J.S.C.
Jean Sousselier Conseil

ProFusion

Programme de **F**usion de données

J.S.C.
Jean Sousselier Conseil

25/02/03

SOMMAIRE

1 INTRODUCTION.

1.1 Généralités sur les fusions de fichiers.

Il s'agit ici de fusions « marketing » concernant 2 populations différentes, et non de fusions « informatiques », concernant 2 ensembles de données relatifs aux mêmes individus.

On dispose de 2 fichiers distincts, contenant des individus en nombre quelconque: un fichier **receveur** et un fichier **donneur**.

Les 2 fichiers contiennent un ensemble de **variables communes**, qui serviront à les rapprocher.

Le fichier donneur contient de plus un ensemble de **variables spécifiques** que l'on voudrait reproduire dans le fichier receveur.

Par exemple, le fichier receveur contient 10 000 personnes, le fichier donneur 1.000 personnes. Les variables communes sont des variables socio- démographiques, et les variables que l'on veut reporter sur le receveur sont des variables de comportement d'achat.

Les applications d'une telle technique sont nombreuses, par exemple :

- enquêtes portant sur des questions différentes, que l'on veut pouvoir croiser.
- enrichissement d'une base Clients avec des données d'enquêtes.

1.2 Renseignements pratiques.

- Les 2 parties qui suivent (Méthode et Logiciel) peuvent être lues indépendamment, ce qui explique certaines redites.

- **Installation du programme** : il suffit de copier le fichier fusion.exe sous un répertoire quelconque, et de créer classiquement le raccourci correspondant.

- **Support technique** : s'adresser à :

Jean Sousselier Conseil
48 rue de Longchamp
75116 Paris
Tél : 01 45 05 15 39
e-mail : sousselier.jean@orange.fr

2 METHODE.

2.1 Généralités.

Le logiciel **ProFusion** réalise une telle fusion. Il permet de rechercher pour chaque individu du fichier receveur un **sosie** dans le fichier donneur : on attribuera alors à chaque individu receveur les variables spécifiques de son sosie.

La recherche de sosie est effectuée ainsi, les variables communes comprenant les variables de contrôle et les variables relais, toutes **nominales** :

- les **variables de contrôle:** définissent des cellules à l'intérieur desquelles les sosies seront recherchés. Par exemple, le département et le sexe: le sosie devra être de même sexe et habiter dans le même département.

- Les **variables relais** : pour chacune de ces variables, on utilise une **matrice de distances** (ou pénalités) entre modalités. Par exemple, si une variable relais est l'âge en 5 catégories, on donnera une distance entre les 18-24 ans et les 25-34, 35-49,50-64,65 et +, puis entre les 25-34 et les 35-49, etc.

Le programme va alors chercher, pour chaque individu du fichier receveur, **un sosie** ayant les mêmes modalités pour toutes les variables de contrôle, et tel que **sa distance en termes de variables relais** soit quasi-minimale, tout en veillant à utiliser une grande variété de donneurs.

La distance entre un receveur et un donneur sera **la somme des distances relatives à chaque variable relais**, compte-tenu des modalités respectives du receveur et du donneur concernant cette variable.

Les **voisins** d'un receveur comprendront le donneur le plus proche ainsi que tous les donneurs dont la distance au receveur n'excède pas d'un **seuil S,** fourni par l'utilisateur, la distance minimale correspondant à ce donneur le plus proche.

Le **sosie** sera choisi parmi les voisins du receveur comme celui ayant le moins servi pour les receveurs déjà traités.

2.2 Calcul des distances.

Elles peuvent être soit fournies manuellement par l'utilisateur, soit calculées automatiquement par le programme. Dans ce dernier cas, on utilise **les variables spécifiques** contenues dans le fichier donneur, et que l'on souhaite transférer dans le fichier receveur.

En effet, la distance entre 2 modalités d'une variable relais (par exemple entre les hommes et les femmes), sera d'autant plus grande que leurs profils selon les variables spécifiques sont différents. Autrement dit, si les réponses des hommes et des femmes aux variables spécifiques sont identiques, la distance entre les hommes et les femmes est nulle, et rien n'empêche de choisir une femme comme sosie pour un homme.

En revanche, si le profil des réponses aux questions spécifiques est différent pour les hommes et les femmes, la distance entre ces 2 catégories doit être grande, et le sosie d'un homme devra être choisi parmi les hommes.

Pour calculer la distance **$d_k(i,j)$** entre les modalités **i et j** de la variable relais **k** , le programme construit, pour chaque variable spécifique **s,** le tableau de contingence ayant 2 lignes (correspondant à **i** et **j**) et autant de colonnes **n_s** que de modalités pour la variable **s**.

Le programme calcule alors le Chi2 de ce tableau, qui témoigne, s'il est grand, de la forte probabilité de non-indépendance des lignes et des colonnes ; au contraire, si le Chi2 du tableau est faible, cela témoigne de la forte probabilité d'identité entre les réponses des modalités i et j pour cette variable **s**.

Il faut noter que ce Chi2 est à **$(n_s - 1)$** degrés de liberté.

La distance **$d_k(i,j)$** relativement à **s** est donc **$Chi2_s(i,j)$,** et finalement la distance recherchée est :

$$d_k(i,j) = \Sigma_s[Chi2_s(i,j)]$$

Une première étape du programme consiste donc à calculer ces matrices de distances, en utilisant les variables spécifiques.

2.3 La recherche d'un sosie

Pour un individu receveur **r**, cette recherche se fait en calculant la distance de cet individu à chaque individu donneur **d.**

Pour cela, on considère pour chaque variable relais **k** la distance **$d_k(i,j)$, i** et **j** étant les modalités respectives de **r** et **d** pour cette variable **k,** et on prend comme distance

$$D(r,d) = \Sigma_k\,[d_k(i,j)]$$

Soit **dmin** le donneur le plus proche, tel que **$D(r,dmin) = \min_d [D(r,d)]$**

L'ensemble des donneurs potentiels, ou **voisins**, comprendra tous les donneurs **d** tels que

$$D(r,d) < (1 + S/100)\, D(r,dmin)$$

S étant le seuil de tolérance indiqué par l'utilisateur.
Le donneur retenu sera finalement le voisin ayant le moins servi au moment où l'on traite le receveur **r.**

2.4 Remarques et conseils d'utilisation :

- L'affectation des donneurs peut donc dépendre de l'ordre dans lequel on traite les receveurs. En fait, cette influence joue très peu sur le résultat final.

- En revanche, le choix du seuil **S** est important ; il est recommandé de ne pas le prendre nul, sauf si le fichier donneur est de grande taille par rapport au fichier receveur ; dans tous les autres cas, on pourra choisir ce seuil entre 10% et 30%, en fonction des tailles respectives des 2 fichiers.

- Le fait d'augmenter **S** a l'inconvénient de choisir des donneurs plus éloignés, mais en revanche présente l'avantage d'ouvrir l'éventail des donneurs, augmentant ainsi la variance des variables spécifiques dans le fichier receveur.

- Il n'y a pas de contraintes particulières **quant à la taille des 2 fichiers** : bien entendu, plus le fichier donneur est grand, meilleure sera la fusion, mais rien n'empêche d'avoir un fichier donneur de 1000 individus, et un fichier receveur de 1 million d'individus. Il faut simplement que le fichier donneur soit « assez grand » pour que chaque receveur puisse y trouver un sosie convenable eu égard à l'ensemble des variables relais, ce qui entraîne que plus il y a de variables relais, plus le fichier donneur doit être grand (et meilleure sera la fusion).

- Le fait d'augmenter le fichier donneur présente aussi l'avantage d'augmenter la variance des variables spécifiques dans le fichier receveur.

- L'existence éventuelle de **coefficients de pondération** pour l'un et/ou l'autre fichier n'a aucune incidence sur le processus de fusion, et peut donc être ignorée.

- Les variables de contrôle doivent être choisies en assez petit nombre pour que la taille des cellules définies par le croisement de toutes leurs modalités soit assez grande pour permettre une bonne expansion des variables relais. En général, on choisit une ou deux variables, considérées comme fondamentales eu égard aux transferts projetés.

