

Evaluation d'impact de l'accompagnement des
demandeurs d'emploi par les Opérateurs Privés de
Placement et le programme Cap Vers l'Entreprise
Annexe technique

Luc Behaghel*, Bruno Crépon[†] et Marc Gurgand[‡]

Septembre 2009

*Ecole d'Economie de Paris (Inra), Crest et J-PAL

[†]Crest et J-PAL

[‡]Ecole d'Economie de Paris, Crest et J-PAL

Annexe A. Suivi du protocole expérimental

Les grandes étapes du protocole ont été présentées dans le corps du rapport : repérer un ensemble de demandeurs d'emploi éligibles pour un accompagnement renforcé, effectuer un tirage au sort entre trois modalités d'accompagnement – accompagnement classique de l'ANPE, accompagnement renforcé par un OPP, accompagnement renforcé par le programme CVE –, orienter en conséquence chaque demandeur, puis suivre les trajectoires des trois groupes expérimentaux ainsi constitués. Cette annexe rend compte des points saillants dans l'exécution de ce protocole. Elle présente l'adaptation du protocole d'orientation des demandeurs d'emploi aux contraintes locales (en particulier pour tenir compte des capacités locales d'accueil des différents dispositifs) ; du fonctionnement de l'outil de constitution des cohortes (OCC) et de la montée en charge des différents programmes ; enfin, elle présente les difficultés rencontrées sur certaines entrées en programme contraires au protocole et leurs conséquences sur l'évaluation.

Ajustement du protocole aux capacités d'accueil locales des différents programmes

Un enjeu, important tout particulièrement en début d'expérimentation, était que le nombre de demandeurs d'emploi orientés vers les différents programmes soit compatible avec les capacités d'accueil de chacun des dispositifs et avec les objectifs fixés contractuellement aux OPP. Cela dépendait tout d'abord du nombre de demandeurs d'emploi éligibles (c'est-à-dire pouvant être orientés vers l'un des trois programmes), mais aussi de la clé de répartition adoptée dans l'envoi vers ces trois programmes. Dans la plupart des cas, la principale contrainte était que soit orienté un nombre suffisant de demandeurs vers les OPP. Mais un dispositif (CVE ou OPP) pouvait se trouver ponctuellement saturé dans un endroit donné, auquel cas seules des orientations vers un suivi classique étaient possibles. Les probabilités d'envoi vers les différents programmes utilisées par OCC ont par conséquent été fixées ALE par ALE (et séparément pour chacune des 3 catégories de population). Elles ont été ajustées mois par mois en fonction de l'évolution de la situation. Au bout de quelques mois, les probabilités d'envoi vers les OPP ont été fortement rehaussées pour faire face aux engagements d'envois pris par l'Unédic. Pour le flux indemnisable et dans les zones où intervenaient les OPP, les probabilités d'envoi vers les OPP étaient élevées, ce qui limitait

fortement les effectifs qui pouvaient être orientés vers CVE ou le parcours classique. Cela implique que dans ces zones, la précision statistique tend à être plus faible. Ainsi, la probabilité de tirage vers OPP était de 0.85 dans 90% des ALE ; la probabilité de tirage vers CVE était alors de 0.06 ou 0.08, le complément étant pour le parcours classique. Dans les zones où les OPP n'étaient pas présents, et sur le flux non indemnisable et sur le stock, la probabilité la plus fréquente était 0.5 vers CVE et 0.5 vers classique, mais dans de nombreux cas, notamment dans le stock, la probabilité d'envoi vers CVE était inférieure, en raison d'un déséquilibre entre le nombre de bénéficiaires potentiels et les capacités d'accueil de CVE.

Ces probabilités déséquilibrées étaient une condition pour que l'évaluation perturbe au minimum de déploiement des programmes. Il est nécessaire de redresser les échantillons des différents groupes pour tenir compte des effets de composition induits par ces probabilités déséquilibrées (certaines ALE sont surreprésentées dans le groupe orienté OPP, d'autres sont sous-représentées ; idem pour l'accompagnement CVE ou le suivi classique) ; les estimateurs utilisés pour ce faire sont présentées dans l'annexe C. Il reste que les tailles déséquilibrées des différents réduisent la précision des estimations qu'il aurait été possible d'obtenir si les groupes avaient été ventilés également (un tiers orientés CVE, un tiers orientés OPP, un tiers orientés classique).

Par ailleurs, un grand nombre de demandeurs éligibles sont entrés dans les dispositifs sans être orientés par l'OCC, surtout en début d'année et dans le flux indemnisable, en raison de la nécessité d'alimenter rapidement les OPP. Ces personnes sont hors expérimentation et cela constitue une perte d'effectif, donc de précision statistique. Mais cela n'entame pas la capacité de l'évaluation à estimer l'effet des programmes sur les bénéficiaires passés par OCC.

Montée en charge du dispositif

La constitution aléatoire des différents groupes expérimentaux s'est effectuée sur 15 mois, de janvier 2007 à mars 2008. Elle a eu lieu dans 393 agences, dans 16 régions au total. Parmi elles, 67 n'ont prescrit que des accompagnements classiques ou CVE, 91 n'ont prescrit que des accompagnements classiques ou OPP et les 235 restantes ont disposé des trois formules d'accompagnement. Au final, ce sont 219 033 demandeurs d'emploi qui rentrent

	total	Répartition par population			Répartition par orientation		
		FI	FNI	STOCK	classique	CVE	OPP
T1 2007	52693	9555	10104	33034	34134	12219	6340
T2 2007	60454	22283	15703	22468	24197	20376	15881
T3 2007	45740	18858	14091	12791	17856	15199	12685
T4 2007	38470	15543	13854	9073	14167	13749	10554
T1 2008	21676	2151	10189	9336	10772	10904	0
total	219033	68390	63941	86702	101126	72447	45460

TAB. 1 – Montée en charge de l’Outil de Constitution des Cohortes

ainsi dans l’évaluation.¹

La montée en charge de l’OCC a été très rapide. Comme l’indique le tableau 1, dès le premier trimestre 2007, l’outil a orienté plus de 52 000 demandeurs d’emploi. Mais il a été d’abord massivement utilisé dans le stock et en requêtes par listes, tandis que la montée en charge dans le flux, indemnisable et non indemnisable, a été progressive. A fin mars 2008, 219 033 demandeurs d’emploi sont passé par OCC, dont 86 702 dans le stock, 63 941 dans le flux non indemnisable et 68 390 dans le flux indemnisable. Au total, 45 460 ont été orientés vers les OPP, 72 447 vers CVE et 101 126 vers l’accompagnement classique.

Entrées en programme non conformes au protocole

Autant il était prévu que des demandeurs d’emploi orientés vers les programmes CVE et OPP n’entrent finalement pas dans ces programmes,² autant le protocole excluait en principe que des demandeurs d’emploi entrent dans un programme d’accompagnement renforcé vers lequel ils n’ont pas été orientés.

En pratique, de telles entrées ont été observées, dans des proportions

¹On exclut de ce décompte les demandeurs d’emploi qui ne sont pas retenus dans le fichier d’analyse, ainsi que les demandeurs d’emploi orientés vers les programmes CVE et OPP sans passer par OCC qui, par définition, ne rentrent pas dans le champs de l’évaluation. Voir l’annexe B pour plus de détail sur cette sélection de l’échantillon d’analyse.

²Voir section 2 du rapport.

faibles mais non négligeables. Le tableau 2 du rapport montre ainsi en particulier que des demandeurs d'emploi orientés vers un suivi classique ont été pris en charge par un OPP dans une proportion de l'ordre de 3% en moyenne. D'autres mouvements de cette nature (demandeurs orientés CVE qui entrent en OPP, orientés classique qui entrent en CVE ou encore orientés CVE qui entrent en OPP) existent mais restent très rares. Les entrées en OPP de demandeurs orientés classique résultent de la forte pression exercée sur les réseaux par la nécessité d'alimenter les opérateurs privés. Il est par exemple arrivé que des listes constituées sans avoir recours à l'OCC aient été transmises aux OPP, alors qu'au même moment les demandeurs étaient orientés classique par l'OCC au cours de leur entretien de PPAE. Cette situation complique encore l'évaluation car le groupe témoin (les personnes orientées classique) contient des personnes prises en charge par les OPP qui sont, à leur tour, particulières. Ces situations ont été très fréquentes en Alsace et, dans une moindre mesure, dans le Centre, en Languedoc-Roussillon et en Provence-Alpes Côte d'Azur, surtout en milieu d'année, lorsque les tensions sur les OPP se sont fait le plus fortement sentir. A l'inverse, l'entrée en OPP des demandeurs orientés classique est restée exceptionnelle en Rhône-Alpes.

