UNE BASE D'ANALYSE LONGITUDINALE DE DONNEES D'ENTREPRISES : Problèmes et résultats

Marie-Christine Parent

Qu'est-ce que la base d'analyse longitudinale des entreprises de Suse ?

Ce vocable regroupe trois bases de données d'entreprises, individuelles et temporelles, articulées grâce à un identifiant unique. Les données contenues dans ces bases sont principalement issues de la source Suse [cf. encadré 1] et collectées sur la période 1984 à 1992 ;

la première base gère les identifiants des entreprises et assure notamment l'interface avec les sources fiscales annuelles ;

la deuxième base comprendra à terme l'intégralité des entreprises du champ des sociétés et quasi-sociétés assujetties à une déclaration fiscale. Pour le moment, elle se limite aux données collectées dans le cadre des bénéfices industriels et commerciaux (BIC) soit 3,5 millions d'entreprises concernées sur la période 1984-1992. Le nombre de variables contenues dans cette base est relativement limité : nature de l'activité exercée (APE), effectif, chiffre d'affaires, production, valeur ajoutée, consommations intermédiaires, masse salariale, actif net et excédent brut d'exploitation. Schématiquement, ces variables permettent d'estimer une fonction de production et donnent quelques éléments d'informations sur les coûts supportés par l'entreprise. Cette base sera mise à jour par la suite régulièrement avec les déclarations fiscales ultérieures, de manière à constituer une base d'analyse sur longue période ;

la troisième base, encore en cours d'élaboration, porte sur les grandes entreprises au sens de Suse (principalement les plus de vingt salariés). Elle comprend davantage de variables d'intérêt (la plupart des soldes intermédiaires de gestion et les principaux ratios d'analyse financière) et des données issues de sources externes comme l'enquête sur les liaisons financières entre les sociétés, LIFI. Pour éviter les biais classiques liés au franchissement du seuil d'interrogation, on garde l'enveloppe du champ des vingt

salariés : toute entreprise ayant au moins une fois franchi le seuil sur la période étudiée reste sélectionnée. Mais cela suppose de modifier la base de données en continu puisque l'ajout d'une année peut conduire à devoir réintégrer sur toute la période, des entreprises qui ont franchi à la hausse le seuil la dernière année. Réciproquement, on s'expose à devoir gérer des entreprises durablement sorties du champ parce qu'une année, elles ont figuré dans le champ. C'est pourquoi, une méthodologie de type "Bridge" reposant sur l'élaboration de bases décennales maintenant le champ constant en glissement peut être adoptée.

Le travail effectué a consisté à rapprocher et à mettre en cohérence les données annuelles pour chaque entreprise concernée. Théoriquement simple, ce travail s'est heurté à des problèmes méthodologiques communs à toute source de données, plus quelques-uns spécifiques aux données d'entreprises.

Quels problèmes méthodologiques ?

Toute source de données, qu'elle soit exhaustive ou par sondage, est confrontée au problème des non-répondants (ou absences). Ce problème a plus ou moins d'impact quand on travaille en coupe instantanée selon que l'on vise ou non la représentativité macro-économique. Il prend une dimension beaucoup plus critique quand on cherche à analyser les données de manière longitudinale (en dimension temporelle). On est ainsi conduit à surestimer de manière non négligeable les mouvements d'entrée-sortie. Dans le cas de la source Suse, la démographie des unités peut ainsi être majorée de 10 % par rapport aux vraies créations-disparitions d'entreprises.

Le rapprochement des mêmes entreprises sur plusieurs années permet de lever en partie cette hypothèque : on peut en effet encadrer les comptes manquants une année donnée ("trou") par les déclarations précédente et suivante de l'entreprise, au moins sur le passé. Il reste alors à mettre en place un modèle d'estimation de ces "trous" suivant que l'on considère que les absences sont aléatoires ou sujettes à certains déterminismes.

Mais, dans le cas des données fiscales, on se heurte au problème du mauvais repérage des unités : l'identifiant de l'entreprise, le Siren, est souvent mal renseigné : de 5 % à 10 % des déclarations fiscales reçues chaque année ne peuvent, faute d'un identifiant correct, être attribuées à une entreprise déterminée. Dans ce cas, nous sommes en présence de "fausses absentes", c'est-à-dire d'entreprises présentes dans la source mais mal repérées.

Le deuxième travail effectué a consisté à affecter à une entreprise, ces données annuelles mal identifiées. C'est possible grâce à l'existence d'un deuxième identifiant des entreprises, géré par la direction générale des Impôts (DGI). Chaque chronique correspondant à une entreprise a, dans la mesure du possible, été associée à un Siren de

manière à faciliter le rapprochement avec des sources extérieures. Toutefois, dans la base de données finale, il reste encore 10 % des chroniques (350 000 entreprises) qui n'ont pas de Siren renseigné sur les neuf années de la base.

Enfin, une fois la base constituée, différents modèles d'interpolation des déclarations absentes ont été mis au point par analyse des systématismes rencontrés.

Quelles utilisations ?

Elles sont nombreuses en matière d'études comme en production.

Production et Comptabilité nationale

- En interne, à la gestion de Suse, la base des identifiants permet d'améliorer l'identification des entreprises "mal sirénisées" ;

- la base des données rassemble l'intégralité des déclarations fiscales reçues de 1984 à 1992. Elle constitue en quelque sorte l'enveloppe théorique des objets économiques susceptibles de faire une déclaration aux BIC. Ce champ théorique des BIC étant élucidé, il est alors possible de mettre la source fiscale en perspective par rapport au référentiel théorique du secteur institutionnel des sociétés et quasi-sociétés et d'améliorer ainsi la cohérence et la représentativité macro-économique de la source annuelle ;

- l'analyse du passé permet de donner aux comptables nationaux gérant le secteur institutionnel des sociétés et quasi-sociétés, des taux de redressement pour absence plus précis que ceux actuellement utilisés. Elle donne également, dans les délais actuels du compte définitif, une liste *individuelle* d'entreprises absentes qui couvre environ les 2/3 des absences d'une année donnée, ce qui permet de redresser les absences de manière individuelle et non plus seulement agrégée ;

- on peut sélectionner à partir du répertoire Sirène, la liste des objets fiscaux potentiels d'une année : il est alors possible de boucler l'exhaustivité des années terminales de la base grâce aux données extraites de Sirène : à l'heure actuelle, on recueille chaque année environ 2,2 millions de déclarations fiscales (BIC, BNC) alors que Sirène recense près de 2,5 millions d'entreprises appartenant au champ SQS.

Études

- tirage d'échantillons représentatifs pour des études sur les entreprises ;

- études sur données de Panel, analyse temporelle, projections de population, modèles de survie ;

- utilisation à des fins régionales : la sélection d'une base régionale est possible grâce à la présence du code département de l'entreprise. Toutefois, ce code département est celui de l'établissement principal de l'entreprise, ce qui pose des problèmes pour les entreprises ayant plusieurs établissements. Dans la base temporelle sur les grandes entreprises (plus de vingt salariés) figurent le nombre d'établissements de l'entreprise, le nombre de ses régions d'implantation ainsi que l'appartenance à un département hors métropole.

Encadré 1

Suse : Système unifié de statistiques d'entreprises

Il a été mis en place à partir des déclarations fiscales des unités relevant des bénéfices industriels et commerciaux et non commerciaux (BIC et BNC) et des enquêtes annuelles d'entreprises (EAE) effectuées par l'Insee et les services statistiques des différents ministères. La confrontation et la mise en cohérence de ces deux sources sont réalisées sur le champ des entreprises de plus de vingt salariés. Ces deux ensembles de données constituent la source fondamentale pour l'élaboration du secteur institutionnel des sociétés et quasi-sociétés de la Comptabilité nationale.

À compter de 1989, on change d'architecture informatique (passage à la version 3 de Suse), ce qui se traduit par la modification d'un certain nombre de programmes de redressement ou de correction automatique. D'autre part, on ne dispose plus des effectifs des entreprises assujetties au régime fiscal forfaitaire. Cela n'a pas d'incidence sur la cohérence macro-économique de la source car il s'agit de petites entreprises et souvent d'entreprises individuelles, sans salarié. D'autre part, depuis 1984, par franchissement automatique du seuil d'imposition fiscal, le nombre d'entreprises forfaitaires a été diminué par deux.