- Pour les variables relais, en revanche, il faut prendre l'ensemble des autres variables communes aux 2 fichiers dont on dispose ; il n'est pas gênant d'avoir au pire une variable inutile, car sans liaison avec les variables spécifiques.

2.5 Vérifications à effectuer.

Des vérifications sont nécessaires pour s'assurer du bien-fondé de la méthode dans le cas considéré. La qualité de la fusion dépend du nombre et de la pertinence des variables relais dont on dispose. Elle se mesure comparativement à la fusion « aléatoire », obtenue en choisissant au hasard chaque sosie. Deux cas sont à considérer :

- Si on dispose d'un échantillon receveur pour lequel on connaît la valeur des variables à transmettre, alors on effectue la fusion sur cet échantillon et on compare les valeurs des variables transmises et observées.

- Sinon, on partage aléatoirement le fichier donneur en deux : une partie x% du fichier donneur deviendra receveur. La valeur de x dépend de la taille du fichier donneur : pour 1000 à 2000 donneurs, on peut prendre x = 30%, mais cette valeur peut être augmentée si le fichier donneur est grand.

Dans les 2 cas, la comparaison des variables transmises et observées peut se faire ainsi sur le fichier test:

- **les marginaux** de toutes les variables transmises et observées doivent être identiques, dans la limite des intervalles de confiance. Pour les variables quantitatives, on compare les moyennes, pour les variables nominales, on compare les distributions.

- **Les bien classés** doivent ensuite être évalués pour chaque variable spécifique, et comparés à ceux obtenus avec une fusion aléatoire. On a le taux de bien classés pour une variable spécifique en constituant le tableau de contingence croisant la variable transmise et la variable observée ; le taux de bien classés est le pourcentage d'individus se trouvant sur la diagonale.

- **Exemple :** considérons le tableau suivant :

	HABITUDES D'ECOUTE RECONSTITUEES TOTAL RADIO											
HABITUDES D'ECOUTE TOTAL RADIO	TOTAL		1. tous les jours		2. presque tous les jours		3. 1 ou 2 fois par semaine		4. moins souvent jamais		5. jamais	
TOTAL	3491 100,0	100,0	1522 43,6	100,0	685 19,6	100,0	166 4,8	100,0	209 6,0	100,0	909 26,0	100,0
1. tous les jours	1525 100,0	43,7	954 62,6	62,7	338 22,2	49,3	52 3,4	31,3	38 2,5	18,2	143 9,4	15,7
2. presque tous les jours	654 100,0	18,7	331 50,6	21,7	158 24,2	23,1	35 5,4	21,1	35 5,4	16,7	95 14,5	10,5
3. 1 ou 2 fois par semaine	163 100,0	4,7	51 31,3	3,4	34 20,9	5,0	15 9,2	9,0	11 6,7	5,3	52 31,9	5,7
4. moins souvent jamais	240 100,0	6,9	57 23,8	3,7	41 17,1	6,0	15 6,3	9,0	25 10,4	12,0	102 42,5	11,2
5. jamais	909 100,0	26,0	129 14,2	8,5	114 12,5	16,6	49 5,4	29,5	100 11,0	47,8	517 56,9	56,9

On observe d'abord une bonne reconstitution des distributions marginales.

Ensuite, le taux de bien classés est :

(954 + 158 + 15 + 25 + 517)/3491 soit 48%.

Avec une fusion aléatoire, ce taux serait :

$(1525^2 + 654^2 + 163^2 + 240^2 + 909^2) / 3491^2$ soit 30%

Le résultat obtenu est donc satisfaisant.

- Enfin, **la méthode garantit la sauvegarde des duplications** entre les variables spécifiques, puisque elles sont transférées dans leur ensemble entre un donneur et un receveur. En revanche, il peut être bon de vérifier que le résultat du croisement des variables spécifiques et des variables relais est bien conservé par la méthode. Pour cela on considèrera ces croisements sur le fichier test, d'une part avec les variables transmises, d'autre part avec les variables observées. Il faut noter que ces comparaisons sont longues et fastidieuses, on se contentera en général d'une ou deux vérifications considérées comme importantes.

Remarque : on peut être tenté, après avoir fait la fusion, de comparer les variables relais observées dans le fichier receveur, et transmises par la fusion. Il faut bien noter que cette comparaison peut n'être pas judicieuse. Par exemple, si la variable SEXE est sans aucune liaison avec les variables spécifiques, il se peut que l'on transfère à un homme les réponses d'une femme, et donc il n'y aura pas similitude entre les variables observée et transmise.

2.6 Remarques sur le logiciel.

Le logiciel accomplit le calcul des distances et la recherche du sosie pour chaque receveur.

En revanche, le logiciel ne fait pas :

- **le transfert des variables spécifiques** : il se contente d'écrire dans le fichier receveur, pour chacun d'eux, l'identifiant du donneur. L'utilisateur devra alors opérer une fusion informatique classique par clé pour transférer les variables souhaitées (spécifiques ou autres).

- **Les différents contrôles et vérifications décrites ci-dessus :** il faut utiliser pour cela un programme de tabulation classique comme Cosi.

3 LOGICIEL.

3.1 Fichiers traités.

Le logiciel accepte comme données **des fichiers ASCII** (receveur et donneur) non délimités, qui doivent avoir l'extension **.ASC**

La description des données est fournie sous forme d'un script de format très simple, construit avec n'importe quel éditeur. Il s'agit du format de sortie automatique du logiciel de traitement d'enquêtes Cosi (exportation « autres formats) ou CoTab (exportation Voxcotab). Ils doivent avoir comme extension **.SCT**, et la 1ère partie de leur nom doit être identique à celle du fichier de données.

Par exemple, les fichiers receveur sont **XXXX.ASC** et **XXXX.SCT,** et les fichiers donneurs sont **ZZZZ.ASC** et **ZZZZ.SCT .**

3.2 Variables.

Les fichiers de données contiennent dans n'importe quel ordre les variables nécessaires à la fusion : **variables de contrôle, variables relais, et une variable Identifiant**.

Ces variables doivent avoir la même description dans les fichiers donneur et receveur, en particulier même nom, en revanche elles peuvent être situées dans des emplacements différents.

Les variables **spécifiques**, utilisées pour le calcul des distances, ne figurent (ou tout au moins ne sont renseignées) **que dans le fichier donneur.**

La variable Identifiant, de type texte (appelé type Littéral ou Texte dans le fichier script) doit être présente dans les 2 fichiers, mais renseignée uniquement dans le fichier donneur ; le logiciel, après avoir trouvé le sosie pour un individu du fichier receveur, renseignera cette variable dans le fichier receveur avec l'identifiant du sosie trouvé.

Les variables de contrôle et relais, de type nominal (appelé type Simple dans le fichier script), doivent être renseignées dans les 2 fichiers. Leurs valeurs possibles sont des nombres entiers positifs ou nuls, dont on doit fournir dans le script les valeurs minimum et maximum. Toutes les valeurs extérieures à cet intervalle seront classées dans la catégorie conventionnelle "Rebut".

Les variables spécifiques peuvent être de type nominal (Simple) ou Multiple, c'est-à-dire acceptant les multi réponses. Ces réponses sont des nombres entiers positifs, dont on doit fournir dans le script les valeurs minimum et maximum.
Toutes les valeurs extérieures à cet intervalle seront classées dans la catégorie conventionnelle "Rebut".
Dans le fichier de données, ces valeurs doivent apparaître dans n champs consécutifs de longueurs identiques.

3.3 Script.

Si les fichiers sont préparés avec Cosi ou CoTab, on obtient un script comme celui figurant dans les pages suivantes.

Il s'agit d'une description en format libre, les lignes introduites par « ; » sont ignorées.
Il y a 2 sections utiles : [Source] et [Variable]. En fait, Cosi génère d'autres sections, mais qui sont ignorées par ProFusion.

Dans la section Source, on doit indiquer **le nom du fichier de données**, qui en général a le même radical que le fichier script.

Dans la section Variable, on donne la liste de toutes les variables qui seront utilisées dans ProFusion.

Le format général pour **une variable Littérale** (**Texte**) est le suivant : en rouge les éléments à renseigner, en noir les éléments fixes, en bleu les éléments à renseigner facultativement) :

```
V = INERV L(6) "Nø INTERVIEW"
F = Saisie(1-6)
```

La variable INERV, de longueur 6 caractères, est située en position 1 à 6 du fichier. Son titre est N° INTERVIEW.