Ces situations sont potentiellement problématiques pour l'évaluation. Leur traitement économétrique est détaillé dans l'annexe C. Elles sont davantage problématiques pour le programme CVE que pour les OPP. En effet le principe de base de l'évaluation est de dire que pour chaque programme, il y a deux populations : le groupe des individus potentiellement éligibles et le groupe de contrôle. S'il n'y avait qu'un seul programme, le fait que des individus affectés au groupe de contrôle entrent dans le programme ne serait pas un problème. Cela réduirait au plus un peu la capacité de détection de l'évaluation et la portée des résultats. On aurait deux populations identiques à l'origine et deux programmes : le programme classique et le programme étudié, auxquels chacune des deux populations participerait avec des taux différents. Il y a là assez d'information pour mesurer la plus-value apportée par le programme. C'est ce qu'explique en détail la section 2 du rapport. Cette section explique aussi que ce que l'on mesure alors est la plus-value apportée par le programme sur une sous-population particulière que l'on définit à cette occasion et qui est fonction des taux d'entrée dans le programme dans les deux populations.

Le problème provient ici du fait que l'on a trois programmes, ce qui rend nécessaire des schémas d'entrée plus contraints : pas d'entrées des personnes

orientées vers le programme classique dans les programme OPP et CVE, et, pour les groupes orientés CVE et OPP, entrées seulement dans le programme vers lequel on est orienté ou, à défaut, le programme classique. On peut néanmoins tirer parti du fait que les écarts ont toujours pour l'essentiel été des entrées chez les OPP. En effet, si on compare le groupe affecté au parcours classique et le groupe affecté aux OPP, le groupe affecté aux OPP entre soit dans le programme classique soit dans le programme OPP. Le groupe affecté au parcours classique entre soit dans le programme classique, soit (pour 3% d'entre eux) dans le programme OPP. Conformément à ce que l'on vient d'expliquer, les problèmes occasionnés sont limités : pour ces deux populations on a en effet deux programmes, OPP ou classique. On veut néanmoins se prémunir d'une trop grande hétérogénéité de la population sur laquelle le programme est évalué. C'est pourquoi il a été décidé de ne pas inclure l'Alsace dans l'évaluation. Pour l'Alsace on a en effet un taux d'envoi des demandeurs d'emploi affectés au parcours classique chez les OPP de 33%, ce qui est considérable.

Pour CVE les problèmes sont potentiellement plus importants. En effet, si on considère les populations affectées à CVE et affectées au parcours classique, on a bien deux populations, mais ces deux populations sont susceptibles d'entrer dans trois programmes : le programme CVE pour les individus affectés au parcours CVE, le programme classique pour les individus affectés au programme classique ou les individus affectés au programme CVE mais n'y entrant pas, mais aussi le programme OPP pour les individus orientés vers le programme classique ou le programme CVE, mais réorientés vers le programme OPP et y entrant. On sent donc bien qu'il y a là un manque d'information. Il est néanmoins possible de le résoudre, à condition de faire l'évaluation des programmes OPP et CVE dans le flux indemnisable de façon conjointe et de supposer que l'effet du programme OPP est en moyenne le même dans les populations entrant dans ce programme. Pour limiter l'impact de ce problème, nous avons éliminé, parmi les régions dans lesquelles les programmes OPP et CVE ont été développés conjointement, la région Midi-Pyrénées pour laquelle 4% des demandeurs d'emplois affectés au programme classique s'étaient retrouvés chez les OPP.

Annexe B : Données et constitution du fichier d'analyse

Enquête téléphonique

Le FH (fichier historique de l'ANPE) sert de base à la mesure de la sortie vers l'emploi car il constitue une première source exhaustive de suivi de plus de 200 000 demandeurs d'emploi (DE) de l'expérience. Néanmoins, il ne suffit pas à mesurer la sortie vers l'emploi de tous, dans la mesure où de nombreuses demandes d'emploi sont annulées sans que le motif soit connu : il peut s'agir d'une sortie vers l'emploi aussi bien que d'une sortie vers l'inactivité. Sur le modèle de "l'Enquête Sortants" mise en oeuvre chaque trimestre par l'ANPE, on a donc procédé à une enquête téléphonique auprès d'un sous-échantillon de demandeurs d'emploi sortis du FH sans qu'on sache s'il s'agissait ou non d'une reprise d'emploi. Cette enquête a été réalisée par un prestataire extérieur, l'institut de sondage Louis Harris 2 (LH2), déjà en charge de l'Enquête Sortants. Chaque mois, une nouvelle vague d'enquête a été effectuée auprès de demandeurs d'emploi récemment sortis du FH. Le pourcentage de demandeurs d'emploi interrogé diffère selon les groupes considérés (DE issus du stock, du flux, et tirés au sort CVE, OPP ou classique) : il a été choisi de façon à maximiser la précision de l'évaluation.

Le questionnaire de l'enquête LH2 était le suivant :

Bonjour suite au courrier de l'ANPE, je suis (Prénom NOM) et je travaille en collaboration de l'ANPE. Nous réalisons actuellement une étude sur le devenir des personnes ayant été inscrites à l'ANPE. Cette étude est très courte. Pouvez-vous m'accorder quelques instants ?

Question 1. *Au mois de (mois) dernier, vous n'avez pas renouvelé votre inscription à l'ANPE. Pour quelle raison ?* Les modalités suivantes étaient possibles : *Reprise d'emploi (y.c. travail à son compte, création d'entreprise)*

- 1. Formation, reprise d'études, stage*
- 2. Service militaire*
- 3. Maladie*
- 4. Congé maternité*
- 5. Congé parental*
- 6. Retraite*

7. *Déplacement ; vacances*
8. *N'est plus indemnisé (aller en Q2)*
9. *Ne cherche plus d'emploi (aller en Q2)*
10. *Ne voit plus l'intérêt d'être inscrit à l'ANPE ; ne souhaitait plus être inscrit à l'ANPE ; découragé (aller en Q2)*
11. *Problème de carte (avait perdu sa carte, a oublié de renvoyer...) (aller en Q2)*
12. *A oublié de téléphoner, ne savait pas qu'il fallait téléphoner (aller en Q2)*
13. *Problème de téléactualisation (par téléphone, Internet, borne "Unidialogue"), ne comprend pas (aller en Q2)*
14. *Radiations administratives (aller en Q2)*
15. *Autres motifs (préciser) (aller en Q2)*

Question 2. *En (mois), étiez-vous en Contrat Aidé ? Consigne enquêteur : si le demandeur ne sait pas s'il est en contrat aidé ou non, citer : CAE : Contrat d'Accompagnement dans l'Emploi, CIE : Contrat Initiative Emploi, CI-RMA : Contrat Insertion Revenu Minimum d'Activité, CAV : Contrat d'Avenir, Contrat d'apprentissage, Contrat de professionnalisation, CJE : Contrat Jeune en Entreprise*

Question 3. *En (mois), avez-vous travaillé en intérim ?*

Question 4. *En (mois), avez-vous pris un emploi même de courte durée ?
Je vous remercie d'avoir répondu à ces questions. Ces données permettront à l'ANPE d'évaluer la part des demandeurs sortis pour reprendre un emploi.*

Le tableau 2 récapitule les grandes données de l'enquête pour un horizon de sortie vers l'emploi de six mois. Il donne différents effectifs et ratios. Dans la première ligne, par exemple, on voit que le nombre d'individus du flux indemnisable (zones avec les deux programmes CVE et OPP) orientés vers un OPP est de 36 027. 12 864 sortent des listes de l'ANPE dans les 6 mois qui suivent. Dans les fichiers administratifs, le motif de sortie n'est pas renseigné ou ne permet pas de savoir s'il y a eu mise en emploi pour 6 274 d'entre eux (soit 49%). On cherche à interroger un échantillon de 32% de ces 6 274 demandeurs, soit 1 980 individus. Le taux de réponses exploitables est de 50%, soit 986 interviews. Au final, on connaît ainsi le motif de sortie de 76% des 12 864 demandeurs sortis des listes ; en ajoutant ceux qui sont restés sur

les listes, le taux de situations connues est de 91% des 36 027 demandeurs initiaux.

Le tableau appelle plusieurs commentaires : tout d'abord, le taux de situations connues est élevé, quelle que soit la population considérée : il est de l'ordre de 90%. Il est indépendant, dans une population donnée, de l'orientation OPP, CVE ou classique (dans les zones communes du flux indemnisable, il est ainsi toujours de 91%). Les principales différences apparaissent dans les comparaisons entre populations. En particulier, les sorties de listes pour motif inconnu sont très fréquentes dans le flux non indemnisable, ce qu'on peut expliquer par le fait que pour un demandeur d'emploi non indemnisé, informer l'ANPE n'a pas d'incidence en termes d'indemnisation par l'assurance chômage.

En résumé, l'effort considérable mené avec l'enquête téléphonique a permis d'obtenir des taux de réponse très élevés, et comparables dans les groupes de traitement et de contrôle. Ces taux de réponse élevés sont indispensables pour tirer tout le parti de l'orientation aléatoire par OCC. Pris ensemble, orientation aléatoire et taux de réponse élevés justifient l'hypothèse que les échantillons de répondants dans les groupes de contrôle et de traitement sont comparables en tout, excepté le traitement. On traite alors la non-réponse résiduelle en la considérant comme aléatoire (*missing at random*) : les non-répondants de chaque groupe (contrôle et traitement) sont représentés par les répondants du même groupe, en utilisant les poids appropriés. Comme cela se fait souvent pour redresser la non-réponse (dans des enquêtes où la non-réponse est en général beaucoup plus élevée), il est également possible de calculer des poids spécifiques par groupes sociodémographiques. Cela a été fait pour vérifier la robustesse des résultats ; il s'avère que ces derniers sont pratiquement inchangés lorsqu'on procède de la sorte.