Constitution de la base d'analyse longitudinale

Cette partie s'attache à décrire la méthodologie utilisée pour constituer les différentes bases de données temporelles constituées à partir de Suse.

Les absences dans Suse

Le tableau 1 donne des éléments de réflexion sur l'exhaustivité de Suse, une année donnée : si on se restreint au champ des entreprises bien identifiées de Suse, en moyenne une entreprise absente de la source fiscale une année donnée a 30 % de "chances" de réapparaître l'année suivante (chroniques dites à "un trou"), 8 % de chances d'être absente deux années de suite et ainsi de suite. Pour l'année 1985, on recense *a priori* 263 926 disparitions (présentes 1984, absentes 1985). 43 % de ces absentes réapparaîtront par la suite, conduisant à un nombre effectif de cessations en 1985 de 149 836 entreprises. La situation symétrique peut être mise en évidence pour les créations.

Ces résultats mettent en lumière les problèmes d'exhaustivité de la source fiscale et un moyen simple de pallier ces absences, au moins pour le passé. Toutefois, l'exhaustivité macro-économique de la source une année donnée n'est pas forcément entachée par ces erreurs : en effet, les entreprises absentes sont d'une part en général des petites entreprises (entreprises forfaitaires) ; d'autre part, elles peuvent être présentes dans la source sous un identifiant Siren erroné : dans ce cas, on parlera de "fausses" absences.

"Vraies" absentes ou "fausses" absentes

Distinguer les "fausses" absentes des vraies, présente plusieurs intérêts : d'abord, compléter des chroniques "à trous" permet de limiter le nombre de doubles comptes générés automatiquement par l'interpolation nécessaire à toute analyse temporelle. D'autre part, les "fausses" et les "vraies" absentes peuvent ne pas suivre des lois de probabilité d'absence équivalentes, puisqu'il s'agit de populations différentes. Les "vraies" absentes peuvent être soit aléatoires, soit obéir à des processus clairement identifiés (réveil d'entreprises en sommeil). Là encore, on est en présence d'un mélange de populations qu'il convient de cerner pour pouvoir construire des modèles de "redressement des trous". Déceler les "fausses" absentes et les réidentifier facilite d'autre part les rapprochements avec des sources extérieures.

La démarche adoptée

Pour séparer dans la source fiscale les "vraies" absentes des "fausses" absentes, sur la période 1984 à 1992, il convient de réidentifier les entreprises mal "sirénisées". L'objectif est de créer une base d'identifiants uniques permettant le suivi de l'ensemble des entreprises présentes dans la source Suse de 1984 à 1992.

Il est d'autre part nécessaire d'améliorer la qualité d'identification ("sirénisation") des fichiers annuels de manière à minimiser les problèmes de rapprochement temporel (éviter les doubles comptes ou les duplications d'entreprises sans Siren). Pour cela, on utilise le fichier des redevables professionnels de la direction générale des Impôts qui répertorie toutes les entreprises attendues au titre de l'impôt sur les sociétés une année donnée. La Direction générale des impôts gère dans ce fichier un indicateur de suivi de l'entreprise, l'IFRP. Ce dernier obéit à des règles de continuité différentes du Siren, même si, théoriquement, un IFRP est créé pour tout changement de Siren. À l'aide de ces deux indicateurs, on cherche à corriger certaines chroniques incomplètes sur le Siren, tout en s'affranchissant au mieux des chroniques tronquées à la suite d'un changement d'IFRP. De manière à bénéficier des gains de qualité du Siren réalisés depuis 1993, grâce à la mise en concordance automatique du fichier des redevables permanents avec Sirène, on rapproche par le biais de l'identifiant de la DGI, l'IFRP, les fichiers BIC des années 1984 à 1992 des fichiers des redevables permanents attendus au titre de 1993 et 1994.

Ces résultats mettent en lumière les problèmes d'exhaustivité de la source fiscale et un moyen simple de pallier ces absences, au moins pour le passé. Toutefois, l'exhaustivité macro-économique de la source une année donnée n'est pas forcément entachée par ces erreurs : en effet, les entreprises absentes sont d'une part en général des petites entreprises (entreprises forfaitaires) ; d'autre part, elles peuvent être présentes dans la source sous un identifiant Siren erroné : dans ce cas, on parlera de "fausses" absences.

"Vraies" absentes ou "fausses" absentes

Distinguer les "fausses" absentes des vraies, présente plusieurs intérêts : d'abord, compléter des chroniques "à trous" permet de limiter le nombre de doubles comptes générés automatiquement par l'interpolation nécessaire à toute analyse temporelle. D'autre part, les "fausses" et les "vraies" absentes peuvent ne pas suivre des lois de probabilité d'absence équivalentes, puisqu'il s'agit de populations différentes. Les "vraies" absentes peuvent être soit aléatoires, soit obéir à des processus clairement identifiés (réveil d'entreprises en sommeil). Là encore, on est en présence d'un mélange de populations qu'il convient de cerner pour pouvoir construire des modèles de "redressement des trous". Déceler les "fausses" absentes et les réidentifier facilite d'autre part les rapprochements avec des sources extérieures.

La démarche adoptée

Pour séparer dans la source fiscale les "vraies" absentes des "fausses" absentes, sur la période 1984 à 1992, il convient de réidentifier les entreprises mal "sirénisées". L'objectif est de créer une base d'identifiants uniques permettant le suivi de l'ensemble des entreprises présentes dans la source Suse de 1984 à 1992.

Il est d'autre part nécessaire d'améliorer la qualité d'identification ("sirénisation") des fichiers annuels de manière à minimiser les problèmes de rapprochement temporel (éviter les doubles comptes ou les duplications d'entreprises sans Siren). Pour cela, on utilise le fichier des redevables professionnels de la direction générale des Impôts qui répertorie toutes les entreprises attendues au titre de l'impôt sur les sociétés une année donnée. La Direction générale des impôts gère dans ce fichier un indicateur de suivi de l'entreprise, l'IFRP. Ce dernier obéit à des règles de continuité différentes du Siren, même si, théoriquement, un IFRP est créé pour tout changement de Siren. À l'aide de ces deux indicateurs, on cherche à corriger certaines chroniques incomplètes sur le Siren, tout en s'affranchissant au mieux des chroniques tronquées à la suite d'un changement d'IFRP. De manière à bénéficier des gains de qualité du Siren réalisés depuis 1993, grâce à la mise en concordance automatique du fichier des redevables permanents avec Sirène, on rapproche par le biais de l'identifiant de la DGI, l'IFRP, les fichiers BIC des années 1984 à 1992 des fichiers des redevables permanents attendus au titre de 1993 et 1994.

Tableau 1

Les "trous" entre 1984 et 1992. Quelques estimations sur le champ des bien Sirénisés de Suse avant toute correction

Disparitions entre n-1 et n(absentes de n)	Total	fausses absentes	"vraies" disparitions	réapparues en 86	réapparues en 87	réapparues en 88	réapparues en 89	réapparues en 90	réapparues en 91	réapparues en 92
prés. en 84 disp. en 85	269522	114092 43.2%	149836 56.8%	78831 29.9%	20793 7.9%	5939 2.2%	3028 1.1%	2190 0.8%	1982 0.7%	1329 0.5%
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 87**	**réap. en 88**	**réap. en 89**	**réap. en 90**	**réap. en 91**	**réap. en 92**	
prés. en 85 disp. en 86	229522	90490 39.4%	139032 60.6%	62552 27.2%	15125 6.6%	5298 2.3%	3357 1.5%	2397 1.0%	1761 0.8%	
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 88**	**réap. en 89**	**réap. en 90**	**réap. en 91**	**réap. en 92**		
prés. en 86 disp. en 87	252166	100976 40.0%	151190 60.0%	68961 27.3%	18201 7.2%	6867 2.7%	4179 1.6%	2768 1.1%		
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 89**	**réap. en 90**	**réap. en 91**	**réap. en 92**			
prés. en 87 disp. en 88	293822	124459 43.4%	169363 57.6%	87755 29.9%	23426 8.0%	7977 2.7%	5301 1.8%			
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 90**	**réap. en 91**	**réap. en 92**				
prés. en 88 disp. en 89	308400	123312 40.0%	182088 59.0%	90625 29.4%	22302 7.2%	30572 10.0%				
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 91**	**réap. en 92**					
prés. en 89 disp. en90	305495	114234 37.4%	191261 62.6%	83662 27.4%	30572 10.0%					
	Total	**fausses ab.**	**"vraies" disp.**	**réap. en 92**						
prés. en 90 disp. en91	312273	102092 32.7%	210181 67.3%	102092 32.7%						
	Total	**fausses ab.**	**"vraies" disp.**							
prés. en 91 disp. en92	291562									