Pour une **variable de type Simple**, avec 2 valeurs 1 et 2 :

```
V = SEXE S(1-2) "SEXE"
F = Saisie(7)
L = 1 - 2
    1 / Hommes
    2 / Femmes
```

Les libellés des codes 1 et 2 peuvent sans inconvénient n'être pas renseignés.

Pour une **variable de type Multiple**, avec 4 valeurs 1 à 4 :

```
V = ACCES M(1-4) "ACCES INTERNET"
F = Saisie(18 * 4)
L = 1 - 4
    1 / A DOMICILE
    2 / LIEU DE TRAVAIL
    3 / A L'ECOLE-UNIVERSITE
    4 / AUTRE LIEU
```

La variable ACCES est renseignée dans 4 champs démarrant en position 18.

Exemple de script complet généré par Cosi / CoTab :

```
;=================================================================
[Source]
Type = ASCII
Fichier = BARO9B.ASC
;=================================================================
[Variable]
V = INERV L(6) "Nø INTERVIEW"
F = Saisie(1-6)

V = SEXE S(1-2) "SEXE"
F = Saisie(7)
L = 1 - 2
     1 / Hommes
     2 / Femmes
V = AGE S(1-5) "AGE"
F = Saisie(8)
L = 1 - 5
     1 / 15-24 ans
     2 / 25-34 ans
     3 / 35-49 ans
     4 / 50-64 ans
     5 / 65 ans et plus
V = ICSP S(1-9) "PCSI  INDIVIDUS"
F = Saisie(9)
L = 1 - 9
     1 / AGRICULTEURS
     2 / ARTISANS,COMMERCANTS
     3 / CHEFS ETP,CADRES,PROFøINT.SUP.
     4 / PROFESSIONS INTERMEDIAIRES
     5 / EMPLOYES
     6 / OUVRIERS
     7 / RETRAITES
     8 / ETUDIANTS
     9 / AUTRES INACTIFS
V = RUDA S(1-9) "REGION UDA"
F = Saisie(10)
L = 1 - 9
     1 / 1 / Region parisienne
     2 / 2 / Nord
     3 / 3 / Est
     4 / 4 / Bassin parisien Est
     5 / 5 / Bassin parisien Ouest
     6 / 6 / Ouest
     7 / 7 / Sud-Ouest
     8 / 8 / Sud-Est
     9 / 9 / Mediterranee
```

V = HAB7 S(1-7) "HABITAT"
F = Saisie(11)
L = 1 - 7
1 / 1/ COMMUNES RURALES
2 / 2/ AGG. - 20.000
3 / 3/ AGG. 20 A 50.000
4 / 4/ AGG. 50 A 100.000
5 / 5/ AGG. 100 A 200.000
6 / 6/ AGG. 200 ET PLUS
7 / 7/ AGG. DE PARIS

V = MICR S(1-2) "EQUIPEMENT ORDINATEUR"
F = Saisie(12)
L = 1 - 2
1 / Oui
2 / Non
V = CINE12 S(0-1) "CINEMA 12 DERNIERS MOIS"
F = Saisie(13)
L = 0 - 1
0 / non
1 / oui
V = FGCI S(1-3) "FREQUENTATION DU CINEMA"
F = Saisie(14)
L = 1 - 3
1 / 1/ASSIDUS
2 / 2/REGULIERS
3 / 3/OCCASIONNELS
V = IT30 S(0-1) "CONNEXION INTERNET 30 DERNIERS JOURS"
F = Saisie(15)
L = 0 - 1
0 / non
1 / oui

V = HABRA2 S(1-3) "HABITUDES DE FREQUENTATION DE LA RADIO"
F = Saisie(16)
L = 1 - 3
1 / AUDITEURS - 30 MN +non auditeurs
2 / AUDITEURS 30 MN A 3 HEURES
3 / AUDITEURS 3 HEURES ET PLUS

V = HABTV2 S(1-3) "HABITUDES DE FREQUENTATION DE LA TELEVISION"
F = Saisie(17)
L = 1 - 3
1 / AUDITEURS - 30 MN + non auditeurs
2 / AUDITEURS 30 MN A 3 HEURES
3 / AUDITEURS 3 HEURES ET PLUS

V = ACCES M(1-4) "ACCES INTERNET"
F = Saisie(18 * 4)
L = 1 - 4
1 / A DOMICILE
2 / LIEU DE TRAVAIL
3 / A L'ECOLE-UNIVERSITE
4 / AUTRE LIEU

V = LIEU M(1-6) "LIEU DE CONNEXION INTERNET"
F = Saisie(22 * 5)
L = 1 - 6
1 / A DOMICILE
2 / LIEU DE TRAVAIL
3 / A L'ECOLE-COLLEGE-LYCEE
4 / A L'UNIVERSITE
5 / CYBERCAFES,LIEUX PUBLICS,AMIS,PARENTS
6 / AUTRE LIEU

V = FREQ[3] S(1-6) "FREQUENCE DE CONNEXION"
F = Saisie(27 / Pas=1)
O =
1 / AU DOMICILE
2 / SUR LE LIEU DE TRAVAIL
3 / DANS UN AUTRE LIEU
L = 1 - 6
1 / 1/TOUS LES JOURS
2 / 2/PRESQUE TOUS LES JOURS
3 / 3/1 A 2 FOIS PAR SEMAINE
4 / 4/1 A 3 FOIS PAR MOIS
5 / 5/MOINS SOUVENT
6 / 6/JAMAIS

V = DATE[3] S(1-4) "DATE DE DERNIERE CONNEXION"
F = Saisie(30 / Pas=1)
O =
1 / AU DOMICILE
2 / SUR LE LIEU DE TRAVAIL
3 / DANS UN AUTRE LIEU
L = 1 - 4
1 / 1/MOINS DE 30 JOURS
2 / 2/MOINS D'UN AN
3 / 3/PLUS LONGTEMPS
4 / 4/NSP

V = DATEPR S(1-7) "DATE DE PREMIERE CONNEXION"
F = Saisie(33)
L = 1 - 7
1 / 1/IL Y A MOINS D'UNE SEMAINE

2 / 2/IL Y A MOINS D'UN MOIS
3 / 3/IL Y A MOINS DE 3 MOIS
4 / 4/IL Y A MOINS D'UN AN
5 / 5/IL Y A MOINS DE 3 ANS
6 / 6/EN 1997 OU AVANT
7 / 7/NSP

V = ACCINT S(1-2) "Accès à internet à domicile"
F = Saisie(34)
L = 1 - 2
1 / Oui
2 / Non

V = NBV S(1-4) "NOMBRE DE VOITURES DANS LE FOYER"
F = Saisie(35)
L = 1 - 4
1 / 1
2 / 2
3 / 3 OU PLUS
4 / AUCUNE

V = POSS2 M(1-8) "Possession"
F = Saisie(36 * 5)
L = 1 - 8
1 / Camescope
2 / Appareil photo
3 / Lecteur CD
4 / Micro Ondes
5 / Console de jeux
6 / Piano
7 / Lave vaisselle
8 / Résidence secondaire

Remarque : Cosi autorise l'emploi de variables dimensionnées, par exemple

V = DATE[3] S(1-4) "DATE DE DERNIERE CONNEXION"
O =
1 / AU DOMICILE
2 / SUR LE LIEU DE TRAVAIL
3 / DANS UN AUTRE LIEU

dont les 3 occurrences équivalent à 3 variables (domicile, lieu de travail, autre).