Constitution du fichier d'analyse

Tous les fichiers utilisent l'identifiant banalisé du Fichier Historique Statistique de l'ANPE (FHS), en date statistique de mars 2009. Le FHS de mars 2009 est en effet le premier à permettre de suivre les 15 cohortes de l'expérimentation pendant les 12 mois qui suivent l'orientation aléatoire par OCC. L'identifiant banalisé du FHS (variable *ident*) est constitué d'un numéro de région à trois chiffres, et d'un numéro individuel à huit chiffres. Il permet en principe de suivre un même demandeur d'emploi au cours de ses différents épisodes de chômage.

Les fichiers source sont les suivants :

1. **res_occ_FH2009T1**. Ce fichier historise les orientations aléatoires effectuées par l'outil OCC entre le 1er janvier 2007 et le 31 mars 2008. Outre l'orientation (CVE, OPP ou CLA), le fichier indique si le DE a été identifié comme faisant partie du flux indemnisable, du flux non indemnisable ou du stock. La date d'orientation et les probabilités d'orientation vers les différents groupes expérimentaux en vigueur dans l'ALE au moment de l'orientation considérée sont également conservées. Ce fichier comporte 10 325 observations. Ont été éliminés de ce fichier 300 doublons (observations avec le même identifiant) et 188 observations hors champ (erreur d'un conseiller qui a fait passer l'ensemble des DE de son portefeuille par OCC). Ont également été éliminés les observations antérieures au 1er janvier 2007 et postérieures au 31 mars 2008, ces dates marquant le début et la fin de la constitution des cohortes de l'évaluation.
2. **Fichiers du FHS datés 2009 T1**. Il s'agit d'un ensemble de fichiers : un par région pour chacun des différents segments du FHS. On utilise ici deux segments : le segment *de*, noté *de_XXX*, et le segment *e0*, noté *e0_XXX* (XXX correspond au numéro de la région). Le segment *de* comporte les caractéristiques socio-démographiques du DE et l'historique des périodes de demande d'emploi recensées (à partir de 1999). Le segment *e0* comporte l'information sur l'activité réduite.
3. **extraction_sorties_2009T1**. Ce fichier empile les extractions effectuées par l'ANPE, à un rythme à peu près mensuel, pour identifier les DE sortis du FH. Cette extraction a été effectuée par les services statistiques de l'ANPE dans les fichiers administratifs, sur un champ constitué de l'ensemble des DE passés par OCC. Une observation correspond à une

sortie des listes du chômage ; un même DE peut donc avoir plusieurs observations. Une sortie est définie par l'annulation de la demande d'emploi, pour quelque motif que ce soit, lorsque aucune nouvelle demande n'est rouverte dans le mois qui suit. On capte ainsi les sorties durables des listes du chômage. Pour chaque sortie, les principales informations sont la date d'annulation de la demande (variable *maxann*) et le motif d'annulation (variable *motann*). Ces variables sont indexées par 6 chiffres correspondant à l'année et au mois de l'annulation de la demande d'emploi.

4. **enq_sortants_2009T1**. Ce fichier comporte les informations issues de l'enquête effectuée par LH2 auprès des DE dont le motif de sortie était inconnu ou imprécis. Il comporte des variables de gestion de l'enquête. En particulier, afin de limiter les coûts d'enquête, seul un échantillon aléatoire de DE sortis pour motif inconnu ont été sélectionnés pour être enquêtés. L'indicatrice *sortieenquete* vaut alors 1 ; lorsque elle vaut 0, cela signifie que le motif de sortie était inconnu mais que ce DE est représenté par d'autres DE dans l'enquête. Parmi les DE tirés pour être interrogés, tous n'ont pas répondu ou n'ont pu être joints. La variable *statutenquete* répertorie cette information. Le questionnaire de l'enquête LH2 figure ci-dessus. La variable *q1* correspond à la première question, utilisée pour identifier les sorties vers l'emploi.
5. **res_cve_opp_FH2009T1**. Ce fichier comporte l'information fournie par l'ANPE sur les entrées dans les programmes CVE et OPP. Seule l'information sur les entrées dans le programme CVE est utilisée dans ce rapport. Pour le programme OPP, l'information tirée des fichiers de l'Unédic, plus complète, est privilégiée. La variable *cve_ent* indique que le DE a signé la charte d'entrée dans le programme CVE et la date d'entrée est consignée dans *datent_cve_opp*.
6. **ori_opp_FH2009T1**. Ce fichier comporte les informations fournies par les OPP et compilées par l'Unédic sur l'ensemble des DE qui leur ont été signalés. En particulier, il permet d'identifier l'entrée dans le programme OPP par la signature d'un formulaire (variable *signature*) ainsi que la date de signature.

Le programme *constitution_base_CVE_OPP.do* comporte en langage Stata l'ensemble des opérations effectuées pour créer la base de données utilisée dans ce rapport. Ce programme est disponible sur demande ; il est largement annoté. On commente ici seulement ses principales étapes.

Etape 1 : Appariement des différents fichiers. Le fichier de référence pour cette étape est *res_occ_FH2009T1*. Toutes les observations apportées par d'autres fichiers mais absentes du fichier tiré d'OCC sortent du champ de l'évaluation ; elles sont donc éliminées du fichier apparié. Le fichier *extraction_sorties_2009T1* comporte l'ensemble des sorties des listes. Dans la plupart des analyses, en cas de sortie multiple après l'orientation par OCC, seule la première est retenue. De même, le segment *de* du FHS comporte plusieurs demandes par individu. Dans la plupart des analyses, on s'intéresse seulement à la demande en cours au moment du passage par OCC. C'est celle-là qu'on retient pour l'appariement.

Etape 2 : Création de variables. La création des variables est commentée dans le programme. On note juste quelques points particuliers.

- Pour l'analyse du flux indemnisable, il est important d'identifier les ALE où co-existaient les programmes CVE et OPP et celles où seul l'un des deux programmes était en vigueur. On utilise pour cela les probabilités moyennes d'orientation vers chacun de ces programmes. Une probabilité nulle d'orientation vers l'un des programmes au cours de l'évaluation signifie que ce programme n'était pas en vigueur. Dans la variable *CATEGORIE3*, les observations concernées sont donc classées comme observations avec un programme seulement.
- Les motifs de sortie de la variable *motann* issue du fichier *extraction_sorties_2009T1* sont trop détaillés ; ils sont recodés dans une nomenclature plus ramassée à 16 postes. La variable *motifFHS* ainsi obtenue peut différer légèrement de la variable *motann* du segment *de* du FHS de mars 2009, dans la mesure où ce dernier fichier a été actualisé par rapport au fichier *extraction_sorties_2009T1* et certains motifs d'annulation mis à jour. De même, certaines annulations de demande ont été annulées ; et certaines nouvelles annulations ont pu apparaître.
- Les variables de résultat (en particulier, la variable de sortie vers l'emploi *EMPLOI*) sont spécifiques à un horizon donné : sortie à 6 mois, à 9 mois, etc. A ce stade, les variables créées le sont à l'horizon le plus éloigné disponible. La restriction à un horizon spécifique est faite ultérieurement, dans les programmes d'analyse des données.

Etape 3 : Sélection de l'échantillon. 265 266 demandeurs d'emploi sont passés par OCC entre le 1er janvier 2007 et le 31 mars 2008 (après éliminations de 300 paires de doublons et d'une observation incohérente).

Néanmoins, l'échantillon mobilisé pour les analyses comporte au maximum 219 033 observations. Des observations doivent en effet être éliminées de l'échantillon d'analyse pour différentes raisons :

1. Pour 10 325 DE, le statut d'éligibilité à l'indemnisation par l'assurance chômage (condition de l'éligibilité au programme OPP) n'a pas été connu à temps pour permettre l'orientation par OCC.
2. Les régions Midi-Pyrénées et Alsace, où le protocole n'a pas été suivi de façon satisfaisante, sont exclues, soit 9 447 DE en moins. Les résultats sont robustes à l'intégration de ces DE dans l'analyse.
3. 8 840 DE sont exclus dans la mesure où ils n'avaient pas de demande d'emploi ouvertes (et étaient donc déjà sortis des listes du chômage) au moment de leur passage par OCC. Il s'agit sans doute d'orientations sur des listes qui n'étaient pas à jour (ou qui ont été mises à jour de manière rétroactive).
4. Sont également exclus 11 330 DE orientés par OCC au 1er trimestre 2008 dans des zones où le programme OPP était en vigueur. Il n'a pas été possible de savoir lesquels de ces DE étaient effectivement entrés en accompagnement OPP. Dès lors, l'analyse de l'impact des programmes CVE et OPP n'était pas possible sur ces zones à cette période.
5. 308 anomalies sont également relevées : des DE qui sont notés comme entrant à la fois dans l'accompagnement CVE et dans l'accompagnement OPP, alors que ces deux programmes étaient exclusifs. Pour 144 observations supplémentaires, il existe un doute sur la date de passage par OCC.
6. Enfin, sont éliminées des observations correspondant à des périodes où OCC a été utilisé bien que le protocole expérimental ne soit pas encore effectif ou ait été suspendu : maintenir l'usage d'OCC permettait de ne pas interrompre le protocole. Dans ces cas, les demandeurs d'emploi étaient orientés avec une probabilité égale à 1 dans l'un des trois groupes, ou à 0 dans l'un des groupes. Néanmoins, cela implique qu'une comparaison entre groupes de contrôle et de traitement n'est alors pas possible. Il faut donc exclure ces observations de l'analyse. Elles sont au nombre de 2877 dans le flux indemnisable, 1710 dans le flux non indemnisable et 1254 dans le stock.