Qualité des sources annuelles

Tableau 2

Avant rapprochement : taux de Siren en double ou non renseignés

(en pourcentage d'entreprises, sur l'ensemble du champ BIC)

1984	1985	1986	1987	1988	1989	1990	1991	1992
17.6	13,5	14,8	13,9	12,2	12,0	9,2	6,1	4,2

Tableau 2 bis

Après rapprochement: gain net dû au rapprochement avec le fichier des redevables professionnels de la fin 1993

(en pourcentage des entreprises mal identifiées)

1984	1985	1986	1987	1988	1989	1990	1991	1992
3,7	13,5	13,7	15,2	14,0	19,9	20,0	31,0	14,2

Méthodologie utilisée pour le rapprochement temporel

Le principe est simple : on sépare le champ des entreprises en deux groupes : celles bien "sirénisées", c'est-à-dire dont le Siren est renseigné et unique sur toute la période. Ces entreprises seront rapprochées temporellement sur le Siren et présentent des chroniques éventuellement incomplètes (compte absent une année donnée) mais dont le Siren est unique et valide sur les neuf années de la base. Les autres entreprises, qui présentent au moins une année un Siren invalide, seront rapprochées sur l'IFRP. Les chroniques temporelles ainsi obtenues, éventuellement incomplètes, sont caractérisées par un identifiant IFRP unique. Les entreprises éliminées au cours de ce double processus seront traitées manuellement. On obtient donc deux séries de chroniques temporelles qu'il faut compléter sur elles-mêmes ou entre-elles. [cf. annexe 1]

La méthode est la même dans chaque cas : chaque groupe de chroniques est caractérisé par l'unicité d'un des deux identifiants (Siren dans le premier cas, IFRP dans le second) . On se sert du deuxième identifiant pour "recoller" les trajectoires.

Si on prend comme exemple l'enrichissement sur lui-même du fichier des Siren valides, on génère pour chaque IFRP rencontré, autant d'enregistrements qu'il y a d'entreprises *différentes* concernées par cet IFRP. On associe à chaque enregistrement IFRP ainsi créé l'intégralité de l'historique de l'entreprise associée. Une entreprise donnée sera donc dupliquée en autant d'enregistrements qu'elle a d'IFRP différents dans sa chronique.

Si un même IFRP est présent plusieurs années dans la chronique d'une même entreprise, l'enregistrement IFRP n'est créé qu'une fois. Par contre, si deux ou plusieurs entreprises différentes sont associées à un même IFRP, on crée autant d'enregistrements qu'il y a de Siren différents.

On attribue alors à chaque enregistrement IFRP une indicatrice de présence annuelle qui vaut 1 si l'entreprise correspondante est présente l'année considérée, 0 sinon. Pour chaque IFRP distinct, on somme ensuite ces indicatrices. Les IFRP qui ne sont rencontrés qu'une fois sur toute la période ne sont pas complétables en interne et sont éliminés. Les IFRP rencontrés plusieurs fois correspondent *a priori* à des chroniques complétables les unes avec les autres. Toutefois, ne sont réellement complétables que les chroniques associées à des IFRP dont les indicatrices annuelles cumulées ne prennent que la valeur 1 ou 0. Une valeur supérieure à 1 signifie en effet que le même IFRP est associé pour la même année à deux entreprises différentes. Le cas caricatural est celui de 1984 où un grand nombre d'IFRP n'ont été saisis que sur 10 caractères au lieu de 19.

Pour compléter les chroniques, on dissocie les enregistrements IFRP en double et on met à jour ces enregistrements deux à deux grâce à l'IFRP. On procède de même pour compléter sur lui-même le fichier des chroniques suivies sur l'IFRP en utilisant le Siren comme identifiant complémentaire. On complète enfin les deux fichiers résultant entre eux en utilisant la même méthode modifiée à la marge.

La base "brute", complétée manuellement par les chroniques exclues des appariements initiaux, contient environ 3,46 millions d'entreprises contre 3,59 millions pour la réunion des deux fichiers initiaux. On a donc reconstitué 130 000 chroniques. A la fin du processus d'enrichissement, on obtient un peu plus de 396 000 chroniques complètes, c'est-à-dire des entreprises présentes les neuf années. Il reste encore de nombreux "trous", mais ayant épuisé l'information disponible, on ne peut que considérer qu'il s'agit d'entreprises vraiment absentes.

On attribue alors à chaque chronique un identifiant unique, caractérisant l'entreprise associée : le SIRIFRP. Cet identifiant est le produit de la concaténation du dernier Siren valide de l'entreprise et de l'IFRP qui lui est associé. Si pour les neuf années de la base, l'entreprise ne présente pas de Siren valide, on garde le dernier Siren renseigné et l'IFRP qui lui correspond. Enfin, les 348 000 chroniques sans Siren seront suivies sur l'IFRP.

L'identifiant ainsi défini est unique. Il servira pour le rapprochement avec les années ultérieures et pour tout rapprochement avec une source extérieure. Le Siren se déduit des neuf premières positions de cet identifiant. A chaque Siren est attribué un code de qualité.

Tableau 3

Base finale des identifiants

Chroniques	Nombre
sans trou	2 762 057
à un trou	519 212
à deux trous	120 898
à trois trous	35 343
à plus de trois trous	26 946
Total	**3 464 456**

On complète ensuite ce fichier des identifiants à partir des informations fiscales disponibles dans les fichiers annuels, grâce au Siren et à l'IFRP d'origine de Suse que l'on a conservé. Le fichier des données ainsi obtenu ne comporte comme identifiant que le SIRIFRP de manière à alléger le volume des données. Cela signifie que pour compléter la base de données avec d'autres données fiscales que celles proposées, il faut revenir à la base des identifiants, qui gère l'articulation entre le SIRIFRP et les différents identifiants annuels.

Tableau 4

Qualité du Siren final

SIREN	Nombre
bon	3 017 593
double	96 794
non renseigné	348 519
mauvaise clef	1 550
Total	**3 464 456**

Le rapprochement temporel des fichiers annuels ainsi corrigés par le double biais du Siren et de l'IFRP permet ainsi de reconstituer des chroniques d'entreprises. Les absences de dépôts de compte, une ou plusieurs années consécutives, sont décelées par cette mise en cohérence annuelle. Il reste alors à boucler les années terminales de la base : il faut en premier lieu repérer les entreprises cessées ou créées. On utilise principalement un fichier des cessations issu de Sirene : le décalage de deux ans entre la constitution de l'année terminale de la base et l'année courante permet en effet d'avoir la plus grande partie des cessations d'entreprises enregistrées à Sirene. D'autre part, Sirene donne la liste des entreprises créées une année donnée, ce qui permet de repérer les entreprises nouvellement créées qui n'ont pas déposé de compte l'année de leur création. On constate en effet que, par le jeu des durées d'exercice, la moitié seulement des entreprises créées une année donnée dépose des comptes l'année de la création. Si

on veut une estimation des entreprises créées l'année terminale de la base et non seulement des entreprises créées qui ont déposé leurs comptes, il convient de les repérer. Toutefois, cela n'est possible que si on est théoriquement à même d'extraire de Sirene, la liste des entreprises concernées. Une des utilisations futures de la base est de contribuer à cerner les "objets" BIC théoriques, c'est-à-dire les catégories d'entreprises susceptibles d'être assujetties à l'impôt sur les sociétés ou de relever du régime des entrepreneurs individuels. Des travaux réalisés dans le cadre du groupe de travail sur le référentiel des entreprises montrent qu'en pratique la correspondance entre les "objets BIC" et le champ des sociétés et quasi-sociétés sélectionné à partir de Sirene n'est pas claire.