Dans le cas où l'on ne passe pas par Cosi / CoTab , on peut générer le script avec un éditeur quelconque, en s'appuyant sur l'exemple ci-dessous :

```
[Source]
Type = ASCII
Fichier = TESTD.ASC

[Variable]
V = NUMPAN L(8)
F = Saisie(1-8)

V = SEXE S(1-2)
F = Saisie(10)

V = AGE S(1-5)
F = Saisie(11)

V = STAT S(1-5)
F = Saisie(12)

V = OCCUP S(1-4)
F = Saisie(13)

V = PCSIA S(1-6)
F = Saisie(14)

V = AGGLO S(1-4)
F = Saisie(15)
```

Chaque variable commune ou spécifique est donc définie par les 2 lignes suivantes, écrites en format libre :

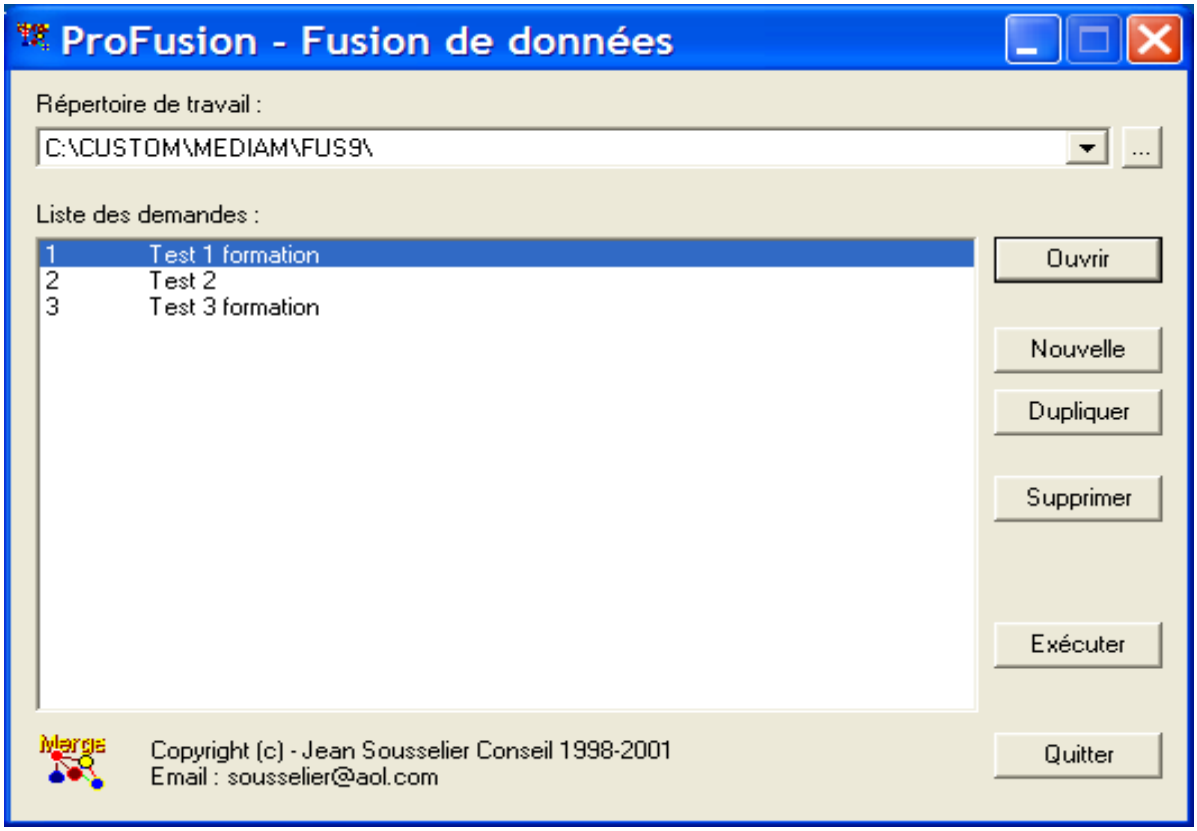
```
V = NOMVAR  S ( i - j )
F = Saisie ( k - l )
```

Avec :

- NOMVAR : nom de la variable (1 à 6 caractères)
- S : type de la variable (S, L ou M)
- i - j : plage de valeurs de la variable de type S ou M
- ou i pour les variables de type L, on donne simplement la longueur
- k - l : position de la variable dans l'enregistrement
- ou k – l * n pour une variable multiple, n désignant le nombre de champs occupés par les réponses.

3.4 Lancement du programme.

Le premier écran est le suivant :



On choisit le répertoire de travail, et on peut créer une nouvelle fusion, et ouvrir, exécuter, dupliquer ou supprimer une étude existante.

Quand on crée une fusion, on lui donne un titre, et un numéro séquentiel sera créé automatiquement.

Sous le répertoire de travail, le logiciel créera les fichiers correspondants :

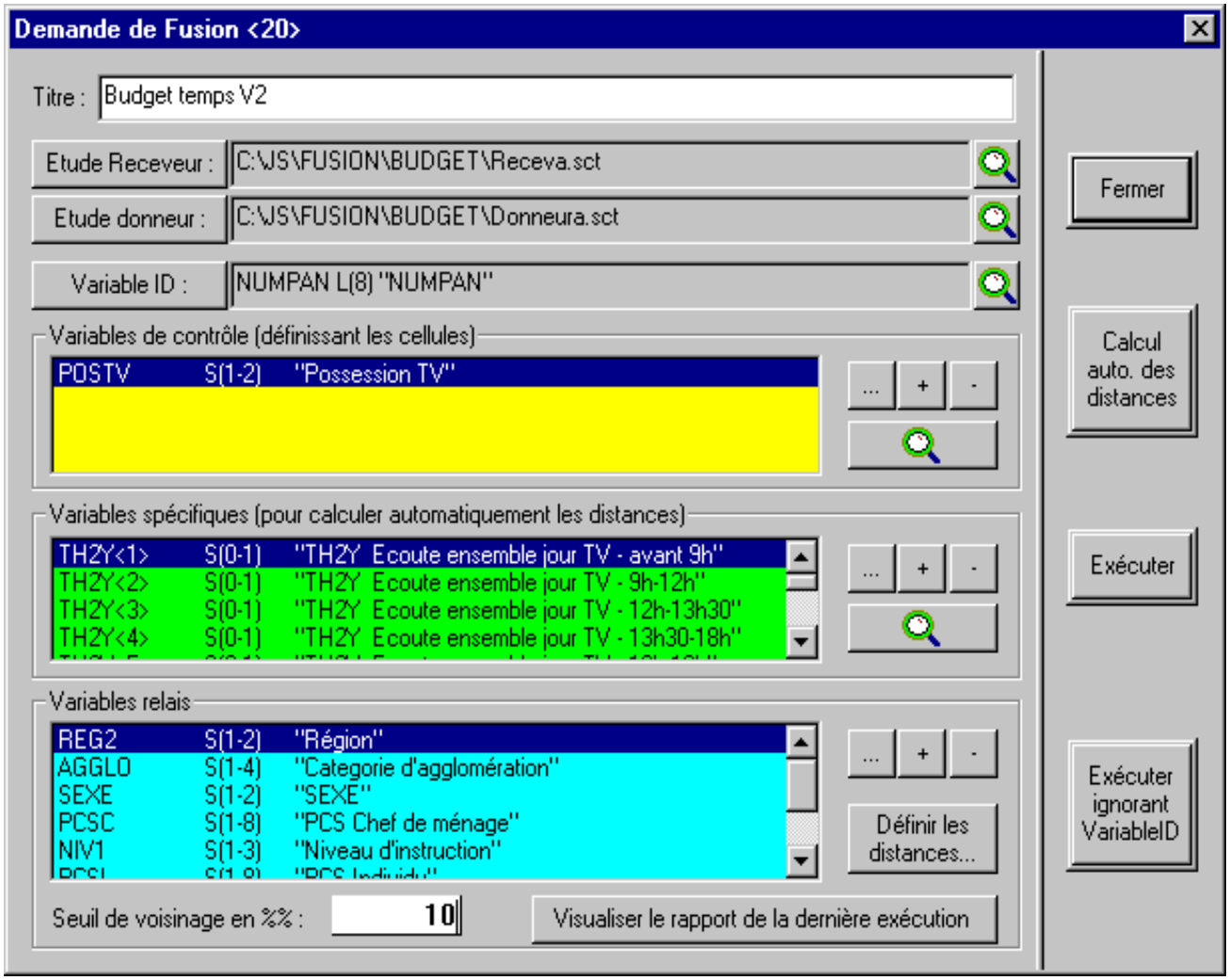
- aux demandes, soit pour l'étude de numéro n : **n.mrg**
- aux résultats, soit pour l'étude de numéro n : **n.log**

(il s'agit de comptes-rendus d'exécution, le résultat principal étant le fichier receveur enrichi du numéro de sosie).

