* : acomptes et soldes comptables *ex-post*
- *O* : consommation d'eau en m³, prime mensuelle de débit et taxe pour les égouts.

Et cela en plus des valeurs prises par le code de service *E* qui concerne l'électricité. Les occurrences de cette variable dépendent du tarif. Parmi les tarifs nous distinguons :

- *Ed(U1)* : consommation en kWh,
- *Ed(U2)* : consommation en kWh, avec double facturation (heures pleines, heures creuses),
- *Ed(P1) Ed(P2)* : même chose que *Ed(U)*, mais pour une activité professionnelle,
- *Ei(B)* : consommation en kWh à 400Volts, avec double facturation (heures pleines, heures creuses) et prime de puissance et la puissance mensuelle appelée en kW.

Dans l'optique quantitative une alternative simple est de considérer les données de facturation comme indicateur du revenu l'individu. On additionne au choix seulement *E* ou *G*, *X*, *F*, *O*, aussi s'ils sont facturés.

Il faut noter que dans le canton de Genève 75 % des clients SIG ont le tarif *Ed(U1)*, ce chiffre monte à 85% parmi les clients solaires ; or, puisque pour ce tarif il n'y a pas de double facturation, nous avons décidé de prendre en compte seulement l'occurrence « *E* », qui indique les kWh consommés et le montant hors taxes en CHF payé.

2.4.1 Segmentation

Un critère logique consisterait à partager les individus selon leur tarif puisque cela distingue assez bien les particuliers, des entreprises collectivités publiques etc. Toutefois le très faible nombre d'entreprises et collectivité rend inutile au départ une telle segmentation de l'échantillon (cf. 2.1). Un autre critère consiste à poser une variable indicatrice dans le cas de la double facturation heures pleines/creuses ; pour cela nous écrivons la variable binaire :

$$D_i(\text{tarif}) = 0 \text{ si Tarif} = Ed(U1) \text{ ou } Ed(P1)$$

$$D_i(\text{tarif}) = 1 \text{ si Tarif} = Ed(U2) \text{ ou } Ed(P2)$$

Nous avons aussi posé la variable binaire *tarifbin* :

$$E_i(\text{tarif}) = 0 \text{ si Tarif} = Ed(U1)$$

$$E_i(\text{tarif}) = 1 \text{ si Tarif} = Ed(U2) \text{ ou } Ed(P1) \text{ ou } Ed(P2) \text{ etc...}$$

Cette dernière variable résulte significative pour différentes estimations de modèles logit dichotomique où la variable dépendante est :

Tranchebin = 0 si l'individu souscrit une tranche

Tranchebin = 1 si l'individu souscrit plus qu'une tranche.

Pour les détails de ces estimations, nous renvoyons à l'annexe 3.

Le tarif *TE* est souvent indicateur d'une villa avec pompe à chaleur (PAC) pour les besoins d'eau chaude sanitaire et chauffage ; s'agissant d'une technologie moderne et efficace, cette tranche de clientèle est susceptible d'avoir un sentiment écologique plus marqué que le reste de la population. Nous avons en conséquence créé une variable binaire pour distinguer les individus appartenant à ce segment.

2.4.2 Des kWh/an à la variation d'une année à l'autre

Une fois obtenues les données de facturation pour l'ensemble des clients solaires, nous songeons à vérifier si non pas le *niveau* mais la *variation* de la consommation explique mieux le choix de combien souscrire. Nous croyons que l'augmentation par rapport aux années précédentes peut expliquer la décision d'adhérer à la Bourse et/ou de souscrire plus qu'une tranche.

Avec cette première base « crée » nous avons essayé de voir si, parmi les gens qui ont adhéré, il y a des facteurs qui influencent la *quantité* d'électricité PV souscrite. L'idée sous-jacente étant que les gens qui :

- ont une grande consommation ;
- ont le double tarif ou le « tout électrique », (*TE*) ;
- habitent des communes « riches ».

sont susceptibles de souscrire plus qu'une seule tranche, ceci car leur revenu – indirectement indiqué par les variables ci-dessus – est supérieur à la moyenne des individus. Les variables rentrent donc dans notre modèle explicatif du choix de *combien* de tranches souscrire.

2.5 Corrélation empirique entre consommation et quantité souscrite

Après avoir enlevé les deux adhérents pour 700 et 50 tranches - évidemment aberrants pour nos fins - nous avons calculé le coefficient de corrélation empirique entre la consommation et le nombre de tranches d'énergie solaire. Le résultat se révèle toute suite très peu encourageant : le coefficient de corrélation est très faible (0.07). Ce résultat nous montre que l'hypothèse d'une corrélation directe consommation-énergie solaire est fautive ; les clients qui souscrivent plus de tranches ne sont pas les plus gros consommateurs.

Une deuxième hypothèse à vérifier concerne le coefficient de corrélation empirique entre la quantité consommée et l'énergie solaire souscrite pour les clients qui ont souscrit *plus* qu'une tranche. Toutefois, dans ce sous-échantillon, le coefficient de corrélation est encore plus faible (0.05).

Pour le sous-échantillon des clients qui ont souscrit trois tranches ou plus le coefficient de corrélation est négatif (-0.1), ce qui pourrait indiquer que ce sont les gens qui consomment le moins d'énergie qui souscrivent le plus d'énergie solaire. Ce résultat est intéressant, car il peut indiquer l'existence d'une catégorie de gens sensibles aux problèmes environnementaux et qui agissent de manière cohérente en souscrivant une quantité d'énergie solaire *en fonction* de la facture d'électricité et pas seulement pour participer à la bourse.

Nous devons toutefois souligner comme le nombre d'adhérents pour plus que trois tranches (80 individus) rend toute conclusion très peu défendable. Nous avons remarqué la valeur négative du coefficient de corrélation pour le sous échantillon de trois tranches ou plus parce que si la grande majorité des adhérents ont souscrit une seule tranche, toute consommation confondue, nous pourrions être emmenés à conclure que la majorité des clients solaires font simplement un geste en faveur de l'environnement, et cela indépendamment de la consommation et du prix de l'énergie solaire. Il s'agit du cas de *lump sum* constituant, d'une certaine manière, le *a priori* dans notre étude.

Un tableau résumant les résultats des coefficients de corrélation :

| | |
|-----------------------------|------------|
| Corrélation totale | 0.07998545 |
| Corrélation 2 tranches ou + | 0.04767103 |
| Corrélation 3 tranches ou + | -0.1039596 |

Tableau 2.8 – Corrélation empirique consommation-tranches

Une dernière possibilité à vérifier est la possibilité d'une logique dans la souscription : l'obtention d'un chiffre rond en additionnant la facture annuelle et le coût de l'énergie solaire. Pour cela nous avons posé la formule :

$$Y = 100 * \text{nombre de tranches} + \text{montant en CHF}$$

Nous avons ensuite calculé la moyenne pour chacune des quantités souscrites, car une éventuelle régularité pourrait être indicatrice d'une logique dans le choix. ; une fois encore ce n'est pas le cas : la moyenne baisse après trois tranches remonte à cinq, rebaisse à six pour remonter nettement.

3 Théories de choix individuel

Nous présentons ici un bref rappel de la théorie du choix individuel, afin de montrer les liaisons avec la modélisation des données discrètes ; nous croyons que ce sera utile pour la compréhension des modèles utilisés dans l'analyse de la bourse solaire.

Les économistes sont en général intéressés aux variables résultantes du comportement d'un grand nombre d'individus, tels la demande pour un bien ou un service. Les *quantités demandées* peuvent être vues comme le résultat de *décisions* prises, dans beaucoup de cas, au niveau individuel. Dans cette perspective, une importante branche de l'économétrie récente s'intéresse à la compréhension, à la modélisation, ainsi qu'à l'éventuelle vérification empirique du comportement individuel, dans des domaines tels le marketing, les assurances, le choix d'école, de la branche d'activité etc.

Si la variable à prévoir peut être envisagée comme le résultat d'un choix individuel, elle sera quasi-forcément discrète. Finalement, dans une optique plus large, la disponibilité et le traitement de données individuelles (sondages) pour chercher à expliquer un choix constitue le « pain » de la micro-économétrie, une discipline relativement nouvelle, qui se penche spécifiquement sur l'individu car, plus ou moins explicitement, retenu au cœur de tout modèle de prévision¹.

3.1 Choix et comportement

Nous considérons un *choix* comme le résultat d'un processus décisionnel qui se décompose selon les étapes suivantes² :

- Définition du problème du choix
- Génération des alternatives
- Évaluation des attributs des alternatives
- Choix
- Implémentation.

A partir de cette représentation, une théorie opérationnelle doit être apte à fournir des procédures pour définir (et codifier) en une formulation cohérente et – pour des fins économétriques – quantifiable :

- Décideur
- Alternatives
- Attributs des alternatives
- Règle de décision

Une remarque sur le comportement : le comportement *observé* n'est pas forcément le résultat d'un processus de décision, car il peut être dû aux habitudes, au conformisme ou l'intuition et, si théoriquement cela pouvait se représenter comme un processus où une seule alternative est générée, la formalisation de tels processus de choix nécessiterait des données très détaillées.

¹ Pour la disponibilité de données du comportement individuel, les données de sondage permettent une modélisation au niveau de chaque consommateur, qui est ainsi caractérisé par ses caractéristiques socio-économiques.

² Pour une description plus exhaustive et des exemples voir Ben Akiva M. and Lerman S. 1985. *Discrete Choice Analysis - Theory and Application to Travel Demand*. MIT Press.

3.2 Modèles et individu

Le décideur peut être un individu, un ménage ou un groupe. Les individus ont généralement une possibilité de choix et des goûts différents, ces derniers pouvant être intégrées dans le modèle pour une éventuelle segmentation des individus.

Les alternatives sont l'ensemble (non vide) des choix possibles de l'individu, elles dépendent de l'environnement de l'individu. Il est possible de distinguer en entre un *ensemble universel* et le sous-ensemble appelé *ensemble du choix*. Le dernier incluant les alternatives accessibles (réalistes) et observables par le modélisateur.

Les attributs des alternatives définissent l'attractivité de chaque choix pour l'individu. Ils sont exprimés dans un vecteur spécifique à l'alternative. L'échelle peut être ordinale ou cardinale³.

La règle de décision est fondamentale. La modélisation en économétrie ne considère qu'une décision basée sur l'utilité ; nous reportons ci-dessous les cinq approches principales de la théorie de la décision.

Domination : une alternative domine les autres si pour au moins un attribut elle est meilleure sans être pire pour aucun autre. Ce critère conduit rarement à un choix unique. Il peut être utilisé dans la phase préliminaire de la modélisation pour éliminer les alternatives inférieures.

Satisfaction : pour chaque attribut il existe un niveau minimal et satisfaisant, selon les attentes du décideur, pour qu'il prenne l'alternative en considération. Ce critère a le même défaut du critère de domination.

Règles lexicographiques : les attributs sont rangés selon une échelle d'« importance », le décideur choisit l'alternative qui est gagnante selon l'attribut le plus important. La combinaison des critères de satisfaction et de règles lexicographiques s'appelle « élimination par les aspects »⁴. Le processus commence avec l'attribut le plus important pour éliminer ensuite les alternatives qui ne satisfont pas au critère de niveau satisfaisant. Si plusieurs alternatives restent, le processus continue avec le deuxième attribut le plus important etc.

Utilité : c'est l'approche adoptée en économétrie. Elle implique que les attributs d'une alternative (ainsi que les caractéristiques du décideur) peuvent devenir *information* et être réduits à un scalaire, l'alternative est codée par un vecteur. Cela permet de définir une seule fonction objectif, l'utilité, qui peut être vue comme un index d'attractivité de l'alternative pour chaque individu.

Historiquement deux approches à la théorie du choix se sont confrontés. D'un côté, en psychologie, le choix a été vu comme le résultat d'un processus probabiliste, à cause des observations expérimentales inconsistantes et non-transitives à expliquer auxquelles les chercheurs font face.⁵ Parmi les économistes, par contre, l'approche néo-classique de la théorie du consommateur s'est pendant longtemps fondé sur l'hypothèse de rationalité de l'individu. L'individu est ainsi parfaitement capable de ranger les alternatives à disposition,

³ Pour un rappel sur les échelles de mesure nous renvoyons aux Annexes.

⁴ Tversky, A. 1972. *Elimination by Aspects : A Theory of Choice*. Psychological Review 79, pag. 281-299.