Ces différents filtres sur les données permettent de constituer un échantillon de référence parfaitement cohérent. Cependant, on a vérifié que les résultats demeurent qualitativement inchangés en ôtant l'un ou l'autre filtre.

Cette annexe présente les paramètres estimés ainsi que les estimateurs utilisés. Ces éléments font référence aux travaux théoriques développés récemment en économétrie pour effectuer des évaluations d'impact de programmes.

Annexe C : Le cadre d'analyse de l'évaluation d'un programme

On s'intéresse à une variable d'output y , comme par exemple la sortie vers l'emploi à une date donnée. Pour définir la plus-value sur la variable y associée à la participation au programme, on introduit deux variables de résultat (ou output) potentiel : $y(0)$ et $y(1)$. Il s'agit des outputs individuels sous chacune des alternatives correspondant à la participation ou non au programme. $y(0)$ est l'output si on ne participe pas au programme, $y(1)$ est l'output lorsque l'on y participe. Ces deux outputs existent pour chaque individu indépendamment de la participation ou non au programme. On définit l'effet causal de la participation comme la différence entre ces deux outputs :

$$c = y(1) - y(0)$$

Cet effet causal ainsi défini a deux caractéristiques centrales :

1. Il est individuel. Il n'y a pas a priori d'homogénéité de l'effet du programme dans la population. Il y a une distribution de l'effet du programme dans la population. Pour certains l'effet est peut être important et pour d'autres plus faible.
2. L'effet causal est inobservable. On n'observe jamais que la situation sous l'une ou l'autre des alternatives de participation au programme. Pour les bénéficiaires, on observe $y(1)$ mais pas $y(0)$ et pour les non-bénéficiaires on observe $y(0)$ mais pas $y(1)$. La difficulté principale de l'évaluation consiste à reconstituer cette situation inobservée.

Les méthodes d'évaluation standards pour évaluer l'effet d'un programme reposent sur des hypothèses fortes nécessaires à la reconstitution de cette situation inobservée. La fiabilité des résultats se mesure de ce fait à l'aune de la vraisemblance de ces hypothèses.

A contrario, le cadre expérimental qui a été développé permet de reconstituer cette situation alternative sans faire d'hypothèses fortes. Il permet en

particulier d'identifier l'effet moyen du programme sur une sous-population des individus qui entrent dans le dispositif. Le cadre expérimental constitué est basé sur le fait que la population éligible initiale est ventilée au hasard en deux sous-populations repérées par une variable $z \in \{0, 1\}$. La population $z = 1$ se voit proposer d'entrer dans le programme. La population $z = 0$ ne reçoit pas cette incitation. Il s'agit par exemple de la ventilation des demandeurs d'emploi du parcours 3 du flux non indemnisable, et on les ventile en deux sous-population à l'aide de l'Outil de constitution des cohortes (OCC). La décision d'entrer dans le programme est notée $T \in \{0, 1\}$.

On introduit deux comportements potentiels d'entrée dans le dispositif associés aux valeurs de z . On a ainsi $T(1)$ et $T(0)$. $T(1)$ représente le comportement d'entrée dans le dispositif lorsque l'on est dans la cohorte affectée au programme. Comme la participation est libre, une partie des individus va accepter d'entrer, une autre n'entrera pas. C'est cette décision qui est retracée par la variable $T(1)$. Lorsque le tirage OCC donne $z = 0$, l'individu est affecté au groupe de contrôle et normalement n'entre pas dans le dispositif. C'est cette situation qui est retracée dans la variable $T(0)$. On a normalement $T(0) = 0$ pour tout le monde.

L'intérêt de la ventilation initiale au hasard est de produire deux sous-populations statistiquement identiques, c'est-à-dire deux sous-populations dans lesquelles n'importe quelle variable est distribuée de façon identique. La seule différence est liée au fait que l'exposition aux programmes a été différente. En particulier, l'orientation aléatoire implique la relation d'indépendance

$$y(0), y(1), T(0), T(1) \perp z$$

Estimation de l'effet du programme

L'estimateur de référence considéré est l'estimateur dit de Wald qui compare la situation moyenne des individus affectés au groupe $z = 1$ avec la situation moyenne des individus affectés au groupe $z = 0$ et normalisé pour tenir compte des différences d'entrée dans le dispositif entre les deux populations :

$$\hat{b}_W = \frac{\bar{y}^{z=1} - \bar{y}^{z=0}}{\bar{T}^{z=1} - \bar{T}^{z=0}}$$

Cet estimateur identifie le paramètre

$$\lim \hat{b}_W = \frac{E(y|z=1) - E(y|z=0)}{E(T|z=1) - E(T|z=0)}$$

Ce paramètre est aussi le paramètre identifié par le système de conditions d'orthogonalité :

$$E\left(\begin{pmatrix} 1 \\ z \end{pmatrix} (y - a - bT)\right) = 0 \quad (1)$$

Proposition 1 *Le paramètre b identifié par le système (1) s'exprime à partir de la distribution des outputs potentiels comme*

$$b = \frac{E(y(1) - y(0))(T(1) - T(0))}{E(T(1) - T(0))} \quad (2)$$

Lorsqu'il y a monotonie, c'est-à-dire lorsque pour tout individu on a $T(1) \geq T(0)$ ce paramètre identifie l'effet moyen du programme sur une sous-population particulière appelée compliers :

$$b = E(y(1) - y(0)) | (T(1) - T(0) = 1) \quad (3)$$

On trouve donc sans surprise que lorsqu'il y a monotonie, les conditions d'orthogonalité précédentes identifient le paramètre LATE : Local Average Treatment Effect.

Preuve de la proposition 1 voir Annexe - Preuve

Identification et estimation des contrefactuels

L'analyse distingue trois populations : la population dite des "never taker", celle des "compliers" et celle des "always taker". Les never taker représente la population des individus qui ne rentreront pas dans le programme, qu'on le leur propose ou non. La population des always takers est formée des individus qui rentrent dans le programme même lorsqu'on ne le leur propose pas. Enfin la population des compliers est la population qui se conforme au tirage : lorsqu'on leur propose de rentrer, ils rentrent, lorsqu'on ne le leur propose pas ils ne rentrent pas. Ces populations ne peuvent pas être identifiées directement par contre, on peut mesurer leur taille et aussi identifier la moyenne de certains outputs potentiels sur ces populations.

En particulier, il est possible d'identifier un quantité d'intérêt examinée dans ce rapport : la situation alternative des compliers en l'absence de programme. On est capable d'identifier l'effet du programme sur cette population, on est aussi capable d'identifier ce qu'aurait été la situation moyenne de cette population en l'absence du programme.

Proposition 2 *La proportion des never-taker P_{NT} , celle des always-taker P_{AT} et celle des compliers P_C sont identifiables. on peut également identifier $E(y(1)|AT)$, $E(y(0)|NT)$, $E(y(0)|C)$ et $E(y(1)|C)$. On a*

$$\begin{aligned}
P_{NT} &= P(T = 0|Z = 1) \\
P_{AT} &= P(T = 1|Z = 0) \\
P_C &= P(T = 1|Z = 1) - P(T = 1|Z = 0) \\
E(y(1)|AT) &= E(y|T = 1, Z = 0) \\
E(y(0)|NT) &= E(y|T = 0, Z = 1) \\
E(y(0)|C) &= \frac{E(y(1 - T)|Z = 0) - E(y(1 - T)|Z = 1)}{P(T = 1|Z = 1) - P(T = 1|Z = 0)} \\
E(y(1)|C) &= \frac{E(yT|Z = 1) - E(yT|Z = 0)}{P(T = 1|Z = 1) - P(T = 1|Z = 0)}
\end{aligned}$$

Dans le cas considéré pour lequel il n'est pas possible pour un individu affecté au parcours classique d'entrer dans le programme, on a nécessairement monotonie puisque $T(0) = 0$. Les compliers représentent simplement les individus décidant d'entrer dans le programme lorsqu'on le leur propose : $T(1) = 1$. On a donc les tailles des populations et les moyennes des outputs suivants :

$$\begin{aligned}
P_{NT} &= P(T = 0|Z = 1) \\
P_{AT} &= 0 \\
P_C &= P(T = 1|Z = 1) \\
E(y(0)|NT) &= E(y|T = 0, Z = 1) \\
E(y(0)|C) &= [E(y|Z = 0) - E(y(1 - T)|Z = 1)] / P(T = 1|Z = 1) \\
E(y(1)|C) &= E(y|Z = 1, T = 1)
\end{aligned}$$

Indépendance conditionnelle à des observables X_c

En pratique les probabilités d'envoi dans les programmes ont fluctué d'une ALE à l'autre et d'une période à l'autre. Cette fluctuation était nécessaire pour satisfaire les objectifs quantitatifs d'envoi aux OPP. Il en résulte qu'on n'a pas la condition d'indépendance précédente sur toute la population dans son ensemble, mais seulement conditionnellement aux zones et périodes de constance des probabilités d'envoi dans les programmes.