3.5 Spécifications d'une fusion.

Ecran principal.

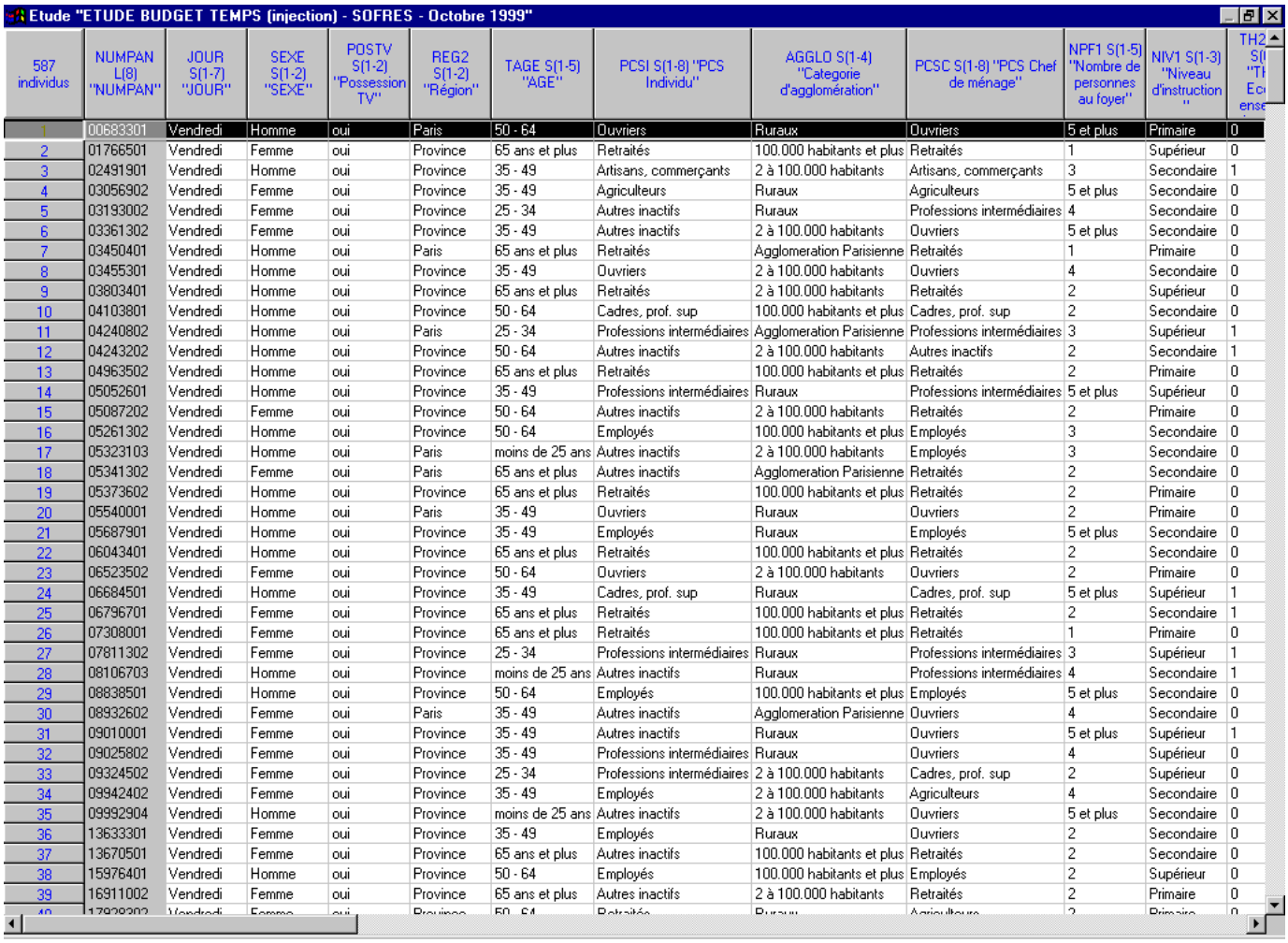
En ouvrant une fusion, ou en créant une nouvelle fusion, l'écran principal est :



Les boutons **« Etude receveur »** et **« Etude donneur »** permettent d'inscrire les noms des fichiers **scripts** correspondants (qui peuvent être sur un autre répertoire que le répertoire de travail).

On ne donne pas le nom des fichiers de données, car ils figurent dans les fichiers scripts.

Les loupes associées permettent de visualiser le contenu de ces fichiers, par exemple :



Etude "ETUDE BUDGET TEMPS (injection) - SOFRES - Octobre 1999"

587 individus	NUMPAN L(8) "NUMPAN"	JOUR S(1-7) "JOUR"	SEXE S(1-2) "SEXE"	POSTV S(1-2) "Possession TV"	REG2 S(1-2) "Région"	TAGE S(1-5) "AGE"	PCSI S(1-8) "PCS Individu"	AGGLO S(1-4) "Categorie d'agglomération"	PCSC S(1-8) "PCS Chef de ménage"	NPF1 S(1-5) "Nombre de personnes au foyer"	NIV1 S(1-3) "Niveau d'instruction "	TH2 S(I "Th Ec ense
1	00683301	Vendredi	Homme	oui	Paris	50 - 64	Ouvriers	Ruraux	Ouvriers	5 et plus	Primaire	0
2	01766501	Vendredi	Femme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	1	Supérieur	0
3	02491901	Vendredi	Homme	oui	Province	35 - 49	Artisans, commerçants	2 à 100.000 habitants	Artisans, commerçants	3	Secondaire	1
4	03056902	Vendredi	Femme	oui	Province	35 - 49	Agriculteurs	Ruraux	Agriculteurs	5 et plus	Secondaire	0
5	03193002	Vendredi	Femme	oui	Province	25 - 34	Autres inactifs	Ruraux	Professions intermédiaires	4	Secondaire	0
6	03361302	Vendredi	Femme	oui	Province	35 - 49	Autres inactifs	2 à 100.000 habitants	Ouvriers	5 et plus	Secondaire	0
7	03450401	Vendredi	Homme	oui	Paris	65 ans et plus	Retraités	Agglomeration Parisienne	Retraités	1	Primaire	0
8	03455301	Vendredi	Homme	oui	Province	35 - 49	Ouvriers	2 à 100.000 habitants	Ouvriers	4	Secondaire	0
9	03803401	Vendredi	Homme	oui	Province	65 ans et plus	Retraités	2 à 100.000 habitants	Retraités	2	Supérieur	0
10	04103801	Vendredi	Homme	oui	Province	50 - 64	Cadres, prof. sup	100.000 habitants et plus	Cadres, prof. sup	2	Secondaire	0
11	04240802	Vendredi	Homme	oui	Paris	25 - 34	Professions intermédiaires	Agglomeration Parisienne	Professions intermédiaires	3	Supérieur	1
12	04243202	Vendredi	Homme	oui	Province	50 - 64	Autres inactifs	2 à 100.000 habitants	Autres inactifs	2	Secondaire	1
13	04963502	Vendredi	Homme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	2	Primaire	0
14	05052601	Vendredi	Homme	oui	Province	35 - 49	Professions intermédiaires	Ruraux	Professions intermédiaires	5 et plus	Supérieur	0
15	05087202	Vendredi	Femme	oui	Province	50 - 64	Autres inactifs	2 à 100.000 habitants	Retraités	2	Primaire	0
16	05261302	Vendredi	Homme	oui	Province	50 - 64	Employés	100.000 habitants et plus	Employés	3	Secondaire	0
17	05323103	Vendredi	Homme	oui	Paris	moins de 25 ans	Autres inactifs	2 à 100.000 habitants	Employés	3	Secondaire	0
18	05341302	Vendredi	Femme	oui	Paris	65 ans et plus	Autres inactifs	Agglomeration Parisienne	Retraités	2	Secondaire	0
19	05373602	Vendredi	Homme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	2	Primaire	0
20	05540001	Vendredi	Homme	oui	Paris	35 - 49	Ouvriers	Ruraux	Ouvriers	2	Primaire	0
21	05687901	Vendredi	Homme	oui	Province	35 - 49	Employés	Ruraux	Employés	5 et plus	Secondaire	0
22	06043401	Vendredi	Homme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	2	Secondaire	0
23	06523502	Vendredi	Femme	oui	Province	50 - 64	Ouvriers	2 à 100.000 habitants	Ouvriers	2	Primaire	0
24	06684501	Vendredi	Homme	oui	Province	35 - 49	Cadres, prof. sup	Ruraux	Cadres, prof. sup	5 et plus	Supérieur	1
25	06796701	Vendredi	Femme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	2	Secondaire	1
26	07308001	Vendredi	Femme	oui	Province	65 ans et plus	Retraités	100.000 habitants et plus	Retraités	1	Primaire	0
27	07811302	Vendredi	Femme	oui	Province	25 - 34	Professions intermédiaires	Ruraux	Professions intermédiaires	3	Supérieur	1
28	08106703	Vendredi	Homme	oui	Province	moins de 25 ans	Autres inactifs	Ruraux	Professions intermédiaires	4	Secondaire	1
29	08838501	Vendredi	Homme	oui	Province	50 - 64	Employés	100.000 habitants et plus	Employés	5 et plus	Secondaire	0
30	08932602	Vendredi	Femme	oui	Paris	35 - 49	Autres inactifs	Agglomeration Parisienne	Ouvriers	4	Secondaire	0
31	09010001	Vendredi	Femme	oui	Province	35 - 49	Autres inactifs	Ruraux	Ouvriers	5 et plus	Supérieur	1
32	09025802	Vendredi	Femme	oui	Province	35 - 49	Professions intermédiaires	Ruraux	Ouvriers	4	Supérieur	0
33	09324502	Vendredi	Femme	oui	Province	25 - 34	Professions intermédiaires	2 à 100.000 habitants	Cadres, prof. sup	2	Supérieur	0
34	09942402	Vendredi	Femme	oui	Province	35 - 49	Employés	2 à 100.000 habitants	Agriculteurs	4	Secondaire	0
35	09992904	Vendredi	Homme	oui	Province	moins de 25 ans	Autres inactifs	2 à 100.000 habitants	Ouvriers	5 et plus	Secondaire	0
36	13633301	Vendredi	Femme	oui	Province	35 - 49	Employés	Ruraux	Ouvriers	2	Secondaire	0
37	13670501	Vendredi	Femme	oui	Province	65 ans et plus	Autres inactifs	100.000 habitants et plus	Retraités	2	Secondaire	0
38	15976401	Vendredi	Homme	oui	Province	50 - 64	Employés	100.000 habitants et plus	Employés	2	Supérieur	0
39	16911002	Vendredi	Femme	oui	Province	65 ans et plus	Autres inactifs	2 à 100.000 habitants	Retraités	2	Primaire	0