⁵ Inconsistance : les mêmes individus ne répètent pas le même choix face aux mêmes alternatives. Transitivité : face à des variations des attributs des alternatives, le choix de l'individu viole les hypothèses de rationalité.

en se basant sur une fonction d'utilité, pour effectuer son choix. Cette hypothèse, défendable quand le choix est entre deux ou trois alternatives, revient à dire que chaque individu peut exprimer ses préférences d'une manière univoque.

Entre des deux approches il est possible d'identifier une dichotomie fondamentale entre l'hypothèse d'un comportement probabiliste et celle qui revient à considérer comme probabiliste le manque d'information sur les facteurs déterminant le choix⁶. Cette dichotomie se reflète dans deux approches :

- la règle de décision est stochastique et l'utilité est constante⁷ ;
- la règle de décision est déterministe et l'utilité aléatoire⁸ ;

A ce propos De Palma et Thisse remarquent que : « *dans un cas comme dans l'autre, les variables de référence sont les probabilités et que celles-ci peuvent être données a priori sans passer par la spécification d'un mécanisme particulier. Toutefois, en tant qu'économistes, on souhaiterait savoir dans quel cas ces probabilités peuvent être construites à partir de la maximisation d'une utilité stochastique* »⁹.

Pour une revue des modèles à utilité constante nous renvoyons à l'article de De Palma et Thisse (1989) ; nous traiterons par la suite des modèles à utilité aléatoire.

3.3 L'approche au choix par l'utilité aléatoire

Le travail de Thurstone¹⁰ est à l'origine la théorie de l'utilité stochastique. Thurstone proposait un modèle explicatif de la variabilité des réponses d'un individu à des stimuli identiques. Il expliquait que « *la variabilité est due aux fluctuations qualitatives entre les occurrences [...] pour un stimulus donné* », ce qui le portait à supposer que « *un stimulus donné provoque une « sensation » ou « état psychologique » qui est la réalisation d'une variable aléatoire* »¹¹. Avec cette hypothèse, la réponse – le choix pour un individu – résulte de la comparaison entre les variables aléatoires engendrés par chaque stimuli.

En économie, nous sommes intéressés par les cas où « stimulus » est une alternative et la réponse est un choix. Dans cette étude plus particulièrement nous essayons d'expliquer un choix qui « révèle » la demande d'énergie solaire des individus. Cette demande devant s'exprimer par *quanta* de 100kWh/an.

Le choix est supposé être rationnel, mais fonction de paramètres inobservables. Face à une alternative simple, l'individu maximise une fonction d'utilité inconnue qui s'écrit :

$$U_0 = \beta_0'X + \varepsilon_0 \quad \text{pour le choix « 0 »}$$

et

$$U_1 = \beta_1'X + \varepsilon_1 \quad \text{pour le choix « 1 »}$$

⁶ Manski, C.F. (1977), "The Structure of Random Utility Models", *Theory and Decision*, **8**, 229-410.

⁷ Luce, R.D. (1959), *Individual Choice Behaviour. A Theoretical Analysis*, New York, Wiley.

⁸ Thurstone, L.L. (1945), "The Prediction of Choice", *Psychometrica*, **10**, 237-253.

⁹ De Palma, A.. et Thisse J.F., (1989), « Les Modèles de choix Discret », *Annales d'Economie et de Statistique*, **14**.

¹⁰ Thurstone, L.L. (1927), « Psychological Analysis », *American Journal of Psychology*, **38**, 368-389.

¹¹ *Ibid.* cité dans De Palma, A.. et Thisse J.F., (1989).

C'est un modèle dichotomique : l'individu choisit toujours le $\max\{U_1, U_0\}$. Prenons le cas où :

$$U_1 > U_0$$

Ce qui revient à poser :

$$\varepsilon_1 - \varepsilon_0 < \beta_0'X - \beta_1'X$$

en écrivant :

$$\varepsilon = \varepsilon_1 - \varepsilon_0$$

et,

$$\beta'X = \beta_0'X - \beta_1'X$$

Les modèles de choix binaire, ou dichotomique, modélisent la probabilité :

$$Pr\{\varepsilon \leq \beta'X\} \quad (1).$$

3.4 Détermination des probabilités de choix

Soit un nombre n de choix possibles, ε_i des variables aléatoire définies sur $]-\infty, +\infty[$ et $f(x_1, \dots, x_n)$ le densité jointe de $\varepsilon_1, \dots, \varepsilon_n$. En généralisant (1), la probabilité de choisir i peut s'écrire :

$$Pr(i) = Pr\{\varepsilon_1 - \varepsilon_i < U_i - U_1, \dots, \varepsilon_n - \varepsilon_i < U_i - U_n\}, \text{ pour } i = 1, \dots, n \quad (2)$$

La densité jointe des $\varepsilon_j - \varepsilon_i$ s'obtient en posant $y_j = x_j - x_i$, pour $j = 1, \dots, n$.¹² La cumulative de $f(x_1, \dots, x_n)$ $g((x_1, \dots, x_n))$ s'écrit :

$$g(y_1, \dots, [y_i], \dots, y_n) \equiv \int_{-\infty}^{\infty} f(y_1 + x_i, \dots, x_i, \dots, y_n + x_i) dx_i, \text{ pour } i = 1, \dots, n \quad (3)$$

la probabilité du choix i est :

$$Pr(i) = \int_{-\infty}^{u_i - u_1} \left[\int_{-\infty}^{u_i - u_i} \dots \int_{-\infty}^{u_i - u_n} g_i(y_1 \dots [y_i] \dots y_n) dy_1 \dots [dy_i] \dots dy_n \right] \quad (4)$$

Le changement de variable :

$$\begin{aligned} x_i &= x \\ z_j &= y_j + x \end{aligned}$$

la probabilité devient :

¹² C'est la procédure utilisée par De Palma, A. et Thisse J.F., (1989).

$$\Pr(i) = \int_{-\infty}^{\infty} F_i(u_i + x - u_1, u_i + x - u_2, \dots, x, \dots, u_i + x - u_n) dx \quad (5)$$

Pour F_i la dérivée partielle par rapport à x_i de la fonction de répartition des variables $\varepsilon_1, \dots, \varepsilon_n$. L'expression ci-dessus s'interprète comme suit : pour chaque valeur de x , l'intégrand est la densité de probabilité que ε_i prenne la valeur désirée x et que $\varepsilon_j \leq u_i + x - u_j$ pour tout $j \neq i$. En supposant les ε_i i.i.d. l'équation dévient :

$$\Pr(i) = \int_{-\infty}^{\infty} f(x) \prod_{j \neq i} F(u_i + x - u_j) dx \quad (6)$$

L'expression (6) définit la vraisemblance d'un modèle de choix discret générale pour n alternatives.

3.4.1 Spécification des probabilités pour le logit

En ayant,

n : individus (décideurs),

J_n : ensemble des alternatives,

$U_{in} = V_{in} + e_{in}$: utilité du $n^{\text{ème}}$ individu pour l'alternative i , composée de :

V_{in} : composante systématique de l'utilité (connue)

e_{in} : composante aléatoire, où :

V_{in} peut être détaillé entre :

Z_{in} : caractéristiques de l'alternative

S_n : caractéristiques de l'individu.

En supposant qu'il ait l'alternative :

1. Four à gaz
2. Four électrique

chacune étant caractérisée par le coût d'achat (p) et le coût de fonctionnement (c)¹³.

L'utilité de chaque alternative s'écrirait :

$$\begin{aligned} U_{\text{gaz}} &= \beta_1 p_{\text{gaz}} + \beta_2 c_{\text{gaz}} + e_{\text{gaz}} \\ U_{\text{ele}} &= \beta_1 p_{\text{ele}} + \beta_2 c_{\text{ele}} + e_{\text{ele}} \end{aligned}$$

Et la probabilité de choisir le four à gaz s'écrit :

$$\Pr_{\text{gaz}} = \frac{\exp\{\beta_1 p_{\text{gaz}} + \beta_2 c_{\text{gaz}}\}}{\exp\{\beta_1 p_{\text{gaz}} + \beta_2 c_{\text{gaz}}\} + \exp\{\beta_1 p_{\text{ele}} + \beta_2 c_{\text{ele}}\}} \quad (7)$$

¹³ Dans un tel cas il n'y a pas des caractéristiques du décideur, tel le revenu, par ex.

Si, par exemple, le choix incluait aussi l'option four à micro-ondes, nous serions dans un cas multinomial (Ch. 5.3), et l'utilité s'écrit :

$$\Pr_{\text{gaz}} = \frac{\exp\{\beta_1 p_{\text{gaz}} + \beta_2 c_{\text{gaz}}\}}{\exp\{\beta_1 p_{\text{gaz}} + \beta_2 c_{\text{gaz}}\} + \exp\{\beta_1 p_{\text{ele}} + \beta_2 c_{\text{ele}}\} + \exp\{\beta_1 p_{\text{micro}} + \beta_2 c_{\text{micro}}\}} \quad (8)$$

3.4.2 Détails sur la spécification du logit dichotomique

L'utilité de l'individu n pour l'option j s'écrit :

$$U_{in} = V_{in} + e_{in} \quad (9)$$

Le modèle logit découle de l'hypothèse que les e_{in} soient identiquement et indépendamment distribuées selon une distribution de Weibull (ou valeur extrême de type 1), *i.e.* :

La cumulative de chaque e_{in} c'est : $\exp\{-\exp\{-e_{in}\}\}$

La densité de chaque e_{in} c'est : $\exp\{-e_{in}\}\exp\{-\exp\{-e_{in}\}\}$.

Donc, la probabilité que l'alternative i soit choisie par l'individu n , P_{in} :

$$\begin{aligned} P_{in} &= \Pr\{V_{in} + e_{in} > V_{jn} + e_{jn}, \forall j \in J_n, j \neq i\} \\ &= \Pr\{e_{jn} < e_{in} + V_{in} - V_{jn}, \forall j \in J_n, j \neq i\} \end{aligned} \quad (10)$$

Pour la résolution de la formulation (2), Train¹⁴ recourt à un artifice pour partager la tache. Il fait l'hypothèse que $e_{in} = s$, ainsi la probabilité que $U_i > U_j$ s'écrit :

$$P_{in} = \Pr\{e_{jn} < s + V_{in} - V_{jn}, \forall j \in J_n, j \neq i\} \Pr\{e_{in} = s\} \quad (11)$$

L'événement recherché est le produit des deux événements ou probabilité jointe. Ces conditions peuvent s'écrire (règle de Bayes) comme la densité de e_{in} évaluée à s , multipliée par la distribution cumulative de chaque e_{jn} (sauf e_{in}) évaluée à $s + V_{in} - V_{jn}$:

$$P_{in} = e^{-s} e^{-e^{-s}} \prod_{\substack{j \in J_n \\ i \neq j}} \left\{ e^{-e^{-(s+V_{in}-V_{jn})}} \right\} \quad (12)$$

En ajoutant $V_{in} - V_{in} = 0$, le produit n'exclue plus la composante i ,

$$P_{in} = e^{-s} \prod e^{\left\{ -e^{-(s+V_{in}-V_{jn})} \right\}} \quad (13)$$

en relaxant l'hypothèse que $e_{in} = s$, nous intégrerons sur toutes les possibles valeurs de s :

$$P_{in} = \int_{s=-\infty}^{\infty} e^{-s} \prod e^{-e^{-(s+V_{in}-V_{jn})}} ds = \int e^{-s} e^{-e^{-s \sum_j V_{in}-V_{jn}}} ds \quad (14)$$

¹⁴ Train, K. (1986), *Qualitative choice analysis*, Cambridge, MIT Press.

Le changement de variable est :

$$e^{-s} = t \quad \Rightarrow \quad -e^{-s} ds = dt$$

$$(-e^{-s}/e^{-s}) ds = dt/t \quad \Rightarrow \quad ds = - dt/t$$

Ainsi l'expression (6) devient :

$$P_{in} = \int_0^{\infty} t e^{-t \sum_j e^{-(V_{in}-V_{jn})}} \left(-\frac{dt}{t} \right) \quad (15)$$

$$P_{in} = \int_0^{\infty} \exp\left\{-t \sum_j e^{-(V_{in}-V_{jn})}\right\} dt \quad (16)$$

et finalement,

$$P_{in} = \frac{\exp\left\{-t \sum_j e^{-(V_{in}-V_{jn})}\right\}}{\sum_j e^{-(V_{in}-V_{jn})}} \Bigg|_0^{\infty}$$

nous obtenons la forme habituelle du modèle logit :

$$P_{in} = \frac{1}{\sum_j e^{-(V_{in}-V_{jn})}} = \frac{e^{-V_{in}}}{\sum_j e^{-V_{jn}}} \quad (17)$$

la somme étant sur j .