Méthodes d'interpolation

Le deuxième objectif du travail effectué visait à améliorer l'exhaustivité macro-économique de la source Suse et d'assurer une meilleure cohérence avec les autres sources disponibles sur les entreprises pour des variables communes comme le nombre d'entreprises, le montant global des effectifs salariés et de la valeur ajoutée. Le rapprochement temporel effectué permet de repérer les comptes absents une année donnée. Il s'agit dans un deuxième temps de proposer et de tester des méthodes d'interpolation de ces données manquantes. L'avantage de disposer de données longitudinales permet de ne pas se limiter aux simples méthodes d'imputation (on affecte l'entreprise à une classe donnée et on lui attribue le compte moyen de cette classe). Deux méthodes d'interpolation sont proposées : la première, développée ici, répond au souci d'améliorer l'exhaustivité macro-économique annuelle de la source. Elle repose sur le principe de l'interpolation linéaire des données. La deuxième méthode repose sur le principe de l'interpolation économétrique des données et répond à des besoins d'études sur données de Panel ou d'analyse temporelle. Comme elle n'a pas encore été testée de manière satisfaisante, elle fait l'objet d'un simple développement en annexe.

L'interpolation fait face à deux types de données absentes : d'une part, une censure totale pour certaines entreprises dont les comptes manquent intégralement pour une année donnée. D'autre part, pour certaines entreprises (données des forfaits ou des enquêtes annuelles d'entreprises), certaines variables ne sont jamais disponibles. On dispose alors d'une information supplémentaire qui est la présence des autres variables de l'entreprise.

Les deux méthodes nécessitent toutefois des "outils" statistiques communs permettant de repérer les entreprises absentes : la date de création, la date de cessation, l'année de premier compte et de dernier compte de l'entreprise permettent ainsi de calculer un vecteur de trajectoires des entreprises permettant de repérer les absences.

Année de premier (resp. dernier) compte
Codifier l'année de premier compte et l'année de dernier compte est simple : comme leur nom l'indique, ces variables donnent la première et la dernière année pour lesquelles

on a des comptes dans la base. Le seul problème qui peut se poser est pour l'année terminale de la base : l'entreprise qui a des comptes en 1992 aura pour année de dernier compte "99" si elle est toujours vivante à la fin de 1992, "92" si elle est morte au cours de l'année 1992. Il faut donc disposer d'une date de cessation pour distinguer ces deux occurrences.

Date de cessation (année, mois)
La majorité des entreprises absentes une année donnée réapparaît dans les deux années qui suivent. Par conséquent, une entreprise dont on n'a pas reçu les comptes depuis 1988, absente sur les quatre dernières années de la base, peut être considérée avec une marge d'erreur faible comme cessée : moins de 2 % des entreprises réapparaissent après quatre ans d'absence. Toutefois, si on connaît la date du dernier compte déposé, celle-ci n'indique rien sur la date de cessation de l'entreprise.

Il est donc nécessaire d'essayer de retrouver par des sources externes les entreprises cessées au cours des trois dernières années de la base. On fait appel à un fichier des cessations de Sirène pour déterminer les entreprises cessées, sur le champ des entreprises bien sirénisées. Ce fichier couvre correctement les cessations de 1987 à 1992. Pour les Siren non valides, on utilise la date de cessation de l'entreprise disponible dans les fichiers Suse ou dans les fichiers de la DGI. En cas de divergence, on privilégie la date de cessation administrative de Sirene dans la mesure où elle est cohérente avec l'année de dernier compte. Pour les entreprises dont on n'a pas retrouvé la date de cessation et qui sont absentes au moins les quatre dernières années de la base, on leur attribue une date de cessation égale à l'année de dernier compte + 1, sans mois de cessation.

Le problème reste par contre entier pour les entreprises absentes la dernière, les deux dernières, les trois dernières années de la base et non déclarées cessées à Sirène. En les considérant comme cessées, dans le premier cas, on majorerait d'environ 30 % les cessations de l'année, de 15 % dans le deuxième cas, de près de 5 % dans le troisième. En l'absence d'une date de cessation explicite, on les garde en activité.

Date de création (année, mois)
On garde le mois de création de l'entreprise de manière à relier la durée écoulée entre la date de création d'une entreprise et la date de dépôt de son premier compte à la durée du premier exercice. La confrontation des sources, Sirène, Suse et direction générale des Impôts permet d'attribuer une date de création aux entreprises. En cas de divergence, on privilégie la date de création de Sirène en gardant la plus ancienne année déclarée. Cette date de création doit être cohérente avec la date de premier compte de l'entreprise. Pour les entreprises pour lesquelles aucune information n'est disponible, on attribue comme date de création l'année de premier compte – 1 sans mois de création.

Calcul des trajectoires
Les quatre variables précédentes permettent de caractériser la chronique temporelle d'une entreprise et d'en repérer les accidents. Aux indicatrices de présence annuelle

dans la base, on substitue un code annuel à plusieurs modalités synthétisant l'état dans lequel se trouve l'entreprise l'année en question. Ces indicateurs annuels ont pour objet de repérer les années manquantes et d'en faciliter l'estimation. En cas d'absence de compte une année donnée, l'indicateur de trajectoire explicite si l'entreprise peut être considérée comme non créée l'année considérée, absente ou cessée. Ne seront interpolées que les entreprises absentes au plus deux années consécutives. Les séquences de trois "trous" et plus ne seront pas complétées. En effet, de manière cohérente avec le fichier de démographie des entreprises géré par Sirène, on fait l'hypothèse que l'entreprise est en sommeil, et peut donc être considérée comme inactive économiquement.

On procède de même pour le début et la fin de la période. Les entreprises absentes au moins les quatre dernières années de la base seront considérées *de facto* comme cessées. Par contre, en l'absence d'une date de décès, les entreprises dont les comptes manquent les trois dernières années seront considérées comme encore juridiquement actives. Mais, les comptes manquants ne seront pas estimés conformément au principe d'activité économique énoncé ci-dessus.

Pour les absences de début de période, la codification se fait de manière symétrique grâce à la date de création. Pour toutes les années antérieures à la date de création l'entreprise sera déclarée "non créée". Pour les années comprises entre la date de création et la date de premier compte l'entreprise sera considérée comme absente. De nouveau, on n'estimera que les comptes absents au plus deux années consécutives. Il convient de manier ces règles d'estimation de début et de fin de vie avec prudence quand il s'agit d'une grande entreprise susceptible d'avoir fait l'objet d'une restructuration.

Tableau 5

Année de premier compte

Année de premier compte	Nombre
84[1]	1 490 935
85	456 242
86	263 940
87	245 319
88	214 854
89	211 183
90	208 095
91	184 145
92	189 741
Total[2]	**3 464 454**

1. Y compris les années de création antérieures à 1984
2. La différence de deux entreprises avec les tableaux de la première partie vient du traitement de la Poste et de France Télécom : dans la base brute, la Poste est présente sous deux Siren différents à la suite de son changement de forme juridique en 1990. Dans la base interpolée, on a reconstitué un enregistrement unique pour la Poste. Il en est de même pour France Télécom.