En double-cliquant dans le corps du tableau, on fait apparaître alternativement les libellés des variables et leurs codes.

En cliquant sur le nom d'une variable, on fait apparaître le tri à plat de cette variable, classé par ordre alphabétique des libellés (ou codes). On peut demander le tri par ordre décroissant des effectifs en cliquant sur la première ligne de la colonne des effectifs.

En cliquant sur « **Variable ID** », on fait apparaître la liste des variables candidates à servir d'identifiant, à savoir les variables littérales présentes dans les 2 fichiers.

Il faut alors en sélectionner une.

Choix des variables.

On sélectionne ensuite les variables de contrôle, les variables spécifiques (si nécessaire) et les variables relais.

Les variables de contrôle et relais sont choisies parmi les variables simples présentes dans les 2 fichiers.

Les variables spécifiques sont choisies parmi les variables simples ou multiples du fichier donneur.

Pour chacune de ces séries, les boutons :

… : permettent de visualiser la variable sélectionnée
\+ : permettent d'ajouter une variable
\- : permettent de supprimer une variable

Calcul des distances.

On peut soit renseigner manuellement les distances, au moyen du bouton « Définir les distances », soit demander le calcul automatique, au moyen du bouton correspondant.

Dans ce cas, l**a matrice des distances** d'une variable relais est calculée ainsi par le logiciel :

- la distance élémentaire d(i,j,k) entre 2 modalités i et j de cette variable relais relativement à la variable spécifique k est prise égale au chi2 du tableau croisant cette variable spécifique k par les 2 modalités i et j de la variable relais.

- la distance totale d(i,j) entre ces 2 modalités i et j de la variable relais est calculée en sommant les distances élémentaires relatives à toutes les variables spécifiques k.

- l'utilisateur peut s'il le désire modifier manuellement cette matrice de distances.

A tout moment, les distances ayant déjà été calculées automatiquement ou bien manuellement, on peut cliquer sur le bouton « Définir les distances », et modifier manuellement les matrices.

Avant de lancer l'exécution, il faut éventuellement renseigner le seuil de voisinage toléré.

3.6 Exécution.

2 boutons sont disponibles pour lancer l'exécution, l'une complète, l'autre en « ignorant la variable ID ».

Dans ce dernier cas, le sosie n'est pas écrit dans le fichier receveur, on dispose simplement du compte-rendu d'exécution (option utilisée pour faire des tests).

Ce compte-rendu se présente ainsi :

```
Rapport d'exécution : Demande N° 3 (26/02/2003 16:30:18)
**********************************************************

RECEVEURS.................:      8000  100.0%
----------------------------------------------------
  AVEC SOSIE..............:      8000  100.0%
    1 voisin..............:      5311   66.4%
    2 voisins.............:      1640   20.5%
    3 voisins.............:       607    7.6%
    4 voisins.............:       271    3.4%
    5 voisins.............:        93    1.2%
    6 voisins.............:        33    0.4%
    7 voisins.............:        15    0.2%
    8 voisins.............:         5    0.1%
    9 voisins.............:         2    0.0%
    10 voisins............:         2    0.0%
    11 à 15 voisins.......:        21    0.3%

DONNEURS..................:      3000  100.0%
----------------------------------------------------
  NON SERVI...............:       270    9.0%

  servi 1 fois............:       644   21.5%
  servi 2 fois............:       773   25.8%
  servi 3 fois............:       520   17.3%
  servi 4 fois............:       336   11.2%
  servi 5 fois............:       194    6.5%
  servi 6 fois............:       115    3.8%
  servi 7 fois............:        62    2.1%
  servi 8 fois............:        43    1.4%
  servi 9 fois............:        18    0.6%
  servi 10 fois...........:        10    0.3%
  servi 11 à 15 fois......:        13    0.4%
  servi 16 à 20 fois......:         2    0.1%
**********************************************************
```

<< RAPPEL DE LA DEMANDE >>

[MERGE]
TITLE=Test 3 formation
STUDY_R=C:\CUSTOM\MEDIAM\FUS9\RATV9B.SCT
STUDY_D=C:\CUSTOM\MEDIAM\FUS9\BARO9B.SCT
VAR_ID=INERV
VARS_CELL=IT30
VARS_SPEC=FREQ<1>,FREQ<2>,FREQ<3>,DATE<1>,DATE<2>,DATE<3>,DATEPR,
ACCES,LIEU,POSS2
VARS_DIST=RUDA,SEXE,FGCI,ICSP,HABTV2,HABRA2,MICR,AGE,CINE12,HAB7
SEUIL=10.0

[RUDA]
CODE=1-9
1-2=58.6
1-3=77.7
1-4=59.1
1-5=91.1
1-6=91.1
1-7=75.8
1-8=86.3
1-9=47.7
2-3=41.3
2-4=41.3
2-5=39.3
2-6=49.6
2-7=59.1
2-8=28.2
2-9=38.2
3-4=35.2
3-5=40.2
3-6=21.6
3-7=37.2
3-8=33.3
3-9=33.1
4-5=63.1
4-6=42.2
4-7=54.3
4-8=43.5
4-9=34.6
5-6=40.7
5-7=38.5
5-8=25.0
5-9=44.5
6-7=33.7
6-8=26.4

6-9=36.3
7-8=41.6
7-9=34.9
8-9=32.9

[SEXE]
CODE=1-2
1-2=140.6

etc.

Il donne des statistiques sur le déroulement de la fusion (nombre de voisins pour les receveurs, nombre de fois qu'ont servi les donneurs), ainsi qu'un rappel de la demande et les matrices de distances.

Quand on demande l'exécution complète, le logiciel **inscrit pour chaque receveur l'identifiant de son sosie**.