Dans cette partie on analyse la situation dans laquelle on a $y(0), y(1), T(0), T(1) \perp z | X_c$ et non plus $y(0), y(1), T(0), T(1) \perp z$. Plusieurs estimateurs peuvent être considérés.

On présente trois estimateurs dérivés de l'estimateur de Wald. Le premier a pour caractéristique de pondérer les observations par l'inverse de la probabilité d'être affecté à son groupe d'affectation. Le second et le troisième introduisent les variables de conditionnement comme variables de contrôle et considèrent différents jeux possibles d'instruments. Ces estimateurs identifient tous des paramètres ayant un sens causal. Il s'agit de moyennes des effets du programme, la différence entre les estimateurs provenant du système de poids retenus. Le premier des estimateurs que l'on considère est celui qui finalement identifie le paramètre naturel auquel on s'intéresse, ici : l'effet du programme sur ses bénéficiaires. C'est cet estimateur qui est utilisé dans ce rapport. On a le résultat suivant :

Proposition 3 *Lorsque l'on a indépendance conditionnelle à des observables $y(0), y(1), T(0), T(1) \perp z | X_c$. Les conditions d'orthogonalité (1), convenablement pondérées :*

$$E \left(w(X_c, z) \begin{pmatrix} 1 \\ z \end{pmatrix} (y - a - bT) \right) = 0 \quad (4)$$

avec

$$w(X_c, z) = \left(\frac{C_1}{P(z=1|X_c)} \right)^z \left(\frac{C_0}{1 - P(z=1|X_c)} \right)^{1-z} \quad (5)$$

identifient le même paramètre que précédemment (où C_1 et C_0 sont des constantes positives).

$$b = \frac{E(y(1) - y(0))(T(1) - T(0))}{E(T(1) - T(0))} \quad (6)$$

Preuve de la proposition 3 voir *Annexe*

On voit donc que finalement rien n'est changé. Il suffit de pondérer les observations par l'inverse de la probabilité d'être affecté au traitement.

Forme de l'estimateur et comparaison avec l'estimateur à variables de contrôle

L'estimateur obtenu par solution du système (4) généralise l'estimateur de Wald. On trouve facilement son expression :

$$\widehat{b}_W = \left(\frac{\overline{y/P(z=1|X_c)}^{z=1}}{\overline{1/P(z=1|X_c)}^{z=1}} - \frac{\overline{y/P(z=0|X_c)}^{z=0}}{\overline{1/P(z=0|X_c)}^{z=0}} \right) / \left(\frac{\overline{T/P(z=1|X_c)}^{z=1}}{\overline{1/P(z=1|X_c)}^{z=1}} - \frac{\overline{T/P(z=0|X_c)}^{z=0}}{\overline{1/P(z=0|X_c)}^{z=0}} \right) \quad (7)$$

Lorsque les variables X_c forment une partition de la population, les poids peuvent être estimés simplement par

$$\widehat{P}(z=1|X_c) = \frac{N_{X_c,1}}{N_{X_c}}$$

L'utilisation de ces poids empiriques conduit à une expression simple de l'estimateur

$$\widehat{b}_W = \frac{\sum_{X_c} \pi_{X_c} (\overline{y}^{X_c, z=1} - \overline{y}^{X_c, z=0})}{\sum_{X_c} \pi_{X_c} (\overline{T}^{X_c, z=1} - \overline{T}^{X_c, z=0})} \quad (8)$$

où $\pi_{X_c} = N_{X_c}/N$ mesure le poids de la cellule X_c .

Si on note $\widehat{b}_{W, X_c} = (\overline{y}^{X_c, z=1} - \overline{y}^{X_c, z=0}) / (\overline{T}^{X_c, z=1} - \overline{T}^{X_c, z=0})$, on voit que l'estimateur s'écrit comme une moyenne pondérée de l'effet du traitement sur chacune des cellules X_c . On a plus précisément

$$\widehat{b}_W = \sum_{X_c} \frac{\pi_{X_c} (\overline{T}^{X_c, z=1} - \overline{T}^{X_c, z=0})}{\sum_{X_c} \pi_{X_c} (\overline{T}^{X_c, z=1} - \overline{T}^{X_c, z=0})} \widehat{b}_{W, X_c} \quad (9)$$

Dans le cas où l'entrée dans le traitement est impossible pour $z=0$, on vérifie simplement que cette expression se simplifie en

$$\widehat{b}_W = \sum_{X_c} \pi_{X_c|T=1} \widehat{b}_{W, X_c} \quad (10)$$

Les poids sont simplement ceux de la loi de X_c conditionnellement à $T=1$. On retrouve le fait que le paramètre identifié est l'effet moyen du programme sur les bénéficiaires. Il est analogue ici à un estimateur qui serait obtenu sur chaque cellule X_c et pondéré ensuite par le poids de la cellule dans la population des bénéficiaires. La forme de la pondération est directement le reflet de l'absence de contrainte et de relation entre les effets du programme d'une population X_c à une autre.

Cet estimateur peut être comparé utilement à l'estimateur à variable instrumentale \hat{b}_c du modèle

$$y = X_c a_c + b_c T + u \quad (11)$$

avec pour variables instrumentales X_c et z , défini comme annulant les conditions d'orthogonalité :

$$E \left(\begin{pmatrix} X_c' \\ z \end{pmatrix} (y - X_c a_c - b_c T) \right) = 0 \quad (12)$$

On trouve aisément l'expression de cet estimateur :

$$\hat{b}_{VI1} = \frac{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} (\bar{y}^{X_c, z=1} - \bar{y}^{X_c, z=0})}{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} (\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0})} \quad (13)$$

$\hat{p}_{1,c} = N_{X_c,1}/N_{X_c}$ représente la probabilité empirique d'affectation au programme et $\hat{p}_{0,c} = N_{X_c,0}/N_{X_c}$ représente la probabilité empirique d'affectation au groupe de contrôle.

Cet estimateur se réécrit comme

$$\hat{b}_{VI1} = \sum_{X_c} \frac{\pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} (\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0})}{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} (\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0})} \hat{b}_{W, X_c} \quad (14)$$

Par rapport à l'estimateur précédent, cet estimateur donne donc d'autant plus de poids à une cellule que sa répartition entre affectation au traitement et aux contrôle est équilibrée.

Un autre estimateur possible est l'estimateur à variable instrumentale \hat{b}_c du modèle

$$y = X_c a_c + b_c T + u \quad (15)$$

avec pour variables instrumentales X_c et $z \times X_c$, défini comme annulant les conditions d'orthogonalité :

$$E \left(\begin{pmatrix} X'_c \\ zX'_c \end{pmatrix} (y - X_c a_c - b_c T) \right) = 0 \quad (16)$$

L'expression de cet estimateur est la suivante :

$$\hat{b}_{VI2} = \frac{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} \left(\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0} \right) \left(\bar{y}^{X_c, z=1} - \bar{y}^{X_c, z=0} \right)}{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} \left(\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0} \right)^2} \quad (17)$$

Il peut lui aussi se réécrire comme une moyenne pondérée des estimateurs \hat{b}_{W, X_c} obtenus cellule par cellule. On parvient à l'expression :

$$\hat{b}_c = \sum_{X_c} \frac{\pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} \left(\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0} \right)^2}{\sum_{X_c} \pi_{X_c} \hat{p}_{1,c} \hat{p}_{0,c} \left(\bar{T}^{X_c, z=1} - \bar{T}^{X_c, z=0} \right)^2} \hat{b}_{W, X_c} \quad (18)$$

On voit que par rapport aux autres estimateurs, on pondère d'avantage les cellules dans lesquelles on a une différence de take-up importante.

Compte tenu de l'expression de la variance de chacun des estimateurs, on peut montrer en outre qu'il s'écrit

$$\hat{b}_c = \sum_{X_c} \frac{1/V_{X_c}}{\sum_{X_c} 1/V_{X_c}} \hat{b}_{W, X_c} \quad (19)$$

où V_{X_c} est la variance de l'estimateur \hat{b}_{W, X_c} . On reconnaît l'estimateur des Moindres Carrés Asymptotiques du modèle imposant l'homogénéité du traitement

$$b_{W, X_c} = b \quad (20)$$

où b_{W, X_c} est la valeur du paramètre dans la cellule X_c .