Tableau 5 bis

Année de dernier compte

Année de dernier compte	Nombre
84	162 692
85	182 575
86	197 724
87	222 222
88	230 769
89	248 766
90	250 805
91	320 882
92[3]	84 630
99[4]	1 563 389
Total	**3 464 454**

3. Entreprise présente en 92 et morte avant le 01 01 1993
4. Entreprise présente en 92 et encore vivante au 01 01 1993

Tableau 6

Date de création

Année de création	Nombre
avant 84	1 600 748
84	366 312
85	236 813
86	235 080
87	203 521
88	206 262
89	206 204
90	180 087
91	152 984
92	76 443

Tableau 6 bis

Date de cessation

Année de cessation[1]	Nombre
84	40 765
85	140 537
86	155 516
87	193 824
88	201 771
89	221 358
90	150 735
91	158 373
92	185 968
Cessées	**1 448 847**

1. Rappelons que, pour les années antérieures à 1987, on ne dispose pas d'un fichier des cessations. Une estimation par défaut des décès de 1985 peut être obtenue par le nombre d'entreprises dont la date de dernier compte est 1984. Elle minore le nombre de décès de l'année 1985 des entreprises cessées en 1985 mais dont la date de dernier compte est 1985. Il en va de même pour 1986. Pour les années postérieures, la qualité de recensement de cessations est bonne sur le champ des entreprises bien "sirénisées". On sous-estime donc à nouveau légèrement le nombre de cessations intervenues entre 1987 et 1992.

Tableau 7

Trajectoires

Situation de l'entreprise	Années								+
	1984	1985	1986	1987	1988	1989	1990	1991	1992
Pas créée	197 405	1 260 585	1 025 505	821 984	61 5722	406 718	229 431	76 445	
Présente	1 490 954	1 671 501	1 733 433	1 762 959	1 718 125	1 690 170	1 658 705	1 615 951	1 646 894
1 trou	352 594	184 835	172 792	153 029	176 058	181 075	165 797	183 694	161 987
2 trous	51 755	86 116	64 105	61 775	70 626	77 037	85 928	150 085	102 599
3 trous et +	71 748	98 727	123 354	121 718	118 712	110 674	175 881	141 176	96 160
Cessée		162 692	345 267	542 991	765 213	995 982	1 148 714	1 297 105	1 456 816

Interpolation linéaire des données

Cette méthode vise à assurer la représentativité macro-économique de la base, c'est-à-dire en "coupe instantanée". Elle propose une estimation lorsque les comptes manquent au plus sur deux années consécutives. Une entreprise absente une année sera redressée par la demi-somme de ses comptes les années précédentes et futures. Une entreprise absente deux années consécutives sera interpolée respectivement par le tiers et par les deux tiers de ses comptes les années précédant et suivant ces absences [cf. encadré 2].

En début et fin de période, l'interpolation est affinée de manière à tenir compte des entreprises nouvellement créées qui n'ont pas encore déposé leurs comptes. En début de période, on redresse les comptes manquants de ces entreprises en leur affectant les comptes de l'année de premier compte au prorata de la durée d'exercice exercée l'année de la création. Les comptes de l'année de premier compte seront également corrigés pour être ramenés à une durée inférieure ou égale à douze mois. Le but est d'éviter de surestimer l'année de la création en dupliquant des comptes correspondants à une durée d'exercice de plus de douze mois. L'hypothèse sous-jacente est que l'activité des entreprises nouvellement créées ne monte pas en puissance progressivement mais qu'elles trouvent immédiatement leur régime de croisière, ce qui est assez vrai pour les effectifs, mais demande à être validé (grâce à l'enquête Sine, portant sur les entreprises nouvellement créées), pour le chiffre d'affaires ou la valeur ajoutée.

Pour l'année terminale, on court le risque inverse de sous-estimer l'activité globale annuelle à cause des entreprises créées en 1992 et qui ne déposeront de comptes que l'année suivante. Pour corriger ce biais, on fait l'hypothèse que le comportement des entreprises créées en 1992 est le même que celui des entreprises créées en 1991 : on ne ramène donc pas à douze mois en 1992, les comptes 1992 des entreprises créées en 1991 et ayant déposé leur premier compte en 1992. On fait l'hypothèse que ces doubles comptes comblent le déficit d'activité des entreprises créées en 1992 qui ne déposeront leur premier compte qu'en 1993.

Progressivement, on tente donc de passer d'un simple rapprochement annuel de sources administratives à une véritable mesure économique de l'activité de l'année, cohérente avec la vision macro-économique. Ces premiers travaux ouvrent donc la voie à une annualisation des comptes d'entreprises qui consisteraient à ramener à douze mois les comptes afférents à des durées d'exercice inférieures ou supérieures (décalage d'exercice). Un ultime prolongement consisterait à ramener dans le cadre de l'année civile les entreprises qui présentent des dates de clôture différentes du 31 décembre.

Divers tableaux comparant les bases brutes et les bases interpolées figurent dans les pages qui suivent. Le premier constat que l'on peut en tirer est que l'interpolation faite à partir de la base brute est loin d'être négligeable, même si elle est naturellement plus forte en nombre d'entreprises qu'en effectif et surtout en chiffre d'affaires. En effet on redresse surtout des petites entreprises dont le poids macro-économique est faible. Un deuxième constat est que les taux de redressement sont plus forts pour les premières années de l'architecture Suse III, 1989 à 1991 ce qui correspond à l'intuition que l'on a de la qualité de la source ces années-là : elles ont en effet été marquées par la grève de la direction générale des Impôts et par le chevauchement de deux applications "Suse II" et "Suse III" pour les exercices 1988 et 1989. D'autre part, l'année 1984 est caractérisée par la très mauvaise qualité du fichier des forfaits. Non seulement beaucoup de Siren sont invalides mais, également, un grand nombre d'IFRP sont tronqués. La suppression nécessaire des couples Siren-IFRP en double a conduit ainsi à éliminer un grand nombre d'entreprises forfaitaires, ce qui a pour conséquence de gonfler artificiellement le volume des créations d'entreprises de l'année 1984. L'interpolation de début

de période permet de récupérer ces comptes. On ne perd ainsi que les entreprises présentes en 1984 et cessées en 1985 ce qui conduit à une légère sous-estimation des cessations de 1985. Mais garder ces enregistrements erronés qui ne seront présents qu'en 1984, aurait conduit à augmenter à tort le nombre de disparitions de l'année 1985.

Encadré 2

Interpolation linéaire sur les un "trou" et deux "trous"1 - un "trou" :

si les comptes de l'année n manquent, estimation de la variable Var (n)

Var (n) = 1/2 [Var (n-1) + Var (n + 1)]

2 - deux trous :

si les deux années consécutives n et n + 1 manquent, estimation de Var (n) et Var (n + 1) :

Var (n) = Var (n-1) + 1/3 [Var (n + 2) – Var (n-1)]

Var (n + 1) = Var (n-1) + 2/3 [Var (n + 2) – Var (n-1)]

3 - estimation des absences en début de vie de l'entreprise

On estime les comptes manquants entre la date de création et l'année de premier compte de l'entreprise.

si l'année de premier compte = n + 1 et l'année de création = n :

Var (n) = [(12-MOISCR) /DUREX] * Var (n + 1)

Var (n + 1) = [12/DUREX] * Var (n + 1)

4 - correction des créations de 1992

Si l'année de création = "91" et l'année de premier compte = "92 :"

Var (91) = [(12-MOISCR) /DUREX] * Var (92)

Var (92) = [12/DUREX] * Var (91)

avec

MOISCR = mois de création de l'entreprise

DUREX = durée du premier exercice

La comparaison entre la base brute et la base interpolée est insuffisante pour juger de la qualité du redressement effectué. Il est nécessaire de se référer à une source externe, la plus proche possible de la source Suse. Seul le répertoire Sirène est à même de donner le nombre d'entreprises une année donnée. Pour les effectifs et la valeur ajoutée, on prend comme référence les séries macro-économiques des comptes trimestriels. Les niveaux annuels sont calculés par moyenne des niveaux trimestriels. Les concepts d'effectifs diffèrent mais on peut supposer que la marge d'erreur est constante.

Les résultats de ce rapprochement montrent tout d'abord qu'alors que la série brute reflète principalement les accidents qui ont marqué la collecte et la production des données définitives, la série des effectifs corrigés retrouve les inflexions conjoncturelles de la série macro-économique de référence (cf. graphique 1). Toutefois, le graphique 2, qui rapporte les effectifs de la base brute et de la base interpolée à la série des effectifs salariés des comptes trimestriels, met en évidence une rupture 1988/1989 correspondant au changement d'architecture Suse. À cause de choix de correction automatique différents pour les deux architectures, l'estimation de la période 1989 à 1992 est légèrement en deçà de la période précédente. Le deuxième constat sur ce graphique porte sur la moindre fiabilité de l'estimation du début et de la fin de la période : 1984 et 1985 sont surestimés vraisemblablement à cause de la qualité globale de 1984.