Il n'y a pas de copie automatique des variables spécifiques correspondantes ; cette opération peut être effectuée par l'utilisateur au moyen d'une fusion de fichiers informatiques classique, puisque les clés de fusion sont connues et renseignées dans les 2 fichiers.

Remarque : avant de lancer l'exécution, il est conseillé de **recalculer systématiquement les distances,** car si l'on a modifié la demande en termes de variables de contrôle, relais ou spécifiques, les matrices ne sont pas à jour.

3.7 Limites:

- nombre maximum de variables de contrôle : 32
- nombre maximum de variables spécifiques : 256
- nombre maximum de variables relais : 256

4 UTILISATION

Ce qui suit s'applique à toute fusion, qu'il s'agisse d'un test au moyen d'un échantillon receveur pour lequel on connaît les valeurs des variables à transmettre (il s'agit éventuellement d'une partie du fichier donneur), ou bien qu'il s'agisse de la fusion réelle à effectuer.

4.1 Utilisation avec Cosi ou CoTab.

On crée d'abord les 2 études receveur et donneur, et on obtient les fichiers données et scripts au moyen du menu « Exportation ». Attention : les variables multiples doivent être **exportées par valeur et non par rang**.

Il faut prendre garde que la variable identifiant figure dans les 2 fichiers, mais ne doit être renseignée que dans le fichier donneur, et que toutes les variables communes (de contrôle ou spécifiques) aient la même description dans les 2 fichiers.

Après avoir réalisé la fusion, on peut créer une nouvelle étude Cosi ou CoTab à partir du script, en utilisant le fichier receveur, puis on récupère les variables spécifiques en fusionnant au sens Cosi cette étude avec l'étude « donneurs », ou pour CoTab en utilisant le fichier donneur comme table externe : attention, **il ne faut récupérer que les variables spécifiques**, et non les variables communes qui écraseraient les variables origines des receveurs.

4.2 Utilisation « manuelle » (sans Cosi ni CoTab)

L'utilisateur crée avec un éditeur les 2 scripts à partir des descriptions des 2 fichiers.
Il doit prendre soin de nommer ces fichiers : receveur.asc et receveur.sct, donneur.asc et donneur.sct.

Après avoir fait la fusion, il peut constituer un **fichier receveur enrichi** des variables à transmettre de la façon suivante, en utilisant les fonctions du gestionnaire de fichiers de Cosi, qui est libre d'accès :

- indexer le fichier donneur à partir des positions de la variable Identifiant
- fusionner avec comme clés les positions de la variable Identifiant, les fichiers receveur et donneur pour obtenir un nouveau fichier receveur : newrecev.asc
- composer le script correspondant newrecev.sct de la façon suivante :
    - ajouter dans receveur.sct les variables de donneur.sct
    - modifier le nom des variables communes ainsi ajoutées pour éviter la confusion avec les variables origine (par exemple, commencer par la lettre Z)
    - modifier les positions de lecture en leur ajoutant la longueur initiale du fichier receveur.
    - modifier dans newrecev.sct le nom du fichier de données, soit newrecev.asc

L'utilisateur pourra alors initialiser une nouvelle fusion en indiquant simplement le nom de ce nouveau fichier receveur, ce qui lui permettra de vérifier les valeurs des variables transmises.

5 EXEMPLE

Etude SOFRES Budget-temps (environ 8 000 individus 15 ans et +)

Le **questionnaire** comprenait en particulier les rubriques suivantes**:**
- signalétique
- présence à domicile par tranches horaires et jours
- habitudes radio-TV
- écoutes radio-TV par quart d'heure pour la veille

Le **problème** est la reconstitution des écoutes radio-TV pour les autres jours de la semaine.

La **méthode** consiste à:
- prendre comme fichier donneur les répondants pour un jour donné (Vendredi par exemple)
- prendre comme fichier receveur les répondants des autres jours

Variables choisies :
- variable de contrôle:
  - possession TV
- variables relais :
  - SEXE
  - AGE
  - Habitudes TV pour le jour à reconstituer.
  - Présence à domicile pour le jour à reconstituer.
- variables spécifiques :
  - écoute radio-TV par quart d'heure

Les **résultats obtenus** ont fait l'objet des contrôles suivants:

Contrôle sur le vendredi.

Constitution du fichier donneur en prenant un individu sur deux au hasard, parmi ceux interrogés sur le Vendredi. Les 8 000 individus sont receveurs.

Comparaison des **résultats calculés** avec les **résultats observés** sur les individus interrogés sur le Vendredi.
Comparaison des résultats de la méthode choisie (méthode complète) avec ceux obtenus avec 2 méthodes simplistes :

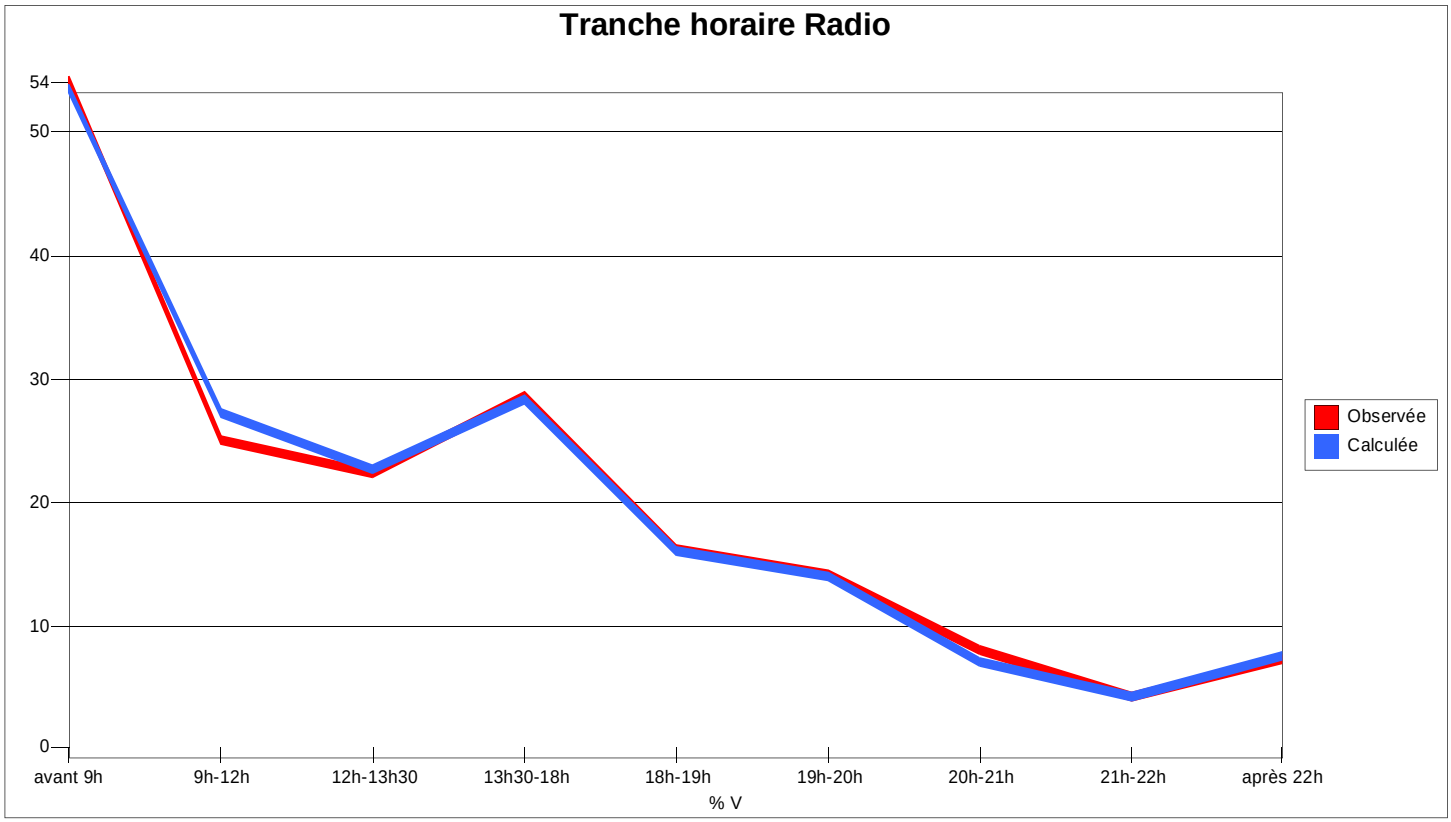
- méthode SEXE-REG-AGE en prenant comme variables relais uniquement SEXE, REGION et AGE