Mesure des contrefactuels et taille de la population des compliers

On peut comme dans le cas précédent mesurer facilement la taille des populations des compliers et les outputs contrefactuels lorsqu'il y a indépendance conditionnelle à des observables. On a le résultat suivant :

Proposition 4 *Les formules obtenues dans le cas inconditionnel dans la proposition 2 se généralisent facilement au cas conditionnel :*

$$\begin{aligned}
P_{AT} &= E \left(T \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right) \\
P_{NT} &= E \left((1-T) \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right) \\
P_C &= E \left(T \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right) - E \left(T \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right) \\
E(y(1)|AT) &= \frac{E \left(yT \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right)}{E \left(T \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right)} \\
E(y(0)|NT) &= \frac{E \left(y(1-T) \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right)}{E \left((1-T) \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right)} \\
E(y(1)|C) &= \frac{E \left(yT \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right) - E \left(yT \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right)}{E \left(T \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right) - E \left(T \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right)} \\
E(y(0)|C) &= \frac{E \left(y(1-T) \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right) - E \left(y(1-T) \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right)}{E \left(T \frac{P(Z=1)}{P(Z=1|X_c)} | Z=1 \right) - E \left(T \frac{P(Z=0)}{P(Z=0|X_c)} | Z=0 \right)}
\end{aligned}$$

Mettre en évidence l'hétérogénéité de l'effet : introduction de variables de conditionnement X_0 et indépendance conditionnelle à X_c

On considère maintenant la question de l'introduction de caractéristiques observables des individus sous la forme de variables X_0 . Ceci est intéressant car cela permettra d'examiner l'hétérogénéité de l'effet. On considère la situation dans laquelle on a randomisation de z conditionnellement à des variables explicatives X_c . On a donc ici $X_0, y(0), y(1), T(0), T(1) \perp z | X_c$.

Compte tenu des hypothèses faites, on a aussi $y(0), y(1), T(0), T(1) \perp z | X_c, X_0$. La démarche d'estimation précédente reste donc valide conditionnellement à X_0 . Donc une première façon d'examiner l'hétérogénéité des effets est de conduire les analyses précédente pour des sous-populations définie comme

prenant un ensemble de valeurs donné : $X_0 \in \chi_0$ On considère les conditions d'orthogonalité

$$E \left(w(X_c, z) \begin{pmatrix} X'_0 \\ zX'_0 \end{pmatrix} (y - X_0a - TX_0b) \right) = 0 \quad (21)$$

avec $w(X_c, z_0) = 1/P(z = z_0 | X_c)$.

Proposition 5 *Sous l'hypothèse d'indépendance $X_0, y(0), y(1), T(0), T(1) \perp z | X_c$ et l'hypothèse de monotonie $T(1) \geq T(0)$, les conditions d'orthogonalité 21 identifient le vecteur des coefficients de la projection orthogonale de l'effet causal du programme sur les caractéristiques X_0 pour la population des compliers.*

$$b = E(X'_0 X_0 | T(1) - T(0) = 1)^{-1} E(X'_0 (y(1) - y(0)) | T(1) - T(0) = 1) \quad (22)$$

Preuve de la proposition 5 Voir Annexe - Preuve

Sondage, non-réponse et variance des estimateurs

Dans la situation où on a $X_0, y(0), y(1), T(0), T(1) \perp z | X_c$, les conditions d'orthogonalité que l'on considèrerait avec information complète sur les variables d'output seraient

$$E \left(w(X_c, z_0) \begin{pmatrix} X'_0 \\ zX'_0 \end{pmatrix} (y - X_0a - TX_0b) \right) = 0$$

avec $w(X_c, z_0) = 1/P(z = z_0 | X_c)$.

Ces conditions d'orthogonalité s'écrivent donc généralement

$$E(h(y, T, z, X_0, X_c, \theta)) = 0 \quad (23)$$

les observations sur la variable y ne sont disponibles que pour les individus sortis et ayant renseigné leur Déclaration de Situation Mensuelle. Pour ceux sortis sans renseigner cette déclaration mensuelle on effectue aussi un sondage et on recherche l'information manquante. D'une façon générale, on effectue un sondage conditionnellement à x_n , la variable indiquant la sortie sans réponse, z et T : $s(x_n, z, T)$, avec $E(s(x_n, z, T)) = \lambda(x_n, z, T)$. Pour $x_n = 0$ on a l'information, si bien que l'on a $s = 1$.

Les conditions d'orthogonalité que l'on considère sont donc

$$E \left(\frac{s(x_n, z, T)}{\lambda(x_n, z, T)} h(y, T, z, X_0, X_c, \theta) \right) = 0$$

et compte tenu du caractère aléatoire du sondage conditionnellement à x_n, z, T , c'est bien le même paramètre que l'on identifie.

Il y a aussi le problème des non réponses. Il y a donc une variable de réponse R indiquant si l'individu répond. On considère que cette variable de réponse est indépendante du résidu $h(y, T, z, X_0, X_c, \theta)$ conditionnellement à x_n, z, T, X_r , où X_r est un ensemble d'information. Sous cette hypothèse les conditions d'orthogonalité

$$E \left(\frac{r(x_n, z, T)}{\rho(x_n, z, T, X_r)} h(y, T, z, X_0, X_c, \theta) \right) = 0$$

avec $r(x_n, z, T) = s(x_n, z, T) R$ une variable indiquant si la variable d'output est renseignée, ce qui résulte du sondage qui est aléatoire et du comportement de réponse qui est supposé aléatoire et $\rho(x_n, z, T, X_r) = \lambda(x_n, z, T) \times (R = 1 | x_n, z, T, X_r)$ identifie le même paramètre que les conditions 23.

Finalement les conditions d'orthogonalité utilisée s'écrivent :

$$E \left(\omega \begin{pmatrix} X' \\ zX' \end{pmatrix} (y - Xa - TXb) \right) = 0$$

où ω est un poids défini par

$$\omega = \frac{s(x_n, z, T)}{\lambda(x_n, z, T)} \frac{R}{P(R = 1 | x_n, z, T, X_r)} \frac{1}{P(z = 1 | x_C)^z} \frac{1}{(1 - P(z = 1 | x_C))^{1-z}} \quad (24)$$

L'estimateur correspondant est défini comme celui annulant les conditions d'orthogonalités :

$$\overline{\omega \begin{pmatrix} X' \\ zX' \end{pmatrix} (y - X\hat{a} - TX\hat{b})} = 0$$

Ceci donne comme estimateur

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left[\begin{array}{cc} \overline{\omega X'_0 X_0} & \overline{\omega X'_0 X_0 T} \\ \overline{\omega z X'_0 X_0} & \overline{\omega z X'_0 X_0 T} \end{array} \right]^{-1} \begin{pmatrix} \overline{\omega X'_0 y} \\ \overline{\omega z X'_0 y} \end{pmatrix}$$

En notant

$$\hat{G} = \begin{bmatrix} \overline{\omega X'_0 X_0} & \overline{\omega X'_0 X_0 T} \\ \overline{\omega z X'_0 X_0} & \overline{\omega z X'_0 X_0 T} \end{bmatrix}$$

et

$$G = \begin{bmatrix} E(\omega X'_0 X_0) & E(\omega X'_0 X_0 T) \\ E(\omega z X'_0 X_0) & E(\omega z X'_0 X_0 T) \end{bmatrix}$$

On a

$$\sqrt{N} \left(\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix} \right) = \hat{G}^{-1} \begin{pmatrix} \overline{\omega X'_0 u} \\ \overline{\omega z X'_0 u} \end{pmatrix}$$

avec $u = y - a - bT$

La variance de l'estimateur est de ce fait

$$V \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = G^{-1} V G'^{-1}$$

avec

$$V = \begin{bmatrix} E(\omega^2 X'_0 X_0 u^2) & E(\omega^2 X'_0 X_0 T u^2) \\ E(\omega^2 z X'_0 X_0 u^2) & E(\omega^2 z X'_0 X_0 T u^2) \end{bmatrix}$$

qui peut être estimé par

$$\hat{V} = \begin{bmatrix} \overline{\omega^2 X'_0 X_0 \hat{u}^2} & \overline{\omega^2 X'_0 X_0 T \hat{u}^2} \\ \overline{\omega^2 z X'_0 X_0 \hat{u}^2} & \overline{\omega^2 z X'_0 X_0 T \hat{u}^2} \end{bmatrix}$$

et on a donc pour estimateur convergent de la matrice de variance

$$\hat{V} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \hat{G}^{-1} \hat{V} \hat{G}'^{-1}$$

Le cas multiprogramme

On reprend le cas initial avec affectation à trois groupes : $1(Z = 0)$ le groupe de contrôle, $Z_1 = 1(Z = 1)$ le groupe affecté au programme 1 et $Z_2 = 1(Z = 2)$ le groupe affecté au programme 2. Une fois affectés, les individus peuvent se comporter de différentes façons et entrer dans les programmes ou non. On a une variable d'entrée dans les programmes T :

$T = 0$ correspond à l'absence de programme, $T = 1$ correspond à l'entrée dans le programme 1 et $T = 2$ correspond à l'entrée dans le programme 2. On a aussi des variables d'output correspondant au programme reçu. $y(0)$ lorsque l'on ne suit pas de programme, $y(1) = y(0) + \Delta(1)$ lorsque l'on suit le programme 1 et $y(2) = y(0) + \Delta(2)$ lorsque l'on suit le programme 2. On a de même des variables de programme potentielles correspondant à l'affectation : $T(0)$ est la variable d'entrée dans le programme lorsque l'on est affecté au groupe $Z = 0$, $T(1)$ celle lorsque l'on est affecté au groupe $Z = 1$ et $T(2)$ celle lorsque l'on est affecté au groupe $Z = 2$.