Comparaisons des bases brutes et interpolées

Tableau 8

Base brute

Année	Chiffre d'affaires (en milliards de francs)	Effectif (en millions de salariés)	Nombre d'entreprises
1984	8020	12,25	1 490 928
1985	8750	12,17	1 671 496
1986	9092	12,10	1 733 430
1987	9681	12,14	1 762 954
1988	10 439	12,22	1 718 121
1989	11 420	12,09	1 690 169
1990	11 976	12,24	1 658 705
1991	12 664	12,37	1 615 950

Tableau 8 bis

Base interpolée

Année	Chiffre d'affaires (en milliards de francs)	Effectif (en millions de salariés)	Nombre d'entreprises
1984	8 551	13,02	1 895 276
1985	9 110	12,81	1 942 447
1986	9 447	12,70	1 970 327
1987	10 044	12,74	1 978 758
1988	10 936	12,92	1 964 805
1989	12 173	12,96	1 948 281
1990	12 854	13,21	1 910 430
1991	13 465	13,26	1 949 729

Tableau 8 ter

Écart relatif Base interpolée/Base brute

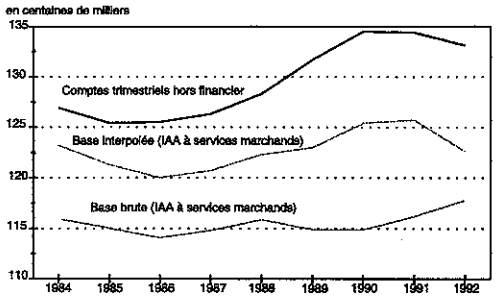
(en %)

Année	Variation relative en chiffre d'affaires	Variation relative en effectif	Variation relative en nombre d'entreprises
1984	6,6	6,3	27,1
1985	4,1	5,2	16,2
1986	3,9	4,9	13,7
1987	3,7	4,9	12,2
1988	4,8	5,7	14,3
1989	6,6	7,2	15,3
1990	7,3	7,9	15,2
1991	6,3	7,2	20,6
1992	3,8	4,3	16,1

Graphique 1

Comparaison de sources

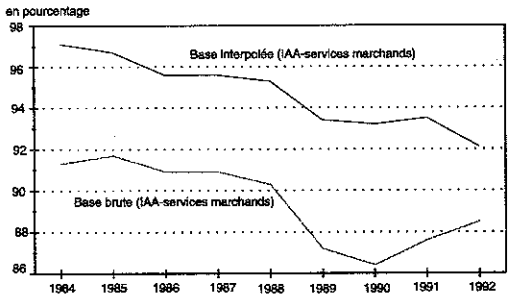
en niveau d'effectif salarié



Graphique 2

Comparaison d'effectifs

Rapport aux comptes trimestriels



Par contre, l'année 1992 paraît légèrement sous-estimée, à cause vraisemblablement du déficit des comptes des entreprises créées en 1992 mais n'ayant déposé leurs comptes qu'en 1993 et ce malgré la correction faite sur l'année 1992. L'ajout de l'année 1993 à la base permettra de tester cette hypothèse.

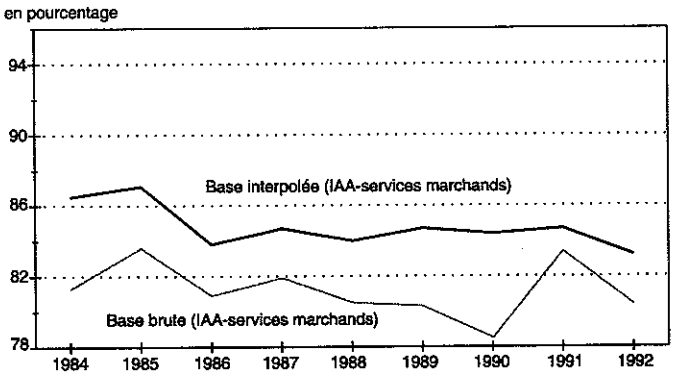
On retrouve des résultats similaires sur la valeur ajoutée : l'interpolation permet en particulier de combler le déficit enregistré dans la base brute les années 1990 et 1991 (cf. graphique 3). L'estimation de 1984 et 1985 est toujours un peu trop forte, celle de 1992 un peu faible. Mais une représentativité macro-économique constante est assurée sur la période 1986 à 1991. L'absence des entreprises assujetties aux Bénéfices Non Commerciaux (BNC) explique l'écart entre l'estimation donnée par la base interpolée qui ne prend en compte que les BIC et les valeurs de la série des comptes trimestriels qui prennent en compte l'ensemble du champ des sociétés et quasi-sociétés.

Par contre, l'estimation du nombre d'entreprises donné par la base diverge profondément de celui de Sirène. À champ comparable (BIC + BNC), la base interpolée ne redonne pas les inflexions de la série du nombre d'entreprises issue de Sirène. Toutes les années sont surévaluées par rapport à Sirène. Cela peut bien sûr provenir de la méthode d'interpolation mais il peut être bon de comparer le champ des entreprises déposant des BIC au champ des entreprises tel qu'il est défini dans Sirène. Un certain nombre "d'objet BIC" comme les loueurs de fonds ne sont en effet pas considérés comme des unités régulières de Sirène alors qu'ils déposent des comptes aux BIC.

Graphique 3

Comparaison de valeur ajoutée

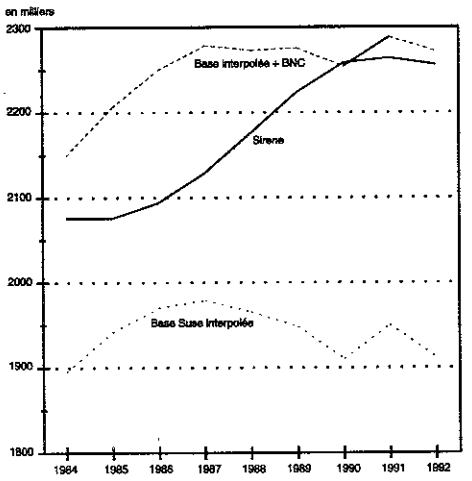
Rapport aux comptes trimestriels



Graphique 4

Comparaison de sources

nombre d'entreprises



Conclusion

L'objet de ce travail était de constituer une base de données temporelles cohérente annuellement avec les estimations macro-économiques. Cet objectif a été en partie réalisé pour des indicateurs de taille comme les effectifs ou la valeur ajoutée, initialisant par là un processus d'annualisation des comptes d'entreprises. Par contre, en ce qui concerne le nombre d'entreprises, qui n'est pas un indicateur qui intéresse la Comptabilité nationale mais qui est une grandeur statistique fondamentale, la méthode d'interpolation proposée conduit à une surestimation du nombre des entreprises une année donnée. D'autre part, la base "brute" d'analyse ouvre la voie à de nombreux travaux statistiques et d'études sur les entreprises. Toutefois, les règles de confidentialité qui régissent les données administratives font que ces données individuelles ne peuvent être diffusées à l'extérieur des services statistiques des ministères. La mise en forme des données convient toutefois très bien à l'élaboration de travaux de cadrage annuels ou de matrices de transition annuelles.

ANNEXE 1

Détail de la méthode de construction des données utilisées

Première étape : Préliminaires

Constitution des deux fichiers de base : un fichier des appariés sur le Siren (SIR8492) et un fichier des appariés sur l'IFRP (FRP8492)

a - Fichier des Siren valides : SIR8492

On apparie sur le Siren et on ne garde que les entreprises qui sont présentes les neuf années avec un Siren toujours valide, soit au total : 3 041 354 entreprises. Parmi elles, 379 780 chroniques sont complètes : elles sont présentes avec un Siren renseigné, unique et identique sur toute la période 1984 à 1992. Elles constituent le sous-fichier nommé SIR8492C dans l'arborescence.