Ce modèle a plusieurs restrictions qui découlent de l'hypothèse *iid* pour tous les e_i . La première vient de l'indépendance entre les différentes alternative, qui donne l'*hypothèse des alternatives non-pertinentes* (IIA), caractérisant le logit ; la deuxième restriction importante concerne l'égalité entre l'espérance et la variance entre e_i et e_j . Ce qui entraîne que tout facteur non observé par le chercheur, qui affecte l'utilité de l'alternative i , n'affecte pas l'alternative j .

Une autre façon de voir ceci est que si les erreurs ont les mêmes espérance et variance, alors les facteurs inobservés de U_i doivent avoir la même variabilité facteurs inobservés de U_j . Cette hypothèse sur les erreurs est très peu défendable lorsqu'il y a raison de croire qu'il existe plusieurs facteurs inobservés pour lesquels la variabilité ne peut être la même que de manière fortuite.

Le modèle probit relaxe ces hypothèses, toutefois, nous le remarquons, la différence substantielle avec les modèles logit n'est guère de passer d'une distribution de Weibull à une normale, mais de l'hypothèse *iid* sur les ε_i à la distribution jointe de toutes les alternatives.

La formulation la plus générale est peut-être celle qui consiste à faire l'hypothèse que les β aient une composante aléatoire ($\beta_n = \beta_n^* + \psi_n$). Cette modélisation permet prendre en compte

les (éventuelles) différences de goût entre les individus ; nous obtenons ainsi le modèle probit multinomial avec effets aléatoires.

Dans la suite du document nous traiterons plus en détail le cas où les erreurs sont distribués selon des distributions normales et de Poisson.

[...] The Church would raise no objections if Galileo stated his theory as a mathematical hypothesis, invented and assumed in order to abbreviate and ease the calculations, provided he did no claim it to be a true description of the world.¹

4 Modèles

Nous avons vu dans le chapitre précédent que les variables économiques engendrées par un processus décisionnel individuel sont souvent discrètes. Nous pouvons voir la variable dépendante *nombre de tranches solaires* comme l'expression révélée et discrète de la demande d'énergie solaire continue de l'individu ; la discrétisation étant due à la contrainte d'acheter par quanta (*i.e.* par tranche de 100kWh/an).

L'hypothèse que nous faisons sur le comportement de l'individu est que la variable de prévision – la quantité d'énergie solaire – est la réalisation discrète de sa demande continue d'électricité verte. L'individu demande de l'électricité solaire pour substituer – en partie ou tout – la quantité annuelle d'électricité conventionnelle (kWh ou CHF). Puisqu'il peu vraisemblable que le client choisisse une quantité supérieure, nous pouvons écrire :

$$Q_{PV} = \min\{\text{consommation annuelle, quantité désirée}\}$$

la quantité désirée peut être fonction de différents facteurs, dont :

- orientation politique,
- revenu,
- éducation,
- sensibilité environnementale, militantisme écologique etc.

L'explication d'une demande caractérisée par une variable dépendante discrète est traité par différents modèles dans la théorie économétrique. Une première distinction concerne tout d'abord, les modèles ordonnés et non ordonnés.²

Dans ce chapitre nous allons présenter la spécification économétrique des modèles que nous allons estimer ; nous présentons d'abord les modèles généraux pour ensuite décrire les versions plus spécifiques et les adaptations des modèles d'intérêt aux données solaires.

4.1 Modèles de choix discret et de dénombrement

Une distinction principale pour les modèles concerne deux types de variables :

1. La variable dépendante est strictement qualitative.
2. La variable est le résultat d'un « comptage » d'occurrence d'un certain événement.

¹ Cardinal Bellarmino, cité dans Lehmann E. L. (1990), "Model Specification : The views of Fisher and Neymann, and Later Developments", *Statistical Science*, Vol. 5, No. 2, 160-168.

² Les modèles discrets prennent souvent en compte le cas où la variable prend la valeur 0, ce que signifierait ne pas adhérer à la bourse solaire, toutefois considérer cette occurrence reviendrait à prendre en compte tous les clients des Services industriels de Genève (environ 400.000 individus).

Dans le deuxième cas, les observations ont deux caractéristiques particulières :

- La variable a un ordre « naturel »
- La magnitude de la variable est significative

Si nous prenons des événements tels le choix du parti politique, le mode de transport ou le lieu du logement ces caractéristiques n'ont pas de signification et la variable est modélisée en termes des distributions des probabilités sur l'ensemble de choix.

Si, par contre, nous considérons un ordre de préférences sur un ensemble de choix ou un choix sur un ensemble ordonné, la variable dépendante se trouve « le long d'une échelle » de mesure (nombre de brevets déposés, de défaillances, de présences etc.).

4.2 Modèles de Poisson

La famille des modèles de Poisson, aussi appelés de comptage (ou dénombrement), regroupe les spécifications cherchant à décrire des phénomènes de type « ordinales », tels : défaillances d'un appareil, visites d'un site, arrosages d'un champ, révoltes etc. Dans tous ces cas la variable est quantitative mais discrète³.

Dans le cas des clients de la bourse solaire, nous essayons de décrire les tranches souscrites en fonction des caractéristiques de l'individu et indépendamment du temps. Cette modélisation se base sur l'hypothèse que tous les clients solaires font le choix du nombre de tranches à prendre dans un intervalle de temps donné. Nous considérons les clients qui ont souscrit en 2000 et qui confirment en 2001, ainsi que les adhérents dès 2001 comme ayant fait un choix « instantané ».

4.2.1 La loi de poisson

Lors de la description d'événements dénombrables une manière « simple » est de considérer que ces événements arrivent *au hasard*. En ayant $Y_{t, t+\delta}$ variable aléatoire qui indique le nombre d'événements dans l'intervalle de temps $(t, t + dt)$ et λ taux d'occurrence des événements, alors les événements sont issus d'un processus de Poisson si :

$$\Pr[Y(t, t + dt)=0]= 1 - \lambda dt + o(dt) \quad (18)$$

$$\Pr[Y(t, t + dt)=1]= \lambda dt + o(dt) \quad (19)$$

$$\Pr[y(t, t + dt)>2]= o(dt) \quad \text{quand } dt \rightarrow 0 \quad (20)$$

Il est à noter que⁴ :

1. La probabilité que plus qu'un événement se produise dans un laps de temps court est négligeable
2. Le taux d'occurrence λ est constant

³ C'est à dire que la différence entre l'occurrence i et $i+1$ a la même valeur pour tout i .

⁴ Dans: Barron, D.N. « The Analysis of Count Data : Overdispersion and Autocorrelation, Journal of the Royal Statistical Society, Series B (Methodological).

3. La probabilité d'un événement dans $(t, t + dt)$ n'est pas affectée par l'histoire précédente et ne dépend pas du temps passé depuis le dernier événement.

Lorsque les événements sont issus d'un processus de Poisson, la distribution des intervalles entre les événements ainsi que du nombre d'événements ont des propriétés particulières :

- la variable aléatoire X_i qui indique l'intervalle de temps entre l'événement $i - 1$ et i est distribuée selon une distribution exponentielle ;
- la durée moyenne entre événements est $\frac{1}{\lambda}$;
- la distribution de X_i est : $Pr(X_i \leq x_i) = F(x) = 1 - \exp\{-\lambda x\}$, pour $x \geq 0$
- la densité : $f(x) = \frac{dF(x)}{dx} = \lambda \exp\{-\lambda x\}$, pour $x \geq 0$

Si Y_t est le nombre d'événements se produisant dans un intervalle de longueur t , alors Y_t est distribué selon :

$$Pr(Y_t = y) = f(y) = e^{-\lambda t} \frac{(\lambda t)^y}{y!}, \quad y = 0, 1, 2, \dots \quad (21)$$

et de moyenne et variance λt .

Plusieurs types d'événements peuvent être représentés par des processus de Poisson et ses généralisations ; nous allons par la suite présenter les extensions principale. Dans le chapitre dédié aux estimations nous employons un tests basé sur la régression pour détecter la sur dispersion.

4.2.2 Le modèle de Poisson

Avec les hypothèses dans 4.2.1, les événements observés (Y_t) dans chaque période ($t = 1, 2, 3, \dots$) sont des variables aléatoires indépendantes distribués selon (4) de paramètre λ . En admettant que les intervalles sont tous de grandeur unitaire nous pouvons écrire la distribution jointe pour tous les événements :

$$L = \prod_{t=1}^T \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!} \quad (22)$$

C'est la vraisemblance pour des variables aléatoires de Poisson ; la log-vraisemblance (sans terme constant) c'est :

$$\log L = \sum_{t=1}^T y_t \log(\lambda_t) - \lambda_t - \log(y_t!) \quad (23)$$

Le variables explicatives $X_{ij}, j = 1, \dots, K$ sont souvent introduites dans le modèle en forme log-linaire pour garantir la non-négativité de λ^5 :

⁵ Une discussion détaillée sur la vraisemblance se trouve dans King, G. (1989), *Unifying Political Methodology : The Likelihood Theory of Statistical Inference*, Cambridge, Cambridge University Press. Un exemple de

$$\lambda_i = \exp(\mathbf{X}_i \boldsymbol{\beta}) = \exp\left(\sum_{k=1}^K x_{ik} \beta_k\right) \quad (24)$$

où \mathbf{X}_i est un vecteur de K éléments (K variables explicatives, incluant la constante) et $\boldsymbol{\beta}'$ est le vecteur des paramètres ($k \times 1$).

La log-vraisemblance s'écrit:

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= -\sum_{i=1}^N \lambda_i + \sum_{i=1}^N y_i \log \lambda_i - \sum_{i=1}^N \log(y_i!) \quad (25) \\ &= \text{cons} - \sum_{i=1}^N \exp(x_i \boldsymbol{\beta}) + \sum_{i=1}^N y_i x_i \boldsymbol{\beta} \end{aligned}$$

d'où les conditions de premier ordre ⁶ :

$$\frac{d\ell}{d\boldsymbol{\beta}} = -\sum_{i=1}^N x_i [\exp(x_i \boldsymbol{\beta}) - y_i] = 0 \quad (26)$$

ce qui est équivalent à :

$$\sum_{i=1}^N x_i [\lambda_i - y_i] = 0 \quad (27)$$

Ce modèle comporte plusieurs restrictions : il y a tout d'abord celle d'égalité entre espérance et variance [$\lambda_i = E(y_i) = V(y_i)$] et ensuite l'hypothèse d'indépendance des événements dans le temps. La littérature considère la possibilité d'une dépendance dynamique des occurrences dans le temps en modélisant le processus appelé dépendance des occurrences ou *occurrence dépendance*.⁷

Dans le contexte de la biométrie les « modèles de contamination » (ou *contagion models*) évaluent la possibilité que l'occurrence d'un événement, tels accidents ou maladies, modifie la probabilité d'occurrence de ces événements dans le futur.⁸

Dans notre contexte les occurrences sont supposées se produire en une fois, le moment où l'individu choisit le nombre de tranches.

Une autre violation de l'hypothèse d'indépendance dans le temps des événements prend en compte la notion de « vagues » (*spells*) ou intervalles, dans lesquels les événements se produisent davantage.⁹ Cette hypothèse est plus réaliste, car nous pourrions imaginer que les gens aient décidé d'adhérer (ou de varier le nombre de tranches) suite à la campagne initiale de Services Industriels ou à un événement majeur de sensibilisation à l'environnement (tel le

specification linéaire est dans : Cameron, C.A. et Trivedi, P.K. (1986), "Econometric Models Based on Count Data : Comparison and Applications of Some Estimators and Tests", *Journal of Applied Econometric*, **1**, 29-53.

⁶ Gourieroux *et al.* remarquent que si \mathbf{X}_i inclut un terme constant l'équation de vraisemblance est : $\sum(\lambda_i - y_i) = 0$.

⁷ Heckman J. and G. Borjas (1980), "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model for heterogeneity and state dependence", *Econometrica* **47**, 247-283.

⁸ Xekalaki, E. (1983), « The univariate generalised Waring distribution in relation to accident theory : proneness, spells or contagion ? », *Biometrics*, **39**, 887-895.

⁹ Voir: Xekalaki (1983) et Cresswell, W. et P. Froggatt (1963), *The Causation of Bus Drivers Accidents*, Cambridge University Press, London.

Festival *Agir 21*, par ex.). La vérification de cette hypothèse pour la campagne d'information SIG nécessite toutefois l'information de la date d'envois des brochures qui n'est pas disponible.