On considère les paramètres c_1 , c_2 , définis par les conditions d'orthogonalité suivantes :

$$E \left(\begin{pmatrix} X'_0 \\ X'_0 Z_1 \\ X'_0 Z_2 \end{pmatrix} (y - X_0 b - X_0 c_1 1(T = 1) - X_0 c_2 1(T = 2)) \right) = 0 \quad (25)$$

Sous certaines hypothèses concernant les comportements d'entrée dans les programmes une fois faite l'affectation aléatoire, ces paramètres s'interprètent encore comme des moyenne des effets des programmes sur des sous populations particulières. On a la proposition suivante :

Proposition 6 *Lorsque les schémas d'affectation au programme satisfont les contraintes $(T(2) = 1) = (T(0) = 1) \leq (T(1) = 1)$ et $(T(1) = 2) = (T(0) = 2) \leq (T(2) = 2)$, les paramètres annulant les conditions d'orthogonalité sont les coefficients de la projection linéaire des effets moyens du traitement sur les caractéristiques X_0 pour la population $(T(1) = 1) - (T(0) = 1) = 1$ en ce qui concerne c_1 et pour la population $(T(2) = 2) - (T(0) = 2) = 1$ en ce qui concerne c_2 .*

Pour que les contraintes soient satisfaites, il faut que la population ne contienne que :

- des never takers $T(0) = 0, T(1) = 0, T(2) = 0$
- des always takers programme 1 $T(0) = 1, T(1) = 1, T(2) = 1$
- des always takers programme 2 $T(0) = 2, T(1) = 2, T(2) = 2$
- des compliers $T(0) = 0, T(1) = 1, T(2) = 2$
- des selective takers programme 1 $T(0) = 0, T(1) = 1, T(2) = 0$
- des selective takers programme 2 $T(0) = 0, T(1) = 0, T(2) = 2$

La population $(T(1) = 1) - (T(0) = 1) = 1$ est composée des compliers et des selective taker 1, la population $(T(2) = 2) - (T(0) = 2) = 1$ est composée des compliers et des selective takers 2.

On voit que les schémas d'entrée compatibles avec l'identification causale sont plus contraints que ceux prévalant lorsqu'il n'y a qu'un seul traitement. Les schémas conservant la symétrie entre les deux programmes sont tels que $T(0) \in \{0\}$, $T(1) \in \{0, 1\}$, $T(2) \in \{0, 2\}$: le groupe de contrôle ne peut entrer dans aucun programme, le groupe affecté au programme 1 a le droit de décliner le programme 1, mais ne peut entrer dans le programme 2, et idem pour le programme 2. Lorsqu'il n'y a qu'un seul traitement, le design expérimental est en outre robuste à des erreurs. Lorsque le programme est proposé à un individu affecté au groupe de contrôle, par erreur plus ou moins intentionnelle, le cadre expérimental continue à produire en général des estimateurs ayant un sens causal, pourvu que l'on ait bien conservé l'affectation initiale. Ici, ce n'est pas possible il est nécessaire que le protocole expérimental soit respecté parfaitement. Les seules sources d'écart au protocole expérimental doivent être des erreurs d'affectation purement aléatoires.

Preuve de la proposition 6 voir Annexe - Preuve

Proposition 7 Soit le cas où il y a des réallocations à partir d'une situation initiale $T^*(j)$ satisfaisant les restrictions d'affectation :

$$\begin{aligned} T(0) &= P_{00}T^*(0) + P_{01}T^*(1) + P_{02}T^*(2) \\ T(1) &= P_{10}T^*(0) + P_{11}T^*(1) + P_{12}T^*(2) \\ T(2) &= P_{20}T^*(0) + P_{21}T^*(1) + P_{22}T^*(2) \end{aligned}$$

avec P_{kj} des variables prenant la valeur 0 ou 1 et représentant la transition entre l'affectation initiale k et l'affectation finale j et satisfaisant : $P_{k0} + P_{k1} + P_{k2} = 1$. Si ces réaffectations sont purement aléatoires, elles ne perturbent pas l'identification des paramètres. Les coefficients c_1 et c_2 s'interprètent comme les coefficients des projections linéaires des effets des programmes sur les mêmes populations qu'avant, mais définies à partir des variables $T^*(j)$

Preuve de la proposition 7 voir Annexe - Preuve

Pour mémoire, il y a bien sûr aussi la situation dans laquelle les effets Δ_1 et Δ_2 sont hétérogènes mais indépendants des variables d'entrée $T^*(j)$

Proposition 8 Sous l'hypothèse $\Delta_1 = X_0c_1 + \varepsilon_1$, avec $E(\varepsilon_1|X_0) = 0$ et $T^*(j) \perp \varepsilon_1 | X_0$ et $\Delta_2 = X_0c_2 + \varepsilon_2$, avec $E(\varepsilon_2|X_0) = 0$ et $T^*(j) \perp \varepsilon_2 | X_0$, les paramètres identifiés par les conditions d'orthogonalité 25 sont les effets moyens du traitement dans la population.

Annexe - Preuves

Preuve de la proposition 1 *La décision observée d'entrer de l'individu est simplement $T = T(0)(1 - z) + T(1)z$. De même l'output observé y est simplement $y = y(1)T + y(0)(1 - T)$. On peut écrire ces deux quantités en fonction seulement des outputs et participations potentielles :*

$$y = y(0) + T(0)(y(1) - y(0)) + (y(1) - y(0))(T(1) - T(0))z \quad (26)$$

$$= g_0 + g_1z \quad (27)$$

$$T = T(0) + (T(1) - T(0))z \quad (28)$$

$$= h_0 + h_1z \quad (29)$$

En considérant les expressions des équations (27) et (29), on a

$$(y - a - bT) = (g_0 - a - bh_0 + (g_1 - bh_1)) = f_0(\theta) + zf_1(\theta) \quad (30)$$

où $\theta = (a, b)$, $f_0(\theta) = g_0 - a - bh_0$ et $f_1(\theta) = g_1 - bh_1$

Les conditions d'orthogonalité (1) se réécrivent donc

$$E \left(\begin{pmatrix} 1 \\ z \end{pmatrix} (f_0(\theta) + zf_1(\theta)) \right) = E \begin{pmatrix} f_0 + zf_1 \\ zf_0 + zf_1 \end{pmatrix} = 0$$

Compte tenu de la relation d'indépendance, on a donc

$$\begin{pmatrix} E(f_0) + E(z)E(f_1) \\ E(f_0)E(z) + E(f_1)E(z) \end{pmatrix} = 0$$

Comme $E(z) \neq 1$ on voit que les conditions d'orthogonalité identifient le paramètre θ tel que

$$E(f_0(\theta)) = E(g_0 - a - bh_0) = 0 \quad (31)$$

$$E(f_1(\theta)) = E(g_1 - bh_1) = 0 \quad (32)$$

On en déduit en particulier que le paramètre b identifie

$$b = \frac{E(g_1)}{E(h_1)} = \frac{E(y(1) - y(0))(T(1) - T(0))}{E(T(1) - T(0))} \quad (33)$$

Preuve de la proposition 3 Compte tenu de l'expression du résidu de l'équation (30), les conditions précédentes (4) s'écrivent

$$\begin{pmatrix} E(w(X_c, z) f_0) + E(w(X_c, z) z f_1) \\ E(w(X_c, z) f_0 z) + E(w(X_c, z) f_1 z) \end{pmatrix} = 0$$

On a compte tenu de la relation d'indépendance

$$\begin{aligned} E(w(X_c, z) f_0) &= E(E(w(X_c, z) f_0 | X_c)) \\ &= E(E(w(X_c, z) | X_c) E(f_0 | X_c)) \end{aligned}$$

De même pour $f = f_0$ ou $f = f_1$ on a :

$$\begin{aligned} E(w(X_c, z) z f) &= E(E(w(X_c, z) z f | X_c)) \\ &= E(E(w(X_c, z) z | X_c) E(f | X_c)) \end{aligned}$$

On voit donc que pour que le paramètre identifié par les conditions d'orthogonalité soit le même, il suffit que l'on ait $E(w(X_c, z) | X_c) = A$ et $E(w(X_c, z) z | X_c) = B$ avec $A \neq B$.