Les chroniques bonnes mais non complètes (fichier SIR8492I), c'est-à-dire pour lesquelles l'entreprise est absente au moins une année en 1984 et 1992, seront complétées grâce à l'IFRP soit en interne, soit à l'aide des chroniques présentant des problèmes. Le rapprochement temporel montre ainsi que près de 320 000 entreprises supplémentaires sont pérennes sur l'ensemble de la période mais sont absentes en cours de période (entre 1985 et 1991) au moins une année, ce qui double le nombre potentiel de pérennes.

b - Fichier des bons IFRP : fichier FRP8492

Il est constitué à partir du complémentaire du fichier SIR8492, par rapport aux fichiers annuels initiaux : toutes les chroniques restantes ont, au moins une année, un Siren non valide, ce qui implique qu'on ne peut pas les apparier sur le Siren. On les apparie donc sur l'IFRP sous réserve de remonter grâce au Siren les chroniques qui changent d'IFRP. Pour cela, on ne garde que les IFRP renseignés et uniques. On obtient ainsi 550 007 chroniques dont l'IFRP est unique. Un peu plus de six mille de ces chroniques sont déjà complètes sur l'IFRP.

Deuxième étape : Complètement des séries bonnes mais incomplètes en interne ([0,b] : SIR8492I)

Dans un deuxième temps on cherche à compléter grâce à l'IFRP les séries bonnes mais incomplètes sur elles-mêmes, c'est-à-dire à partir des mêmes séries bonnes mais incomplètes. Pour cela, on génère pour chaque IFRP rencontré autant d'enregistrements qu'il y a d'entreprises différentes concernées par cet IFRP. On associe à chaque enregistrement IFRP ainsi créé l'intégralité de l'historique de l'entreprise associée. Une entreprise donnée sera donc dupliquée en autant d'enregistrements qu'il y a d'IFRP différents dans sa chronique.

Remarque :

Si un même IFRP est présent plusieurs années dans la chronique d'une même entreprise, l'enregistrement IFRP n'est créé qu'une fois. Par contre, si deux ou plusieurs entreprises différentes sont associées à un même IFRP, on crée autant d'enregistrements qu'il y a de Siren différents.

On attribue alors à chaque enregistrement IFRP une indicatrice de présence annuelle qui vaut 1 si l'entreprise correspondante est présente l'année considérée, 0 sinon. On somme ensuite ces indicatrices par IFRP différents. Les IFRP qui ne sont rencontrés qu'une fois sur toute la période ne sont pas complétables en interne et sont éliminés. Les IFRP rencontrés plusieurs fois correspondent *a priori* à des chroniques potentiellement complétables les unes avec les autres. Toutefois, ne sont réellement complétables que les chroniques associées à des IFRP dont les indicatrices annuelles cumulées ne prennent que la valeur 1 ou 0. Une valeur supérieure à 1 signifie en effet que le même IFRP est associé pour la même année à deux entreprises différentes. Le cas caricatural est celui de 1984 où un grand nombre d'IFRP n'ont été saisis que sur 10 caractères au lieu de 19. Le nombre de doubles est donc considérable. Ces chroniques ne sont pas complétables par l'IFRP. Le fichier des chroniques de Siren complétables en interne sur l'IFRP est intitulé CORIFRP dans l'arborescence.

Pour compléter les chroniques, on dissocie les enregistrements IFRP en double dans ce fichier et on met à jour ces fichiers deux à deux sur l'IFRP. Au total, on obtient 28 042 chroniques complétées en interne, dont 6 800 chroniques complètes (présentes les neuf années). Ces chroniques complétées sur l'IFRP présentent pour caractéristique de ne plus présenter de Siren unique. On leur attribue donc un numéro qui fera office d'identifiant pour la suite.

On crée ensuite le fichier des complémentaires à CORIFRP. Pour cela on repart du fichier des chroniques complétables avant complètement qui contient la liste des Siren de ces chroniques. On peut donc créer le complémentaire de CORIFRP grâce au SIREN à partir de SIR8492I.

On réunit ensuite ces deux fichiers (fichier SIR8492M). Ce fichier n'est plus complétable en interne sur les chroniques valides mais incomplètes mais peut éventuellement être complété grâce au fichier FRP8492.

Troisième étape : Complètement des séries à Siren non valide ([0, 1, 2, 3, b] fichier FRP8492)

On procède de façon symétrique pour compléter sur elles-mêmes les chroniques appariées sur l'IFRP contenues dans le fichier FRP8492. La différence est que, puisque l'IFRP est unique, renseigné et identique, on recherche les chroniques candidates au complètement grâce au Siren. 680 chroniques sont ainsi complétées (fichier CORSIR).

On réunit ensuite le fichier des Siren non valides complétées sur lui-même (fichier CORSIR) et son complémentaire par rapport au fichier initial FRP8492 (fichier NCORSIR). On obtient le fichier FRP8492M. Il est important de souligner que les entreprises de CORSIR comme celles de CORIFRP ne présentent plus ni Siren, ni IFRP unique sur la période. Il faut donc leur attribuer un identifiant spécifique permettant de les différencier.

Quatrième étape : Complètement mutuel des séries incomplètes bonnes et mauvaises

La dernière étape consiste à rapprocher les deux fichiers complétés sur eux-mêmes après les étapes 2 et 3 (fichiers SIR8492M et FRP8492M). La même méthode est utilisée. On choisit de rapprocher ces chroniques sur l'IFRP, c'est-à-dire de compléter les chroniques bonnes mais incomplètes par les autres, mais la méthode est symétrique et donnerait le même résultat en complétant les séries à Siren non valide par les autres sur le Siren.

On duplique donc les IFRP de chaque fichier, en associant à chaque IFRP créé la chronique complète de l'entreprise correspondante. Les deux fichiers sont rapprochés sur l'IFRP, de manière à obtenir la liste des IFRP présents simultanément dans les deux fichiers. Cette liste d'IFRP correspond aux chroniques susceptibles d'être complétées mutuellement. Les enregistrements correspondant à ces IFRP, sélectionnés dans chacun des deux fichiers, sont cumulés de manière à ne conserver que les IFRP ne présentant pas d'indicatrice annuelle supérieure à 1. Chaque enregistrement IFRP de la liste des IFRP complétables est ainsi caractérisé par deux identifiants : son numéro d'appartenance à SIR8492M et son numéro de classement dans FRP8492M. Ces numéros permettent de revenir aux fichiers initiaux et d'en extraire les chroniques non complétables "en externe" : ce sont les fichiers finals SIR8492F et FRP8492F.

Pour les enregistrements correspondant à des IFRP complétables (présents au moins deux fois), on dissocie les multiples et on met à jour chaque fichier ainsi créé deux à deux (par la procédure UPDATE). Le fichier final est le fichier CORALL. Il comprend les chroniques qui ont été complétées en externe : en particulier 1 555 enregistrements de CORIFRP sont ainsi complétées à la fois en interne et en externe et correspondent à l'arrivée à 611 chroniques. 94 829 chroniques bonnes mais incomplètes ont par ailleurs pu être complétées en externe.

On crée alors la base SUSEREF de gestion des identifiants par concaténation des quatre fichiers finals : SIR8492C, CORALL, FRP8492F et SIR8492F. Cette base, complétée manuellement par les chroniques exclues des appariements initiaux, contient environ 3,46 millions d'entreprises contre 3,59 millions pour le total de SIR8492 et de FRP8492. On a donc reconstitué 130 000 chroniques. A la fin de ces quatre étapes, on obtient un peu plus de 396 000 chroniques complètes, c'est-à-dire des entreprises présentes les neuf années.