4.2.3 L'égalité entre moyenne et variance

La restriction principale concerne la relation (24), faute de laquelle les conséquences sont semblable à celles en cas d'hétéroskédasticité pour le modèle linéaire : les variances estimées sont inconsistantes et les éventuels tests non valables.¹⁰

Pour ce qui concerne l'égalité entre moyenne et variance de Y_i étant donné X_i , une manière d'inclure l'éventuelle sur-dispersion consiste à reformuler λ_i de manière à qu'il inclut un terme aléatoire :

$$\ln \lambda_i = X_i \beta + \varepsilon_i \quad (28)$$

Le terme d'erreur ε_i peut refléter différents erreurs de spécification, comme des variables explicatives omises.¹¹ Cette stochasticité du modèle doit être distinguée de celle du modèle de Poisson, qui est reliée au caractère aléatoire de la variable endogène Y_i .

En écrivant $g(\varepsilon_i)$ la densité de ε_i , la densité marginale de Y_i s'obtient en intégrant par rapport à ε_i :

$$\begin{aligned} \Pr[Y_i = y_i] &= \int \Pr[Y_i = y_i | X_i, \varepsilon_i] g(\varepsilon_i) d\varepsilon_i \\ &= \int \frac{e^{-\exp(X_i \beta + \varepsilon_i)} e^{(X_i \beta + \varepsilon_i)^{y_i}}}{y_i!} g(\varepsilon_i) d(\varepsilon_i) \end{aligned} \quad (29)$$

4.2.4 Binomiale négative

L'expression (4) définit une extension du modèle de Poisson dépendant de $g(\varepsilon_i)$. Quand cette dernière est une distribution gamma l'expression (29) a une forme fermée et nous obtenons le modèle de régression *binomiale négative*¹². En effet, ayant $g_{\lambda_i}(\varepsilon_i)$, ou de manière équivalente $f(\lambda_i)$, distribution gamma, l'intégration de (4) conduit au modèle binomiale négatif. En posant :

$$\lambda_i \sim \text{Gamma}(\phi_i, \nu_i)$$

avec la densité¹³ :

¹⁰ Cameron, A.C. et Trivedi P.K., (1990), "Regressions-Based Tests for Overdispersion in the Poisson Model", *Journal of Econometrics*, **46**, 347-364, page 347.

¹¹ Voir: Gourieroux, C., A. Monfort and A. Trognon (1984), "Pseudo maximum likelihood methods: applications to Poisson models", *Econometrica* **52**, 701-720. Dans Hausman, J., B.H. Hall and Z. Griliches (1984), "Econometric models for count data with an application to the patents-R & D relationship", *Econometrica*, **52**, 909-938, il y a le cas de stochasticité du paramètre pour des données de panel.

¹² Le modèle a été dérivé à l'origine par Greenwood, M. et Yule, G.U. (1920), "An inquiry into the nature of Frequency Distributions of Multiple Happenings, With Particular References to the Occurrence of Multiple Attacks or Repeated Accidents", *J. of the Royal Stat. Soc, ser.A*, **83**:255-79.

¹³ Nous reprenons la notation de Cameron, A.C. et P.K. Trivedi (1986), "Econometric models based on count data : comparisons and applications of some estimators and tests", *Journal of applied econometrics* Vol. **1**, 29-53, qui utilisent la paramétrisation "indice" de la fonction Gamma.

$$f(\lambda_i) = \frac{1}{\Gamma(\nu_i)} \left(\frac{\nu_i \lambda_i}{\phi_i} \right)^{\nu_i} \exp\left(-\frac{\nu_i \lambda_i}{\phi_i} \right) \frac{1}{\lambda_i} \quad (30)$$

Où ν est appelé *index* ou paramètre de précision. L'expression (4) devient :

$$\Pr[Y_i = y_i] = \int \Pr[Y_i = y_i | \lambda_i] f(\lambda_i) d\lambda_i \quad (31)$$

$$= \frac{\Gamma(y_i + \nu_i)}{\Gamma(y_i + 1)\Gamma(\nu_i)} \left(\frac{\nu_i}{\nu_i + \phi_i} \right)^{\nu_i} \left(\frac{\phi_i}{\nu_i + \phi_i} \right)^{y_i} \quad (32)$$

Avec :

$$E[Y_i] = \phi_i \quad (33)$$

et :

$$\text{Var}[Y_i] = \phi_i + \frac{1}{\nu_i} \phi_i^2 \quad (34)$$

Cette formule permet de mettre en évidence un taux de sur-dispersion par rapport au modèle de Poisson standard :

$$\text{Var}[y] / E[y_i] = \{1 + \alpha E[y_i]\} \quad (35)$$

En posant $\alpha = 1/\nu$.

4.2.5 Maximum de vraisemblance des modèles de Poisson

Les modèles de Poisson peuvent être estimés par la méthode du maximum de vraisemblance, la fonction de vraisemblance s'écrit :

$$L = \prod_i^N \left[e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \right] = \prod_i^N \exp[-\lambda_i + y_i x_i \beta - \log(y_i!)] \quad (36)$$

ce qui correspond à la log vraisemblance :

$$\log L = -\sum_{i=1}^N \lambda_i + \beta_1 \sum_{i=1}^N y_i + \sum_{j=2}^K \beta_j x_j \sum_{i=1}^N y_i x_{ij} - \sum_{i=1}^N \log(y_i!) \quad (37)$$

Pour l'estimation les moments empiriques $\mu_j = \sum_{i=1}^N x_{ij} y_i, j = 1, 2, \dots, K$, sont utilisés.

Cameron et Trivedi (1986) proposent un test de spécification pour choisir entre les modèles de Poisson et des modèles plus généraux, étant donné que l'espérance est correcte : $\mu_i \equiv E[y_i / X_i, \beta]$ La procédure exploite la relation entre moyenne et variance des modèles de Poisson incluant l'hétérogénéité, *i.e.* $E[y_i / X_i, \beta]$ et $\text{Var}[y_i / X_i, \beta]$ pour un l donné.

Ce qui revient à tester la significativité de $\alpha \neq 0$.

Nous avons opéré ce test dans le chapitre 5.

4.3 Modèles de choix discret

L'origine des modèles traitant de variables discrètes est ancienne : le premier modèle probit a été conçu pour la première fois en 1860 par Fechner en psychophysique. La théorie a ensuite été formalisée par Fisher (1937) et Bliss (1934, 1935, 1938), notamment pour ce qui concerne la vraisemblance et l'inférence, le dernier ayant donné le nom.¹

Les modèles de choix discret ont ensuite été développés par des biologistes pour le traitement de l'information des expériences de laboratoire qui comportait souvent l'établissement d'une relation entre dose et réponse, cette dernière étant caractérisée par des états (*e.g.* vivant, malade, mort)².

4.3.1 Modèles de choix discret binaires

Les modèles binaires ou dichotomiques sont employés lorsque la variable dépendante résulte d'un choix entre deux possibilités seulement et la variable peut prendre deux valeurs : 0 ou 1.

Nous avons vu (Ch. 4.1) que l'emploi d'un modèle qualitatif individuel revient à supposer que le choix de l'individu provienne d'une *règle de décision* ; la règle permet d'associer la réalisation de la variable dépendante (0 ou 1) aux variables explicatives.

Une telle approche pour notre problème d'expliquer les souscriptions à la bourse solaire consiste à imaginer le choix comme binaire : la personne ayant décidé d'adhérer à la bourse est-ce qu'il choisit une tranche ou plus³ ?

La formalisation des modèles binaires est présentée ci-dessous en supposant que la réalisation de $\{y_i = 1\}$ soit associée à des valeurs élevées des x_i . Une première conséquence de cette hypothèse est qu'il existe un *seuil*, dépendant de $x_i\beta$, au-delà duquel la proportion des $\{y_i = 1\}$ l'emporte sur $\{y_i = 0\}$.

Afin d'aboutir à une spécification économétrique (probabiliste), on ajoute un terme aléatoire u_i à l'explication déterministe. En notant le seuil c , la règle de décision s'écrit :

$$\begin{aligned} x_i\beta + u_i > c & \Leftrightarrow \text{proportion } \{y_i = 1\} \text{ élevée} \\ x_i\beta + u_i \leq c & \Leftrightarrow \text{proportion } \{y_i = 1\} \text{ faible} \Leftrightarrow \text{proportion } \{y_i = 0\} \text{ élevée}^4. \end{aligned}$$

En termes des probabilités :

$$\begin{aligned} \text{Prob}(y_i = 1) &= \text{Prob}(x_i\beta + u_i > c) = 1 - \text{Prob}(u_i \leq c - x_i\beta) \\ \text{Prob}(y_i = 0) &= \text{Prob}(x_i\beta + u_i \leq c) = \text{Prob}(u_i \leq c - x_i\beta). \end{aligned}$$

¹ Fechner, G.T. (1860) *Elemente der Psychophysik* (2 vols.). Breitkopf und Härtel, Leipzig. Fisher, R.A. (1935) "The case of zero survivors in Probit Analysis", *Ann. Appl. Biol.* Bliss, C.I. (1934), "The method of Probits", *Science* **79**, 38-39; Bliss, C.I. (1935), "The calculation of dosage-mortality curve" *Ann. Appl. Biol.* **22**, 134-167. "The comparison of dosage-mortality data" *Ibid.*; Bliss, C.I. (1938), "The determination of dosage-mortality curves from small numbers", *Quart. Journal of Pharmacy & Pharmacology* **11**, 192-216.

² Aitchison, J. et D. Silvey (1957), « The Generalisation of Probit Analysis to the case of Multiple Responses », *Biometrika*, **44**, 131-140.

³ Cette démarche nous laisse également la possibilité de « dichotomiser » autrement le choix : (1,2 ou 3,4,...N) ou (1,2,3 ou 4,5...N) etc.

⁴ Nous supposons que la valeur seuil est la même pour toutes les observations.

Il est cependant important de remarquer que « *le modèle ne détermine pas exactement la réalisation de l'événement $\{y_i = 1\}$ ou $\{y_i = 0\}$, mais fournit une mesure théorique de la proportion d'observations pour lesquelles cet événement s'est réalisé. Ensuite, puisque cette mesure théorique (la probabilité) est croissante dans son argument, la probabilité que $y_i = 1$ sera croissante en les composantes de x_i dont les paramètres associés sont positifs, et décroissante pour celles dont les paramètres sont négatifs* »⁵.

Nous présentons par la suite les différents modèles binaires.

4.3.2 Logit binaire

Le modèle logit binaire ou dichotomique suppose que le terme aléatoire u_i suive une distribution logistique ; les fonctions de répartition et densités sont respectivement :

$$\Lambda(x_i\beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \quad (38)$$

et

$$\lambda(x_i\beta) = \frac{e^{x_i\beta}}{(1 + e^{x_i\beta})^2} \quad (39)$$

L'avantage principal, comparé au modèles probit (voir 5.2.2), est que la fonction logistique est plus facile à calculer numériquement que celle de la distribution normale - notée Φ - ce qui implique que :

$$1 - \Lambda(x_i\beta) = \Lambda(-x_i\beta) = \frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}} = \frac{1}{1 + e^{x_i\beta}} \quad (40)$$

L'inversibilité directe de la densité est aussi de grande utilité puisque :

$$\log\left(\frac{p_i}{1 - p_i}\right) = x_i\beta \quad (41)$$

La vraisemblance du modèle s'écrit :

$$L(y, x, \beta) = \prod_{i=1}^N \left[\frac{1}{1 + e^{x_i\beta}} \right]^{1-y_i} \left[\frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \right]^{y_i} \quad (42)$$

et la log vraisemblance :

$$\log L(y, x, \beta) = \sum_{i=1}^N \left\{ (1 - y_i) \log \left[\left(1 + e^{x_i\beta} \right)^{-1} \right] + y_i (x_i\beta) - y_i \log \left(1 + e^{x_i\beta} \right) \right\} = \quad (43)$$

$$= - \sum_{i=1}^N \left\{ \log \left[1 + e^{x_i\beta} \right] - y_i x_i \beta \right\} \quad (44)$$

$$= - \sum_{i=1}^N \left\{ \log \left[1 + e^{x_i\beta} \right] - y_i x_i \beta \right\} \quad (45)$$

⁵ A. Thomas, *Econométrie des variables qualitatives*, Dunod 2000, page 52.