En effet dans ce cas, les équations précédentes se résument à

$$\begin{pmatrix} A & B \\ B & B \end{pmatrix} \begin{pmatrix} E(f_0) \\ E(f_1) \end{pmatrix} = 0$$

$$\begin{aligned} E(w(X_c, z) | X_c) &= A = w(X_c, 1) P(z = 1 | X_c) + w(X_c, 0) P(z = 0 | X_c) \\ E(w(X_c, z) z | X_c) &= B = w(X_c, 1) P(z = 1 | X_c) \end{aligned}$$

Il en résulte que tout système de poids de la forme

$$w(X_c, z_0) = \frac{C_{z_0}}{P(z = z_0 | X_c)} \quad (34)$$

permet d'identifier les paramètres annulant $E(f_0)$ et $E(f_1)$. En particulier pour $A = 1$ et $B = 2$, on a le système de poids proposé.

Preuve de la proposition 5 Là aussi, on peut réécrire l'expression du résidu $(y - X_0 a - T X_0 b)$ en utilisant les équations 27 et 29, on a

$$(y - X_0a - X_0bT) = (g_0 - X_0a - X_0bh_0 + z(g_1 - X_0bh_1)) = f_0(\theta) + zf_1(\theta) \quad (35)$$

où $\theta = (a, b)$, $f_0(\theta) = g_0 - X_0a - X_0bh_0$ et $f_1(\theta) = g_1 - X_0bh_1$

Les conditions d'orthogonalité ?? se réécrivent donc

$$E \left(w(X_c, z) \begin{pmatrix} X'_0 \\ zX'_0 \end{pmatrix} (f_0(\theta) + zf_1(\theta)) \right) = E \left(w(X_c, z) \begin{pmatrix} X'_0f_0 + zX'_0f_1 \\ zX'_0f_0 + zX'_0f_1 \end{pmatrix} \right) = 0$$

Compte tenu de la relation d'indépendance, les conditions se réécrivent comme

$$\begin{pmatrix} 2E(X'_0f_0) + E(X'_0f_1) \\ E(X'_0f_0) + E(X'_0f_1) \end{pmatrix} = 0$$

Ceci provient du fait qu'en raison des poids choisis on a $E(w(X_c, z)) = 2$ et $E(zw(X_c, z)) = 1$. On voit donc que les conditions d'orthogonalité identifient le paramètre θ tel que

$$E(X'_0f_0(\theta)) = E(X'_0g_0 - X_0a - X_0bh_0) = 0 \quad (36)$$

$$E(X'_0f_1(\theta)) = E(X'_0g_1 - X'_0X_0bh_1) = 0 \quad (37)$$

On en déduit en particulier que le paramètre b identifie

$$b = E(X'_0X_0h_1)^{-1} E(X'_0g_1) \quad (38)$$

Comme $h_1 = T(1) - T(0)$ et $g_1 = (T(1) - T(0))(y(1) - y(0))$. Sous l'hypothèse de monotonie, on voit que l'on a

$$b = E(X'_0X_0 | T(1) - T(0) = 1)^{-1} E(X'_0(y(1) - y(0)) | T(1) - T(0) = 1) \quad (39)$$

et on voit que le paramètre b s'interprète naturellement comme le vecteur des coefficients de la projection orthogonale de l'effet du programme sur les caractéristiques X_0 . Remarquons que si à partir de ces estimateurs, on voulait remonter à un paramètre global, il faudrait réintégrer par rapport à la distribution des variables de conditionnement sachant que l'on est un complier.

Preuve de la proposition 6 On a :

$$T = T(0) + (T(1) - T(0))Z_1 + (T(2) - T(0))Z_2$$

soit

$$\begin{aligned} 1(T = 1) &= 1(T(0) = 1) + (1(T(1) = 1) - 1(T(0) = 1))Z_1 + (1(T(2) = 1) - 1(T(0) = 1))Z_2 \\ &= h_0 + h_1Z_1 + h_2Z_2 \end{aligned}$$

et

$$\begin{aligned} 1(T = 2) &= 1(T(0) = 2) + (1(T(1) = 2) - 1(T(0) = 2))Z_1 + (1(T(2) = 2) - 1(T(0) = 2))Z_2 \\ &= i_0 + i_1Z_1 + i_2Z_2 \end{aligned}$$

on a en outre pour la variable y

$$\begin{aligned} y &= y(0) + \Delta_1 1(T = 1) + \Delta_2 1(T = 2) \\ &= y(0) + \Delta_1 (h_0 + h_1Z_1 + h_2Z_2) + \Delta_2 (i_0 + i_1Z_1 + i_2Z_2) \\ &= y(0) + \Delta_1 h_0 + \Delta_2 i_0 + (\Delta_1 h_1 + \Delta_2 i_1) Z_1 + (\Delta_1 h_2 + \Delta_2 i_2) Z_2 \\ &= g_0 + g_1Z_1 + g_2Z_2 \end{aligned}$$

Si on considère maintenant les conditions d'orthogonalité 25, on a

$$\begin{aligned} y - X_0b &- X_0c_1 1(T = 1) - X_0c_2 1(T = 2) = g_0 + g_1Z_1 + g_2Z_2 - X_0b \\ &- X_0c_1 (h_0 + h_1Z_1 + h_2Z_2) - X_0c_2 (i_0 + i_1Z_1 + i_2Z_2) \\ &= g_0 - X_0b - X_0c_1 h_0 - X_0c_2 i_0 \\ &+ (g_1 - X_0c_1 h_1 - X_0c_2 i_1) Z_1 + (g_2 - X_0c_1 h_2 - X_0c_2 i_2) Z_2 \\ &= f_0 + f_1Z_1 + f_2Z_2 \end{aligned}$$

Compte tenu des hypothèses d'indépendance, on a

$$\begin{pmatrix} E(X'_0 f_0) + E(X'_0 f_1) E(Z_1) + E(X'_0 f_2) E(Z_2) \\ E(X'_0 f_0) E(Z_1) + E(X'_0 f_1) E(Z_1) \\ E(X'_0 f_0) E(Z_2) + E(X'_0 f_2) E(Z_2) \end{pmatrix} = 0$$

On en conclut donc que l'on doit avoir

$$\begin{pmatrix} E(X'_0 f_0) \\ E(X'_0 f_1) \\ E(X'_0 f_2) \end{pmatrix} = 0$$

Ceci conduit donc en particulier aux deux équations :

$$\begin{aligned} E(X'_0 \Delta_1 h_1) + E(X'_0 \Delta_2 i_1) &= E(X'_0 X_0 h_1) c_1 + E(X'_0 X_0 i_1) c_2 \\ E(X'_0 \Delta_1 h_2) + E(X'_0 \Delta_2 i_2) &= E(X'_0 X_0 h_2) c_1 + E(X'_0 X_0 i_2) c_2 \end{aligned}$$

Le cas intéressant est celui dans lequel on a $h_2 = 0$ et $i_1 = 0$. Cela correspond à $(T(2) = 1) = (T(0) = 1)$ et $(T(1) = 2) = (T(0) = 2)$. On vérifie que ceci n'est possible que si on a des taker uniformes (choix de 0 quelle que soit l'affectation, choix de 1 quelle que soit l'affectation et choix de 2 quelle que soit l'affectation), des "compliers" : $T(0) = 0$, $T(1) = 1$ et $T(2) = 2$ et des selective takers $T(0) = 0$, $T(1) = 0$ et $T(2) = 2$ ou $T(0) = 0$, $T(1) = 1$ et $T(2) = 0$. Alors les paramètres s'interprètent comme précédemment comme les coefficients de la projection des effets sur les X pour les $h_1 = 1$ et $i_2 = 1$.

Preuve de la proposition 7

On a en effet

$$\begin{aligned} h_1 &= (T(1) = 1) - (T(0) = 1) \\ &= (P_{11} - P_{01})((T^*(1) = 1) - (T^*(0) = 1)) + (P_{12} - P_{02})((T^*(2) = 1) - (T^*(0) = 1)) \\ &= (P_{11} - P_{01})((T^*(1) = 1) - (T^*(0) = 1)) \\ h_2 &= (P_{21} - P_{01})((T^*(1) = 1) - (T^*(0) = 1)) + (P_{22} - P_{02})((T^*(2) = 1) - (T^*(0) = 1)) \\ &= (P_{21} - P_{01})((T^*(1) = 1) - (T^*(0) = 1)) \\ i_1 &= (P_{11} - P_{01})((T^*(1) = 2) - (T^*(0) = 2)) + (P_{12} - P_{02})((T^*(2) = 2) - (T^*(0) = 2)) \\ &= (P_{12} - P_{02})((T^*(2) = 1) - (T^*(0) = 2)) \\ i_2 &= (P_{21} - P_{01})((T^*(1) = 2) - (T^*(0) = 2)) + (P_{22} - P_{02})((T^*(2) = 2) - (T^*(0) = 2)) \\ &= (P_{22} - P_{02})((T^*(2) = 2) - (T^*(0) = 2)) \end{aligned}$$

et on en déduit le résultat facilement.