Les modalités de remontée des chroniques incomplètes

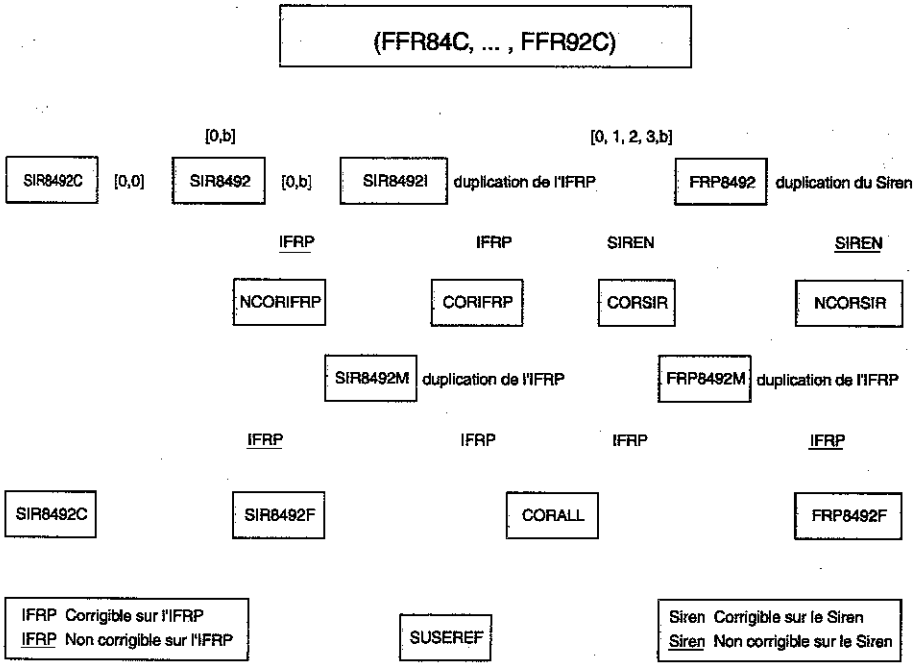


Tableau 9

Bilan des complètements effectués

Origine de la correction	Nombre de chroniques corrigées
Chroniques bonnes et complètes au départ	379 780
Chroniques bonnes, incomplètes, non complétables	2 510 501
Chroniques mauvaises, non complétables	452 110
Chroniques bonnes, complétées sur elles-mêmes seulement	27 431
Chroniques bonnes, complétées par des mauvaises chroniques seulement	94 795
Chroniques mauvaises complétées sur elles-mêmes	599
Chroniques bonnes, complétées en interne et en externe	611

Le tableau précédant donne l'origine des corrections faites sur ces chroniques. Il reste encore de nombreux "trous", mais ayant épuisé l'information disponible on ne peut que considérer qu'il s'agit d'entreprises vraiment absentes. On crée alors un identifiant unique pour chacune des chroniques de SUSEREF, caractérisant chaque entreprise du fichier. Cet identifiant est le produit de la concaténation d'un SIREN et d'un IFRP de l'entreprise. Quand le Siren est unique, on prend ce SIREN et le dernier IFRP renseigné. Si l'IFRP est unique (cf. FRP8492F), on prend cet IFRP et le dernier SIREN bon (rebut à 0) renseigné.

Dans les autres cas, dans un souci de prolongement ultérieur de la base, on prend le Siren valide le plus récent (ou sinon le dernier Siren renseigné) et l'IFRP qui lui est associé.

L'identifiant ainsi défini est unique. Il servira pour le rapprochement avec les années ultérieures et pour tout rapprochement avec une source extérieure (exemple LIFI). Le Siren peut se déduire à partir des neuf premières positions de cet identifiant. Il reste non renseigné pour 379 000 entreprises environ, malgré les rapprochements effectués sur les neuf années. Ce fichier des identifiants contient le SIRIFRP, le Siren qui s'en déduit et pour chaque année le Siren et l'IFRP d'origine de Suse ainsi que le Siren du fichier des redevables permanents associés. À chacun des Siren est attribué un code de qualité.

On complète ensuite ce fichier à partir des informations fiscales disponibles dans les fichiers annuels, grâce au Siren et à l'IFRP d'origine. Le fichier des données en sortie ne comporte comme identifiant que le SIRIFRP de manière à alléger le volume des données. Cela signifie que pour compléter la base de données avec d'autres données fiscales que celles proposées, il faut revenir à la base SUSEREF, qui gère les différents identifiants annuels. Le Siren et l'IFRP de Suse seront utilisés pour tout rapprochement avec les fichiers des BIC.

Remarque sur la qualité de l'année 1984 : l'année 1984 pose des problèmes spécifiques car la qualité des forfaits est très mauvaise. Les Siren sont très mal renseignés mais aussi, les IFRP sont tronqués. La suppression nécessaire des couples Siren-IFRP en double (près de 240 000) conduit ainsi à éliminer un grand nombre d'entreprises forfaitaires conduit à augmenter à tort le nombre de disparitions de l'année 1985. Toutefois, la qualité macro-économique de la source en effectifs ou en chiffre d'affaires est peu affectée par la perte de ces petites entreprises.

Tableau 10

Année 1984	Avant correction	Après correction	Variation relative
Nombre d'entreprises	1 730 025	1 492 215	-14,7%
Chiffre d'affaires (en milliards)	8 049	8 021	-0.3%
Effectif moyen (en milliers)	12 237	12 213	-0.2%

ANNEXE 2

Méthode d'interpolation économétrique

Contrairement à l'approche précédente, cette méthode privilégie une approche individuelle et temporelle des entreprises. La méthode d'interpolation linéaire a pour inconvénient de lisser les fluctuations subies par les entreprises. L'estimation économétrique est plus souple, toutefois elle est très lourde à mettre en œuvre sur de gros volumes de données et elle nécessite d'avoir les données sous forme de panel. C'est pourquoi, elle doit être limitée à estimer les entreprises n'ayant été absente qu'une seule fois dans leur chronique. Les comptes de fin et de début de période ne sont d'autre part pas redressés.

Sur données de Panel

On estime les effectifs et la valeur ajoutée manquants pour les chroniques à "un trou" seulement.

On procède en deux étapes. On estime en premier deux variables "directrices" d'intérêt de l'entreprise toujours demandées. Par exemple, si on élimine les forfaits, les effectifs et la valeur ajoutée :

$$EFF_{it} = \alpha_i + \beta_i\ EFF_{it-1} + \gamma_i\ EFF_{it+1} + \delta_i\ VA_{it-1} + \eta_i\ VA_{it+1} + \varepsilon_{it}$$

$$EFF_{it} = \alpha_i + \beta_i^{'} EFF_{it-1} + \gamma_i^{'}\ EFF_{it+1} + \delta_i^{'}\ VA_{it-1} + \eta_i^{'}\ VA_{it+1} + \varepsilon_{it}^{'}$$

avec u_i, un effet fixe d'entreprise. On estime par la PROC GLM

Les effectifs ou la valeur ajoutée manquants une année donnée sont estimés à partir des effectifs et de la valeur ajoutée non manquants les années précédentes et suivant le "trou", en contrôlant par l'équation estimée sur les entreprises dont tous les effectifs et les différentes valeurs de la valeur ajoutée sont renseignés ces trois années. Un effet fixe d'entreprise permet d'éliminer les biais dus à l'hétérogénéité des données. Les coefficients sont biaisés mais cela n'est pas gênant dans un objectif de prévision.

On estime ensuite toutes les autres variables d'intérêt à partir des estimations réalisées dans la première étape. L'intérêt d'avoir estimé les coefficients pour les variables directrices non manquantes est que l'on peut également estimer dans cette étape les comptes pour lesquels la censure est partielle : par exemple dans les enquêtes annuelles d'entreprises, on ne dispose pas du bilan de l'entreprise. L'actif net est donc systéma-

tiquement manquant pour les Enquêtes isolées (sans enregistrement BIC correspondant). Cette méthode permet d'estimer l'Actif net dans la deuxième étape en même temps que les variables qui sont aléatoirement manquantes.

$$CAHT_{it} = \alpha_i + \beta_i \; CAHT_{it-1} + \gamma_i \; CAHT_{it+1} + \delta_i \, EFF_{it} + \eta_i \; VA_{it+1} + \varepsilon_{it}$$

ou

$$ACTIFN_{it} = \alpha_i + \beta_i \; ACTIFN_{it-1} + \gamma_i \; ACTIFN_{it+1} + \delta_i \, EFF_{it} + \eta_i \; VA_{it+1} + \varepsilon_{it}$$