4.3.3 Probit dichotomique

L'hypothèse de base du modèle probit dichotomique est que les erreurs sont normalement distribuées. Les fonctions de répartition et densités d'une loi normale centrée et réduite sont respectivement :

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/2} \quad (46)$$

et

$$\phi(X) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad (47)$$

La probabilité s'exprime ainsi :

$$\begin{aligned} \text{Prob}(y_i = 1) &= \text{Prob}(u_i > -x_i\beta) \\ &= \text{Prob}(u_i < x_i\beta) \\ &= \text{Prob}(u_i/\sigma < x_i\beta/\sigma) \\ &= \Phi(x_i\beta/\sigma) \end{aligned} \quad (48)$$

Cette formulation empêche l'identification des paramètres structurels β et σ séparément ; pour cette raison une solution habituelle consiste à normaliser l'écart type à 1 (voir 4.7.2).

4.4 Généralisation de la vraisemblance pour les modèles dichotomique

En posant :

$$\begin{aligned} \wp\{y_i\} &= P_i^{y_i}(1-P_i)^{1-y_i} = P_i & \text{si } y_i = 1 \\ &= 1 - P_i & \text{si } y_i = 0. \end{aligned} \quad (49)$$

La probabilité conjointe de l'échantillon c'est :

$$l = \wp\{y_1, y_2, y_3, \dots, y_N = Y\} = \prod_{i=1}^N \wp\{y_i\} = \prod_{i=1}^N [P_i^{y_i}(1-P_i^{1-y_i})] \quad (50)$$

En posant $P_i = F(\beta'X_i)$, et en écrivant le logarithme de la fonction :

$$\log l = \sum_{i=1}^N [y_i \log F(\beta'X_i) + (1-y_i) \log [1-F(\beta'X_i)]] \quad (51)$$

les dérivées de premier ordre par rapport aux paramètres β donnent leur valeur pour lequel la log-vraisemblance est maximisée, *i.e.* les estimateurs du Maximum de Vraisemblance :

$$\frac{\partial \log l}{\partial \beta} = \sum_{i=1}^N \frac{y_i - F(\beta'X_i)}{F(\beta'X_i)[1-F(\beta'X_i)]} f(\beta'X_i) X_i \quad (52)$$

avec $f(\cdot)$ fonction de densité.

Il est du de remarquer qu'il s'agit d'une somme de fonctions de $\beta'X_i$ qui dépend directement de l'indice i . Posons :

$$\sum_{i=1}^N \frac{[y_i - F(\beta'X_i) + f(\beta'X_i)]}{F(\beta'X_i)[1 - F(\beta'X_i)]} = \mu_i \quad (53)$$

S'agissant de fonctions typiquement non linéaires il se peut que la solution, si elle existe, ne soit pas unique ; l'unicité dépend de la hessienne, qui doit être définie négative. Elle s'écrit :

$$E \frac{\partial^2 L}{\partial \beta \partial \beta'} = E \left[\frac{\partial \mu_i}{\partial (x' \beta)} x_i x_i' \right] = - \sum_{i=1}^n \frac{f^2(x' \beta)}{F(x' \beta)[1 - F(x' \beta)]} x_i x_i' = \sum_{i=1}^N h_i(x' \beta) x_i x_i' \quad (54)$$

le signe de la hessienne dépend donc indirectement de μ_i .

Cas particuliers :

Uniforme : $F_i = \beta'X$ $f_i = 1$; $f'_i = 0$

$$h_i = - \left[\frac{y_i}{F_i^2} + \frac{1 - y_i}{(1 - F_i)^2} \right] f_i^2 < 0$$

Logistique :

$$F_i = \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)} ; f_i = \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)} \left(1 - \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)}\right) ; f'_i = \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)} \left(1 - \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)}\right) \left(1 - 2 \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)}\right) ;$$

$$h_i = - \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)} \left(1 - \frac{1}{\left(1 + e^{-\frac{\pi}{\sqrt{3}} \beta'x}\right)}\right)$$

Normale :

$$F_i = \int_{-\infty}^{\beta'x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt ; f_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\beta'x)^2}{2}} ; f'_i = -\hat{\beta}' x_i \frac{1}{\sqrt{2\pi}} e^{-\frac{(\beta'x)^2}{2}} ;$$

Le calcul itératif des estimateurs se fait par la méthode du gradient ; l'idée de base des algorithmes étant d'orienter la « direction » de recherche de sorte à que elle soit perpendiculaire au gradient de la solution précédente.

4.5 Modèles séquentiels

Les modèles séquentiels sont basés sur l'hypothèse que les occurrences (dans notre cas le nombre de tranches) découlent d'une séquence de choix binaires⁶. Dans les modèles séquentiels l'individu qui choisit l'alternative k est supposé avoir exprimé la préférence pour $k - 1$ sur $k - 2$, $k - 2$ sur $k - 3$ etc.

Le problème consiste donc à spécifier $k - 1$ choix binaires pour chaque couple d'alternatives, en voyant la variable endogène comme issue d'un processus qui poursuit jusqu'au nombre final de tranches achetées. Son principal avantage étant la facilité d'estimation.

Toutefois, certains auteurs, mettent en garde du fait que le modèle est valable seulement si les composantes aléatoires, qui influencent les réponses à chaque étape, sont indépendantes.⁷ D'autres estiment que cette hypothèse est difficile à respecter, en vue de l'ordre même qui a aidé à structurer le modèle.⁸ Une description de modèle séquentiel appliquée au degré de scolarité est présenté dans l'ouvrage de Maddala.⁹

La formulation séquentielle est généralement utilisé pour incorporer le phénomène de la dépendance d'état - *state dependence* - caractérisant, par ex., chômage, maladies etc.¹⁰

D'autres possibilités pour modéliser des processus de choix séquentiel sont offertes par le modèle logit emboîté (*nested logit*), permettant de varier l'ensemble de choix à chaque étape.¹¹

Dans le cas de la bourse solaire ce phénomène pourrait seulement être analysé dans un cadre dynamique. Pour cela, en modélisant le choix de souscrire, nous utilisons le fait qu'il s'agit d'un choix ordonné et homogène : les tranches ont la même « grandeur » - en kWh - et coût. L'individu n'est pas « lié » à un nombre de tranches plutôt qu'à un autre et nous croyons que le processus de choix puisse être mieux approché par des *modèles ordonnés*, que nous allons expliquer et employer par la suite.

4.6 Modèles ordonnés

Dans l'explication des modèles des variables qualitatives, la théorie des modèles polythomiques (ou multinomiaux) opère une distinction entre modèles ordonnés et non ordonnés. Les modèles ordonnés sont employés lorsque la variable dépendante (la réponse observée) reflète l'appartenance à un des intervalles successifs d'une variable latente continue inobservée.

⁶ Pour une exposition détaillé voir : Amemiya, T., « Advanced econometrics », Cambridge University Press, 1985, Ch. 6.

⁷ Maddala, G.S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge, University Press, pag. 51.

⁸ Small, K. A., (1987), "A Discrete choice model for ordered alternatives", *Econometrica*, Vol 55, No. 2, 409-424.

⁹ Maddala G.S. *Ibid.* pag. 49.

¹⁰ Kahn, L. et Morimune K. (1979), « Unions and Employment Stability : A Sequential Logit Approach », *International Economic Review*, **20**, 217-235.

¹¹ McFadden, D. (1980), "Econometric Models of Probabilistic Choice" in C.F. Manski and D. McFadden eds. , *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass. : MIT Press.

Les modèles ordonnés s'appliquent à l'explication, entre autre, du nombre de voitures achetées, des états-reponse aux doses d'un insecticide, ou tout autre ordre de préférences (ou tolérances). D'un point de vue plus général, les modèles ordonnés ont acquis un rôle important dans le traitement des données de sondage, où le "jugement" se trouve sur une échelle qualitative ordonnée.

4.6.1 Spécification des modèles ordonnés

Un modèle ordonné est par définition polythomique. Sa modélisation probabiliste se base sur l'hypothèse d'inobservabilité de la variable dépendante d'intérêt Y .¹² L'analyste peut observer seulement une version discrète et ordinale (Z) de la variable précédente, pour laquelle le modèle linéaire est inadéquat.

Pour mieux comprendre, il suffit d'imaginer l'état d'avancement d'une maladie avec un certain détail : l'état de l'individu sera caractérisé par différents adjectifs (e.g. "mauvais", "bon", "très bon"). Le modélisateur, intéressé à standardiser l'information disponible et à employer des méthodes probabilistes, va *coder* les états en une échelle qualitative ordonnée telle : 0, 1, 2,...N.

Pour un individu j , la variable théorique d'intérêt satisfait :

$$Y_j = \beta' X_j + u_j$$

avec l'hypothèse que le terme d'erreur u soit normal multivarié, soit : $u \sim N(0, \sigma^2 I)$

La variable latente Y_j représente l'état réel de l'individu j , inobservable. La variable observée Z_j est catégorielle avec M catégories-reponses : $R_1, R_2 \dots, R_m$. Les réponses résultent de la variable inobservée Y de la manière suivante : il existe $M + 1$ nombres réels, $\mu_0, \mu_1, \dots, \mu_M$, avec $\mu_0 = -\infty, \mu_M = +\infty, \mu_0 \leq \mu_1 \leq \dots, \leq \mu_M$ et tels que :

$$Z_j \in R_i \Leftrightarrow \mu_{i-1} \leq Y_j \leq \mu_i \text{ pour } 1 \leq j \leq M.$$

Puisque Z est ordinale il peut être représenté par une série de variables dichotomiques pour l'estimation :

$$Z_{jk} = \begin{cases} 1 & \text{si } Z_j \in R_k \\ 0 & \text{sinon} \end{cases}$$

pour $1 < j < n$ et $1 < k < M$.

Nous pouvons dès lors écrire :

$$Z = \begin{bmatrix} Z_{11} & \cdots & Z_{1M} \\ \vdots & & \vdots \\ Z_{n1} & \cdots & Z_{nM} \end{bmatrix}$$

¹² Le modèle est du à McElvey, R. et Zavoina, W (1975), « A Statistical Model for the Analysis of Ordinal Level Dependent Variables », Journal of Mathematical Sociology, Summer, **4**, pp. 103-120.

Les probabilités pour $l \leq i \leq M$ et $l \leq j \leq n$ sont :

$$\mu_{i-1} < Y_j < \mu_i \Leftrightarrow \mu_{i-1} < \sum_{i=0}^K \beta_i X_{ij} + u_j < \mu_i \Leftrightarrow \frac{\mu_{i-1} - \sum_{i=0}^K \beta_i X_{ij}}{\sigma} < \frac{u_j}{\sigma} < \frac{\mu_i - \sum_{i=0}^K \beta_i X_{ij}}{\sigma}$$

et, puisque u_j est normal multivarié, nous pouvons écrire :

$$Pr(Z_{jk} = l) = Pr(Z_{jk} \in R_k) = \Phi \left[\frac{\mu_k - \sum_{i=0}^K \beta_i X_{ij}}{\sigma} \right] - \Phi \left[\frac{\mu_{k-1} - \sum_{i=0}^K \beta_i X_{ij}}{\sigma} \right] \quad (55)$$

Avec $\Phi(t)$ la cumulative normale :

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

4.6.2 Le problème de l'identifiabilité

Dans ce modèle il y existe un problème d'identifiabilité des paramètres structurels, puisque toute transformation lineaire de l'échelle de reference Y appliquée aux paramètres μ_0, \dots, μ_M résulte en le même modèle ; les probabilités sont en effet invariantes à une transformation linéaire affine de la forme :

$$\begin{aligned} Y^* &= a + bY, \\ \mu^* &= a + b\mu \end{aligned}$$

Ceci car :

$$Prob(y_i^* = l) = P_{i1}^* = Prob(y_i^* < \mu^*) = Prob(a + by_i < a + b\mu_i) = Prob(y_i < \mu_i)$$

La procédure courante consiste à assumer, sans perte de généralité, que $\mu_l = 0$ et $\sigma = 1$.¹³

Le modèle d'intérêt devient :

$$Pr[Z_k = l] = \Phi \left[\mu_k - \sum_{i=0}^K \beta_i X_{ij} \right] - \Phi \left[\mu_{k-1} - \sum_{i=0}^K \beta_i X_{ij} \right] \quad (56)$$

pour l'estimation des $Q = M + K - 1$ paramètres $\mu_2, \dots, \mu_{M-1}, \beta_0, \dots, \beta_K$.

Les μ_i peuvent s'interpréter comme des *seuils* qui déterminent le J choisit par l'individu ; l'estimation est faite par le maximum de vraisemblance.

Nous rappelons les variantes principales parmi les modèles ordonnés¹⁴:

¹³ D'autres contraintes d'identification sont reportées dans Jackman, S. (Spring 2000) « Models for Ordered Outcomes », Political Science 200C, page 4 du document.

- logit ordonnée
- probit ou logit ordonnée pour données de panel avec effets aléatoires
- probit ou logit ordonnée avec stratification¹⁵
- probit ou logit ordonnée pour données censurés
- probit ou logit ordonnée avec hypothèse d'hétéroschélasticité selon deux formes :
 - Var[ε_i] = w
 - Var[ε_i] = {exp[$\gamma'z_i$]}²
- probit ou logit ordonnée avec mécanisme de sélection.

4.6.3 Les estimateurs du maximum de vraisemblance

Concernant la notation, nous écrivons :

$$Y_{j,k} = \mu_k - \sum_{i=0}^K \beta_i X_{ij} \quad (57)$$

$$\Phi_{j,k} = \Phi(Y_{j,k}) \quad (58)$$

L'expression (2.10) peut s'écrire :

$$\Pr[Z_{j,k} = 1] = \Phi_{j,k} - \Phi_{j,k-1} \quad (59)$$

La vraisemblance de Z, étant donné les paramètres c'est :

$$L = L(Z | \beta_0, \dots, \beta_K, \mu_2, \dots, \mu_{M-1}) = \prod_{j=1}^n \prod_{k=1}^M (\Phi_{j,k} - \Phi_{j,k-1})^{Z_{j,k}} \quad (60)$$

et la log- vraisemblance L*:

$$L^* = \log L = \sum_{j=1}^n \sum_{k=1}^K Z_{j,k} \log(\Phi_{j,k} - \Phi_{j,k-1}) \quad (61)$$

Le problème consiste à maximiser la log-vraisemblance par rapport aux paramètres μ_2, \dots, μ_{M-1} , β_0, \dots, β_K , sous les contraintes : $\mu_2 \leq \mu_3, \dots, \leq \mu_{M-1}$.

Nous posons :

$$N_{j,k} = \frac{1}{\sqrt{2\pi}} e^{-\frac{Y_{j,k}^2}{2}} \quad \text{pour } 1 \leq k \leq M \quad \text{et } 1 \leq j \leq n \quad \text{et} \quad (62)$$

$$\delta = \begin{cases} 1 & \text{si } = j \\ 0 & \text{si } \neq j \end{cases}$$

les dérivées partielles sont :

¹⁴ Basée sur: W.H. Greene *LIMDEP 7.0 User's Manual*, Econometric Software Inc. pag. 486.

¹⁵ LIMDEP 7.0 peut estimer jusqu'à 9 strates, à l'intérieur desquelles les paramètres sont les mêmes, tandis que les seuils varient.

$$\frac{\partial}{\partial \beta_u} \Phi_{j,k} = N_{j,k} \left[\frac{\partial}{\partial \beta_u} Y_{j,k} \right] = -N_{j,k} X_{u,j} \text{ pour } 0 \leq u \leq K \quad (62)$$

$$\frac{\partial}{\partial \mu_u} \Phi_{j,k} = N_{j,k} \left[\frac{\partial}{\partial \mu_u} Y_{j,k} \right] = N_{j,k} \delta_{u,k} \text{ pour } 2 \leq u \leq M-1 \quad (63)$$

et

$$\frac{\partial}{\partial \beta_u} N_{j,k} = Y_{j,k} N_{j,k} X_{u,j} \text{ pour } 0 \leq u \leq K \quad (64)$$

$$\frac{\partial}{\partial \mu_u} N_{j,k} = -Y_{j,k} N_{j,k} \delta_{u,k} \text{ pour } 2 \leq u \leq M-1 \quad (65)$$

Ces expressions permettent le calcul des Q dérivés partielles de la log-vraisemblance par rapport aux paramètres inconnus, ainsi que la matrice $Q \times Q$ des dérivées secondes. L'expression analytique des dérivées premières :

$$L^* = \frac{\partial L}{\partial \beta_u} = \sum_{j=1}^n \sum_{k=1}^M Z_{j,k} \left[\frac{N_{j,k-1} - N_{j,k}}{\Phi_{j,k} - \Phi_{j,k-1}} \right] X_{u,j} \quad (66)$$

pour $0 < u < K$, et

$$\frac{\partial L^*}{\partial \mu_u} = \sum_{j=1}^n \sum_{k=1}^M Z_{j,k} \left[\frac{N_{j,k} \delta_{k,u} - N_{j,k} \delta_{k-1,u}}{\Phi_{j,k} - \Phi_{j,k-1}} \right] \quad (67)$$

pour $2 < u < M-1$.

Finalement, les dérivées secondes s'écrivent¹⁶ :

$$\frac{\partial^2 L^*}{\partial \beta_u \partial \beta_v} = \sum_{j=1}^n \sum_{k=1}^M Z_{j,k} \left[\frac{(\Phi_{j,k} - \Phi_{j,k-1})(N_{j,k-1} Y_{j,k-1} - N_{j,k} Y_{j,k})}{(\Phi_{j,k} - \Phi_{j,k-1})^2} - \frac{(N_{j,k-1} - N_{j,k})^2}{(\Phi_{j,k} - \Phi_{j,k-1})^2} \right] X_{u,j} X_{v,j} \quad (68)$$

$$\frac{\partial^2 L^*}{\partial \beta_u \partial \mu_v} = \sum_{j=1}^n \sum_{k=1}^M Z_{j,k} \left[\frac{(\Phi_{j,k} - \Phi_{j,k-1})(N_{j,k} Y_{j,k} \delta_{k,v} - N_{j,k} Y_{j,k-1} \delta_{k-1,v})}{(\Phi_{j,k} - \Phi_{j,k-1})^2} - \frac{(N_{j,k-1} - N_{j,k})(N_{j,k} \delta_{k,v} - N_{j,k-1} \delta_{k-1,v})}{(\Phi_{j,k} - \Phi_{j,k-1})^2} \right] X_{u,j}$$

$$\frac{\partial^2 L^*}{\partial \beta_u \partial \mu_v} = \sum_{j=1}^n \sum_{k=1}^M Z_{j,k} \left[\frac{(\Phi_{j,k} - \Phi_{j,k-1})(N_{j,k-1} Y_{j,k-1} \delta_{u,k-1} \delta_{v,k-1} - N_{j,k} Y_{j,k} \delta_{u,k} \delta_{v,k})}{(\Phi_{j,k} - \Phi_{j,k-1})^2} - \frac{(N_{j,k} \delta_{u,k} - N_{j,k-1} \delta_{u,k-1})(N_{j,k} \delta_{v,k} - N_{j,k-1} \delta_{v,k-1})}{(\Phi_{j,k} - \Phi_{j,k-1})^2} \right]$$

Le calcul numérique des fonctions objectif est fait par différents algorithmes tels : Broyden, Davidon, Fletcher et Powell ou Berndt, Hall, Hall et Hausman, Newton-Raphson et pente

¹⁶ Tiré de McKelvey et Zavoina (1975) page 108.

maximale (*steepest descent*) le programme LIMDEP utilise par défaut l'algorithme de Broyden.

Le chapitre suivant décrit les différentes formulations de probit ordonné que nous avons essayé. Dans tous les cas la significativité des seuils est nette.

5 Estimations

5.1 Descriptif des procédures

Nous avons essayé les spécifications suivantes :

1. Nous avons spécifié un modèle dichotomique pour vérifier si des variables explicatives *discriminent* significativement les individus qui ont souscrit une tranches des autres.
2. Nous avons ensuite essayé un modèle de Poisson avec et sans hétérogénéité en supposant que le mécanisme de souscription soit représentable comme un *dénombrement*.
3. Nous avons finalement spécifié un modèle ordonné pour décrire le mécanisme d'adhésion à la bourse par des *seuils*.

Des premières estimations probit et logit dichotomiques ont été faites avec une base partielle obtenue en été 2001 et constituée du sous-échantillon des adhérents 2001 (env. 60 individus). Les résultats ont globalement montré peu de significativité des paramètres : dans un premier temps la variable « consommation¹ » semblait jouer un rôle pour le choix de souscrire plus qu'une tranche, mais les estimations successives rejetaient cette hypothèse.²

5.1.1 Variables créées

Nous avons créé les variables suivantes :

- « *dmytarte* », variable dichotomique prenant la valeur « 1 » si l'individu a le tarif « tout électrique ». Le choix est motivé par le fait que tel tarif est souvent présent chez les gens habitant des villas ayant un système de chauffage avec pompe à chaleur ;
- « *dmytarU2* », variable dichotomique qui indique le mode de tarification différencié entre heures pleines et heures creuses ; elle est attribué aux activités professionnelles comme aux particuliers;
- « *genrebin* », variable dichotomique prenant la valeur « 1 » en cas de (co) signature par « madame » dans le contrat d'adhésion à la bourse solaire³ ;
- « *txkwh* », variable indiquant la variation en pourcentage entre la consommation 1999 et 2000
- « *dmygenf* », binaire, qui prends la valeur 1 si l'individu habite dans la commune de Genève
- « *txkwh* », quantitative, indiquant la variation de la consommation entre les années 1999 et 2000.

5.1.2 Données aberrantes

Une des premières opérations avant d'estimer un modèle est la vérification des plages des données, afin de limiter la dispersion des variables de consommation (CHF et kWh) ou

¹ La valeur a été reportée d'une façon quelque peu « artisanale » : en prenant la période d'intérêt sur l'interface *Data Warehouse* en service aux SIG pour chaque individu.

² Pour les détails des différentes spécifications essayé voir Annexe # *Les premières spécifications*.

³ La commande est : `CREATE;If(genrequa=3)genrebin=0;(else)genrebin=1$`

d'exclure des fausses données. Nous avons affiné l'estimation en éliminant des données extrêmes, c'est à dire. :

- 4 individus de tarif $Ei(B)$, moyenne tension (-4 dans l'échantillon),
- les 15 sociétés de l'échantillon (- 11 dans l'échantillon),
- les 14 individus entre 6 et 700 tranches (-13 dans l'échantillon)
- le individu avec consommation annuelle supérieure à 7.000 CHF (l'échantillon reste pareil)
- les individus avec consommation annuelle inférieure à 200 CHF (l'échantillon diminue de 56 individus)⁴.

L'échantillon final est ainsi composé de 569 individus.

5.2 Modèles dichotomiques

L'hypothèse sur le processus de choix étant que l'individu a deux façons d'adhérer à la bourse :

- « faire un geste » = il achète une tranche ;
- « investir » = l'individu souscrit un montant supérieur, fonction de :
 - consommation ;
 - sensibilité environnementale, orientation politique ;
 - scrupules, mauvaise conscience ;
 - autres variables inobservables, telles la richesse ou le revenu.

Dans un modèle qualitatif dichotomique la variable d'intérêt est binaire et elle représente le choix entre « *une tranche* » ou « *plusieurs tranches* » ou, plus précisément, la probabilité de les souscrire.

D'un point de vue formel, la variable de prévision est codé ainsi :

$$Y_i = 0 = \text{Probabilité}(\text{individu } i \text{ achète 1 tranche}) = F(\beta X)$$
$$Y_i = 1 = \text{Probabilité}(\text{individu } i \text{ achète plus qu'une tranche}) = 1 - F(\beta X)$$

Le modèle dichotomique fournit un seuil, fonction des variables explicatives, qui discrimine entre les deux types de réponses (adhérents).

La fonction de répartition $F(.)$ est soit la loi normale soit la logistique.

Nous présentons ci-dessous les estimations de type logit et probit effectuées.

⁴ Le tarif E(B) consomment pour des dizaines de milliers de francs (pour les autres il s'agit de milliers), d'autre part la logique de choix d'entités telle *sociétés* ou *collectivités publiques* ne peut pas être rapportée à un simple individu (décision d'entreprise, stratégie d'image ou électorale etc.). La commande LIMDEP c'est : REJECT;Genrequa=4 | tranc0>4 | chf00>5\$.

| Variable (t-valeur) | Probit | Logit |
|---------------------|----------------------------------|-----------------------------------|
| ONE | -0.536 (-2.623) | -0.8709 (-2.565) |
| CENTCOM | -0.004 (-1.324) | -0.007 (-1.290) |
| CHF99MIL | 0.206 (1.725) | 0.341 (1.746) |
| DMYTE | 0.310 (0.763) | 0.477 (0.727) |
| DMYU2 | 0.473 (2.355) | 0.771 (2.373) |
| DMYGENF | 0.144 (0.983) | 0.235 (0.951) |
| TXVARK | -0.000 (0.000) | 0.000 (-0.194) |
| GENREBIN | -0.035 (-0.311) | -0.058 (-0.306) |
| % succès réf. 70% | 71% | 71% |

Tableau 5.1 – Estimation pour les modèles probit et logit dichotomique

Les deux spécifications sont pratiquement équivalentes, ce qui est évident en regardant le critère des prédictions justes. Aucune estimation est satisfaisante, car seule la variable *dmyU*, qui indique la double tarification, est significative. De plus, le signe négatif du coefficient de la variable *centcomm* est surprenant, car nous aurions attendu que la richesse de la commune influence positivement les souscriptions. Une raison pour cela peut être le fait que la commune de Genève, qui fait partie des communes « riches », incorpore 32% des tranches souscrites alors que son poids démographique est de 42% : une grande partie des adhérents genevois ont donc pris une tranche seulement ce qui a engendré un coefficient global négatif. Pour les détails, nous renvoyons au tableau complet des centimes communaux (Annexe 3).

Nous remarquons que le fait d'avoir une double tarification affecte le choix de souscrire plusieurs tranches ce qui peut confirmer que le choix se fasse en fonction de la facture et non pas simplement pour adhérer (5.2). Nous nous serions attendus à avoir aussi un résultat significatif pour le coefficient du tarif tout électrique (*dmyte*), les estimations successives vont conforter cette hypothèse.

5.3 Modèles de Poisson

Après avoir spécifié le processus de choix comme binaire (une tranche ou plus), nous avons estimé un modèle de Poisson. L'idée sous-jacente est que le choix de combien de tranches souscrire peut être approché par un processus de dénombrement de type : 1, 2, ..., N. Une critique essentielle à cette approche étant que les processus de Poisson ont lieu dans le temps ; nous justifions notre choix d'employer tel modèle en faisant l'hypothèse que le « dénombrement » des tranches à souscrire se fasse dans le laps de temps qui passe entre le moment où l'individu prend la décision d'adhérer et quand il choisit le nombre définitif de tranches.

Nous avons estimé la spécification Poisson et binomiale négative (4.2.3) pour vérifier si une hétérogénéité de type gamma pouvait expliquer le comportement des individus. A ce propos,

nous rappelons que la principale faiblesse du modèle est que même avec la moyenne et la variance du modèle correctement spécifiés, si la distribution sous-jacente n'est pas la binomiale négative, l'estimateur du MV est inconsistant (Cameron et Trivedi 1990).

Une possibilité alternative, offerte par le logiciel Limdep est d'assumer une hétérogénéité normale. Nous avons essayés cette option, toutefois toutes les estimations effectuées ont été insatisfaisantes. Dans la fin du sous-chapitre nous avons testé l'hypothèse de surdispersion du modèle, qui résulte confirmée par le test de signification du paramètre d'hétérogénéité effectué dans le chapitre 5.4.2.

5.3.1 Poisson simple et binomiale négative

Nous avons estimé les deux spécifications, avec la base des données des clients solaires décrite en 5.2.1, pour toutes les variables disponibles. Les résultats sont présentés ci-dessous :

| Coefficient (t-valeur) | Poisson | Binomiale négative |
|-----------------------------------|--------------------|-----------------------------------|
| ONE | 0.407 (0.126) | 0.408125 (1.955) |
| CENTCOM | -0.001 (0.002) | -0.001 (-0.447) |
| CHF99MIL | 0.023 (0.071) | 0.023 (0.188) |
| DMYTE | 0.396 (0.234) | 0.396 (1.420) |
| DMYU2 | 0.201 (0.123) | 0.202 (1.142) |
| DMYGENT | 0.059 (0.090) | 0.059 (0.411) |
| TXVARKWH | -0.000 (0.000) | -0.000 (-0.190) |
| GENREBIN | -0.058 (-0.820) | -0.058 (-0.503) |
| Alpha | | 0.000 (0.000) |

Tableau 5.2 – Estimation pour les modèles de Poisson et binomiale négatif

Le modèle est nettement insatisfaisant : seule la constante est significative pour la spécification binomiale négative, en plus la log-vraisemblance est très proche de la valeur pour le modèle sans variables explicatives. Nous devons souligner que lors des estimations avec la base de données avec les sociétés et collectivité publiques les coefficients de consommations étaient significatifs ; ce qui nous confirme la non robustesse du modèle aux données extrêmes.

Les autres variantes de modèles de Poisson fournissent des résultats semblables : la variables de consommation n'est pas significative ainsi que toutes les autres combinaisons de variables présentés. Dans toutes les spécifications seulement la constante est significative. Ce résultat est confirmé avec des spécification Poisson avec une variable explicative à la fois en plus de la constante.

5.3.2 Les non-résidents à Genève

Nous avons essayé d'estimer un modèle pour les non-résidents dans la commune de Genève (276 individus) pour différentes variables. L'idée étant de vérifier s'il y a des changements dans cette composante de la population. Il est en effet possible qu'en enlevant les résidents de la commune de Genève, qui « noient » l'information disponible, des caractéristiques puissent s'afficher. Une autre possibilité est de voir si les propriétaires des villas et les habitants des communes rurales marquent une différence par rapport à la population totale, mais les résultats se sont encore une fois révélés insatisfaisants :

| N=276 | Variable | Coefficient | b/St.Er |
|----------------------|----------|-------------|---------|
| MCO | Constant | 1.594 | 13.282 |
| | Centcom | 0.002 | -0.922 |
| Poisson | Constant | 0.469 | 4.360 |
| | Centcom | -0.001 | -0.691 |
| Bin. Négative | Constant | 0.469 | 3.221 |
| | Centcom | -0.001 | -0.496 |
| | Alpha | 0.000 | 0.083 |

Tableau 5.3 – Poisson et NegBin pour les non-résidents à Genève

Nous avons enfin estimé un modèle pour les non résidents à Genève en excluant seulement les individus avec une facture supérieure à 5000CHF avec toutes les variables. Les résultats sont présentés ci-dessous :

| N=279 | Variable | Coefficient | b/St.Er |
|------------------|----------|---------------|---------|
| Poisson | Centcom | -0.004 | -3.227 |
| Log-vr. | Chf00mil | 0.460 | 9.195 |
| -472.9461 | Txvarkwh | -0.000 | -1.152 |
| Res. log-vr. | Genrebin | 0.338 | 3.899 |
| -534.5859 | Dmttarte | -0.958 | -3.895 |
| | Dmytarpr | 0.718 | 6.552 |
| Neg. Bin. | Centcom | -0.001 | -1.03 |
| Log-vr. | Chf00mil | 0.374 | 8.538 |
| -442.4954 | Txvarkwh | -0.000 | -0.565 |
| Log-vr. Res. | Genrebin | 0.253 | 1.841 |
| -472.9461 | Dmttarte | -0.674 | -2.114 |
| | Dmytarpr | 0.628 | 4.616 |
| | Alpha | 0.184 | 3.414 |

Tableau 5.4 – Poisson et NegBin pour les non-résidents à Genève avec CHF2000<5000

Les variables significatives à 5% sont marquées en gras. Il est à remarquer qu'avec cet échantillon les coefficients associés aux dummy pour le genre et le type de tarif sont significatives, par contre ni les centimes communaux, ni la variation de la consommation ont une influence. Le fait que le type de tarif et la consommation soient très significatifs (davantage la double tarification que le tout électrique), nous révèle l'importance d'une catégorie d'individus qui sont plus conscients que le reste des adhérents du *niveau* de consommation et choisissent d'acheter de l'énergie verte en conséquence. Ces derniers éléments nous confortent dans l'hypothèse – testé dans les premières spécifications dichotomiques - que les tarifs différenciés rendent les gens plus conscients de la quantité et qualité de courant qu'ils consomment. Il faut aussi dire que la consommation moyenne pour cette sous population est 30% plus élevée que dans l'échantillon complet.

5.3.3 Test de surdispersion du modèle de Poisson

Nous avons effectué un test pour vérifier la sur-dispersion des données. Le test est basé sur une approche par régression de l'égalité entre moyenne et variance (Cameron et Trivedi 1990). L'hypothèse est :

$$H_0: \text{Var}[y_i] = \mu_i \quad (1)$$

$$H_A: \text{Var}[y_i] = \mu_i + \alpha g(\mu_i) \quad (2)$$

avec les possibilités⁵ : $g(\mu) = \mu$ et $g(\mu) = \mu^2$.

En récrivant (2) comme :

$$E[(y_i - \mu_i)^2 - y_i] = \alpha \cdot g(\mu_i)$$

Avec les $\mu_i = \exp\{X_i\beta\}$ le test serait un t-test pour $\alpha = 0$ de la régression :

$$(y_i - \mu)^2 - y_i = \alpha \cdot g(\mu) + \varepsilon$$

Les μ_i étant inobservés ils sont remplacés par une estimation (basée sur l'estimation de β du modèle Poisson). Il est à remarquer qu'il n'est pas nécessaire d'avoir une estimation consistente de β pour la consistance de α (Cameron et Trivedi 1990).

Numériquement⁶, la statistique du test T_{opt} est fournie en testant la significativité du coefficient de la régression MCO de $z_i = w_i + \varepsilon_i$. Où:

$$z_i = \frac{[(y_i - \mu_i)^2 - y_i]}{\sqrt{2\mu_i}}$$

et

$$w_i = \frac{g(\mu_i)}{\sqrt{2\mu_i}}$$

L'avantage principal du test est qu'il ne requiert pas la spécification de la distribution sous l'hypothèse alternative, il construit la statistique en partant de l'erreur empirique (μ) sous H_0 . Nous avons opéré ce test en partant de l'estimation d'un modèle avec la variables explicative *CHF1999* (plus la constante) ; l'estimation est présentés ci-dessous :

| | Coeff. | Std.Err. | t-ratio | P-value |
|-----------------|---------------|-----------------|----------------|----------------|
| ONE | 0,324 | 0,056 | 5,727 | 0.000 |
| CHF99MIL | 0,020 | 0,069 | 0,291 | 0,770 |
| DMYU2 | 0,184 | 0,121 | 1,519 | 0,128 |
| DMYTE | 0,367 | 0,231 | 1,587 | 0,112 |

Tableau 5.5 – Résultats de l'estimation avec la variables explicative *CHF1999*

⁵ Ces deux possibilité correspondent aux modèles NEGBIN I et NEGBIN II dans Cameron et Trivedi 1990.

⁶ Limdep® Version 7.0 page 606.

Ensuite, la régression linéaire du paramètre de sur-dispersion fournit :

| | Coeff. | Std.Err. | t-ratio |
|-----|-----------|-----------|----------|
| WI1 | -0,538288 | 0,0347181 | -15,5045 |
| WI2 | -0,364526 | 0,0240432 | -15,1613 |

La variance du modèle peut enfin s'écrire :

$$\text{Var}(y_i/X_i, \beta) = \mu_i - 0.538\mu_i$$

$$\text{Var}(y_i/X_i, \beta) = \mu_i - 0.36\mu_i^2.$$

Les coefficients négatifs montrent une forte significativité pour les deux paramètres de sous-dispersion estimés. Un commentaire peut concerner le type des données : les troncatures opérés ont également soustrait de la variabilité aux variables explicatives, mais il est évident que la leçon principale de l'approche de Poisson est que les occurrences doivent être prises dans une dimension temporelle. Ceci constitue la principale faiblesse du modèle de dénombrement dans le cas des données solaires.

Dans la suite du chapitre nous allons présenter les résultats pour différentes spécifications de modèles Probit ordonnés.

5.4 Probit ordonné

Ce chapitre se compose de deux parties : dans la première nous avons estimé la variable dépendante discrète *tranches souscrites* (1, 2,...), dans la deuxième nous avons considéré une version discrète du *taux de substitution* du courant PV par rapport aux kWh conventionnels consommés dans l'année 1999.

Le fait de prendre une variable construite avec une donnée exogène (la consommation électrique de l'individu en 1999) a pour but de calibrer la variable d'intérêt – la demande d'énergie solaire - au niveau de consommation à la consommation de l'individu. Cette démarche - a notre avis - explique mieux le choix de *combien* souscrire pour chaque individu.

5.4.1 Les tranches souscrites

Le choix de la forme de la variable dépendante s'est révélé vite important : si d'un côté nous sommes bornés aux valeurs comprises entre 1 et 5, puisque cela représente environ 95% des effectifs, de l'autre il se peut que différentes répartitions puissent mieux expliquer le processus du choix.

Par conséquence, afin d'inclure toutes les possibilités, nous avons créé les variables *trantri*, *tranquad* et *trancin* pour représenter les choix de 1 à 3, de 1 à 4 et de 1 à 5 respectivement⁷.

Nous avons essayé une spécification probit trichotomique ordonné, en sélectionnant :

- les individus avec une facture annuelle comprise entre 300 et 5000 francs (517 sur 653)
- les clients individuels (15 sociétés et collectivité publiques exclues)

⁷ Pour être exactes, LIMDEP estime les modèles ordonnés avec la variable d'intérêt codée: 0, 1,...J.

- les acheteurs de moins que 6 tranches (15 individus)

Les résultats présentés ci-dessous concernent un échantillon de 501 individus :

Variable expliquée: TRANTRI0

| N=501 | Variable | Coeff. | t-ratio |
|---------------------------|----------|--------|-----------|
| Probit | ONE | -0.478 | -2.294 |
| | CENTCOM | -0.004 | -1.248 |
| | CHF00MIL | 0.103 | 0.933 |
| | DMYTE | 0.571 | 1.589 |
| | DMYU2 | 0.446 | 2.254 |
| | DMYGENF | 0.204 | 1.315 |
| | TXVARKWH | -0.001 | -1.158 |
| | Mu(1) | 0.839 | 11.227 |
| Log likelihood function | | | -392.0092 |
| Restricted log likelihood | | | -400.6575 |

A part deux individus, le modèle ne prévoit que des souscription pour une tranche, le taux de succès est 70%, mais il est proche de la référence (modèle avec constante).

Pour la même sélection nous estimons un modèle pour expliquer le choix jusqu'à quatre tranches (variable expliquée : *TRANQUAD*)⁸ ; les coefficients sont :

| N=501 | Variable | Coefficient | b/St.Er | Mean of X |
|---------------------------|----------|-------------|-----------|--------------|
| | ONE | -0.469 | 0.207 | 0.0233027 |
| | CENTCOM | -0.004 | 0.003 | 0.24164 |
| | CHF00MIL | 0.085 | 0.116 | 0.462649 |
| | DMYTE | 0.552 | 0.355 | 0.119885 |
| | DMYU2 | 0.430 | 0.204 | 0.0353462 |
| | DMYGENF | 0.187 | 0.153 | 0.221364 |
| | TXVARKWH | -0.001 | 0.000 | 0.128535 |
| | Mu(1) | 0.839 | 0.074 | 2.88658e-015 |
| | Mu(2) | 1.254 | 0.102 | 2.88658e-015 |
| Log likelihood function | | | -423.8896 | |
| Restricted log likelihood | | | -432.3681 | |

Le modèle avec la variable *trancin* n'affiche pas des meilleurs résultats : les coefficients sont dans l'ensemble pas très significatifs ; le modèle ne prédit que des souscriptions pour une tranche d'électricité solaire.

Par rapport aux modèles précédents, la significativité du paramètre associé au montant de la facture pour l'année 2000 (*chf00mil*) baisse légèrement, tandis que les variables binaires indicatrices du tarif augmentent. Les valeurs sont quasi identiques aux modèle pour quatre tranches. L'option de la cinquième tranches est donc pratiquement sans influence pour nos fins de modélisation.

Une spécification avec des variables significatives est présentée ci-dessous, l'échantillon est composé des individus avec une facture d'électricité inférieure à CHF 5000 et qui ont souscrit jusqu'à 5 tranches, la variable expliquée est *trantri0* :

⁸La commande est : CREA ; If (tranches=1) tranquad=0 ; if (tranches=2) tranquad=1 ; If (tranches=3) tranquad=2 ; If (tranches<3) tranquad=3\$.

| N=501 | Variable | Coefficient | b/St.Er | P[Z >z |
|---------------------------|----------|-------------|------------|--------------|
| Probit | ONE | -0.584226 | -9.23995 | 2.88658e-015 |
| | DMYTE | 0.731553 | 2.72721 | 0.00638721 |
| | DMYU2 | 0.446757 | 2.39588 | 0.0165803 |
| | Mu(1) | 0.83142 | 11.3173 | 2.88658e-015 |
| Log likelihood function | | | - 395.0965 | |
| Restricted log likelihood | | | - 400.6575 | |

Malheureusement, le tableau des prévisions indique que le modèle ne prédit que des souscriptions pour une tranche.

5.4.2 Modèle pour les non-résidents dans la commune de Genève

Comme auparavant (5.4.2) nous voulons vérifier si les individus résidents dans la commune de Genève biaisent nos estimations. Nous avons ainsi essayé un modèle pour les non-résidents dans la commune de Genève. L'échantillon est maintenant de 251 individus ; les résultats sont présentés ci-dessous :

| N=251 | Variable | Coefficient | b/St.Er | P[Z >z | Mean of X |
|-------|----------|-------------|---------|---------|-----------|
| | ONE | -0.624 | -3.170 | 0.001 | |
| | CENTCOM | -0.005 | -1.592 | 0.111 | 48.092 |
| | CHF00MIL | 0.305 | 2.952 | 0.003 | 1.02 |
| | DMYU2 | 0.445 | 2.013 | 0.044 | |
| Seuil | Mu(1) | 0.895 | 8.242 | 0.000 | |

La négativité du coefficient de la variable *centcomm* est à imputer à la disproportion entre communes riches et pauvres ; en effet puisque des communes très peuplées et « pauvres », telles, Bernex ou Onex ont souscrit des montant très variés, le coefficient indique une relation négative entre la variable et les souscription. Les résultats prédits et observés sont :

| | | Predicted | | | |
|--------|-----|-----------|---|-------|--|
| Actual | 0 | 1 | 2 | Total | |
| 0 | 174 | 0 | 0 | 174 | |
| 1 | 53 | 0 | 2 | 55 | |
| 2 | 21 | 0 | 1 | 22 | |
| Total | 248 | 0 | 3 | 251 | |

Ce qui montre que le modèle explique 69% des tranches souscrites, cette valeur étant égale à celle d'un modèle incluant seulement la constante.

5.4.3 Probit ordonné, essais avec une seule variable explicative

Toutes les estimations avec une seule variable explicative montrent une forte significativité de la constante, la variable explicative étant non significative ainsi qu'une faible capacité prédictive : le modèle ne prévoit que des souscription pour une tranche.

5.5 Substitution du solaire vs. conventionnel

Insatisfaits par les résultats obtenus jusqu'ici, nous avons créé une variable représentative de la demande d'électricité solaire en fonction de la consommation de l'année précédente. La variable indique le rapport en pourcentage entre les kWh solaires achetés en 2000 et les kWh

consommés l'année précédente. Les valeurs obtenues sont ensuite regroupés par segments, de façon à obtenir une variable discrète - *subs0123* - qui vaut :

- de 0.1% à 5% = 0
- de 5.1% à 10% = 1
- de 10.1% à 25% = 2
- plus que 25% = 3

L'estimation produit les résultats suivants :

| | Coeff. | Std.Err. | t-ratio | P-value |
|---------------------------|----------|-----------|-----------|--------------|
| ONE | 2.7547 | 0.179614 | 15.3368 | 2.88658e-015 |
| CHF99MIL | -4.17538 | 0.234029 | -17.8413 | 2.88658e-015 |
| DMYTE | 1.99858 | 0.613652 | 3.25686 | 0.00112651 |
| Mu(1) | 1.39537 | 0.0910833 | 15.3197 | 2.88658e-015 |
| Mu(2) | 2.57086 | 0.125335 | 20.5119 | 2.88658e-015 |
| Log likelihood function | | | -485.7255 | |
| Restricted log likelihood | | | -691.9307 | |

| Actual | Predicted | | | | Total |
|--------|-----------|-----|----|---|-------|
| | 0 | 1 | 2 | 3 | |
| 0 | 194 | 45 | 0 | 0 | 239 |
| 1 | 31 | 134 | 0 | 0 | 165 |
| 2 | 8 | 53 | 14 | 0 | 75 |
| 3 | 0 | 13 | 5 | 3 | 21 |
| Total | 233 | 245 | 19 | 3 | 500 |

La taux de succès est bien de 70% qui se compare très favorablement avec le 47% obtenu pour le modèles avec la seule constante.

Nous avons ensuite regroupé les individus ainsi :

- de 0.1 % à 10 % = 0
- de 10.1 % à 25 % = 1
- plus que 25 % = 2

en créant la variable *subs012* pour estimer un modèle trichotomique ordonné.

Les résultats ont montrés une puissance prédictive proche de celle obtenue avec un modèle basé sur les proportions (83% contre 75%).

5.6 Discussion

Les estimations effectués nous ont montré que la meilleure spécification est celle d'un modèle probit ordonné ou la variable d'intérêt est la part d'électricité substituée. Parmi les variables disponibles il y a le type de tarif et, dans une moindre mesure la consommation, qui expliquent l'effort d'adhésion des client de la bourse.

6 Conclusions

6.1 Les souscriptions à la bourse solaire

Nous nous sentons obligé à ce stade à revenir à notre problème : l'explication du *processus d'adhésion* à la bourse solaire des SIG. Il est nécessaire de remarquer que les clients ont fait leur choix d'une manière personnelle et leurs motivations peuvent être les plus variées, telles l'orientation politique, l'influence du contexte familial, l'activité professionnelle ou autre. Nous ne prétendons pouvoir détailler cela; avec les données administratives et comptables à disposition nous avons essayé de proposer des modèles pour expliquer au mieux combien de tranches d'électricité d'origine photovoltaïque l'individu *i* a acheté.

Les données ont été traitées comme transversales. En réalité il s'agit d'un processus dynamique - démarré en mai 2000 – et qui peut être temporellement découpé en année d'adhésion. Pour des raisons de numerosité de l'échantillon nous avons adopté des modèles statiques, n'ayant pas assez de clients en 2001, ni de variations des quantités souscrites. Cependant, le temps a une importance puisque le choix d'adhérer (et de combien) peut être motivé par des facteurs situables dans le temps.

Nous devons également remarquer que des études ultérieures devraient :

- inclure la dimension temporelle en employant des modèles pour données de panel,
- compléter l'information par des questionnaires à envoyer à un échantillon de la clientèle (politique, présence d'enfants, etc.).

6.2 Spécificité des données traitées

Les derniers signaux du marché de la bourse solaire (étude de satisfaction, entretien personnel) indiquent que les souscriptions ralentissent. Ce phénomène n'est probablement pas imputable aux faibles quantités livrées en 2000 et 2001 (23 et 21% respectivement), mais plutôt à la fin de l'effet de la campagne d'information; en plus en 2002 les SIG pourront satisfaire la demande (et les adhérents) avec la station du Parking de l'Etoile qui comble le déficit d'offre.

Un autre changement concerne l'intérêt grandissant de la part d'entreprises, en concertation avec des communes, pour la réalisation de nombreux projets. Quels sont les motivations de ces acteurs ? Nous pouvons en identifier quelques unes:

- L'obligation du rachat par les SIG ;
- L'image au sens très large (nous y reviendrons);
- Pour l'entreprise privée la certitude du rachat (1e CHF/kWh par les SIG) d'un côté et de l'autre l'attractivité de l'accès au crédit et subventions grâce aux projets sponsorisés par des communes ;
- L'économie d'échelle, (a) en termes d'amortissement de la dette, car par ex. 50.000 CHF sur 20 ans pour un individu sont plus lourds que 500.000 CHF pour une commune (cela valant aussi pour les taux d'intérêt et les subventions) ; (b) en termes réels, puisque certains composants d'une centrale PV, tels un onduleur, sont quasiment indépendantes de la taille et coûtent donc proportionnellement beaucoup plus pour une petite installation.

6.3 Le courant vert en Suisse

Le courant vert est considéré comme la clé du succès dans un marché libéralisé de l'électricité¹. Il est en effet vraisemblable que les consommateurs qui peuvent choisir l'origine de l'électricité consommée vont progressivement substituer le nucléaire et fossile pour de l'électricité d'origine photovoltaïque et du petit hydraulique.

Nous rappelons que des études ultérieures devraient mener une analyse dynamique pour inclure l'effet de la campagne d'information de SIG et le mécanisme de choix pour le courant vert en général ; nous croyons également souhaitable que l'information sur la disponibilité à payer pour du solaire et du courant vert obtenue suite à l'Étude de satisfaction des clients de la bourse solaire², soit intégrée dans des études ultérieures.

¹ P. Schaub information personnelle, 25 février 2002.

² Le document est fourni en annexe (confidentiel).