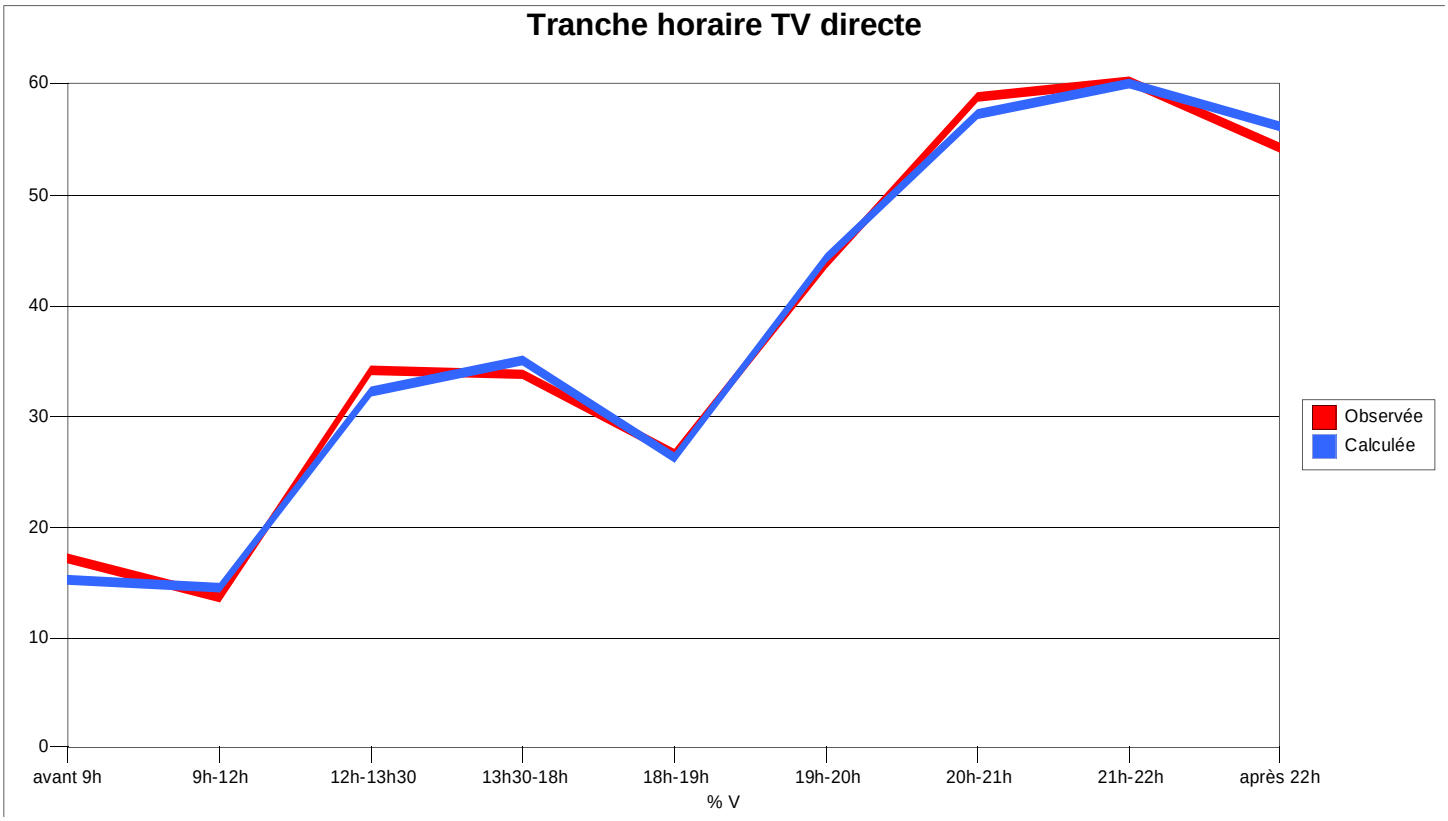
- méthode HASARD : neutralisation complète des variables relais, tous les donneurs étant équidistants de chaque receveur.

Ci-après, figurent quelques courbes comparant résultats observés et calculés dans les différents cas.

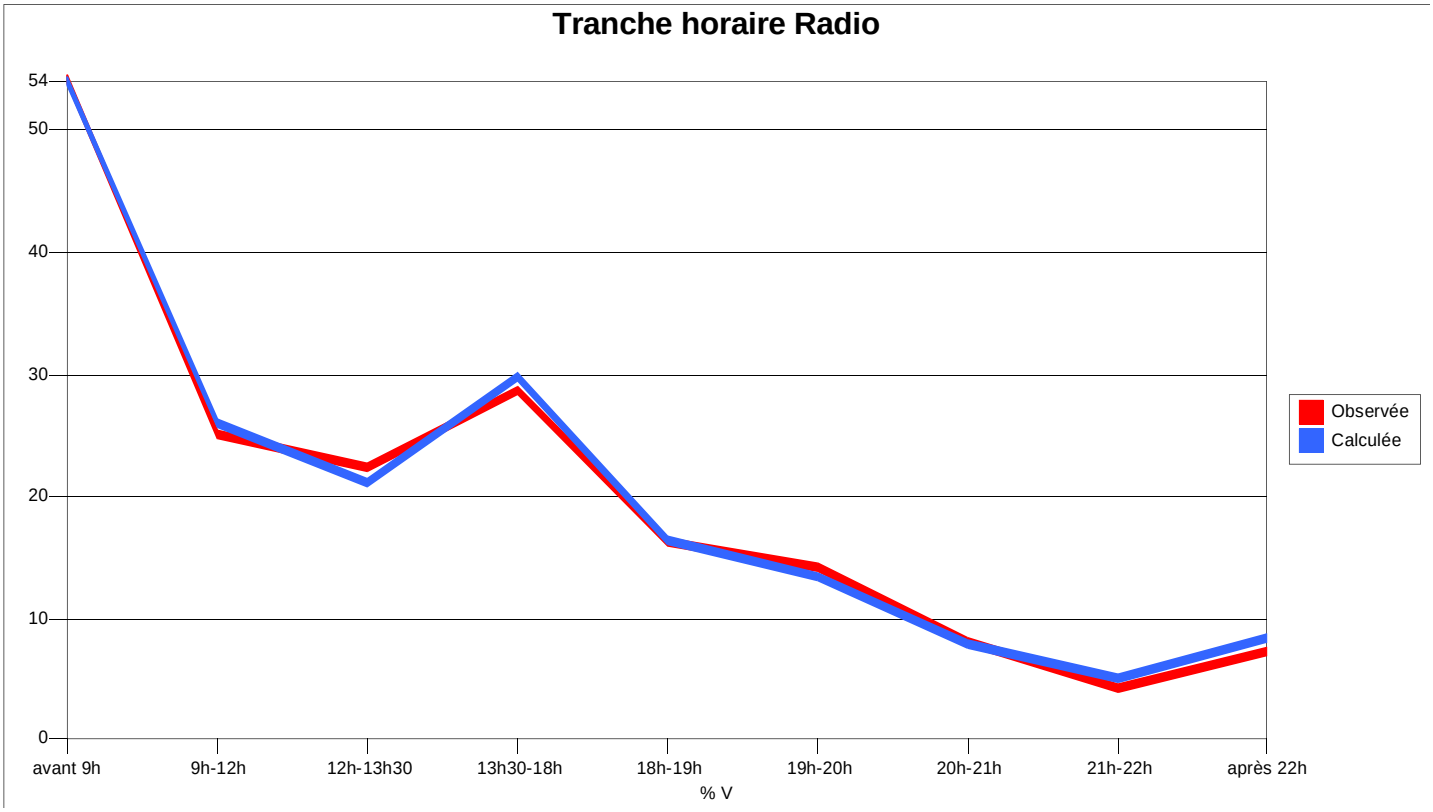
Il s'agit des écoutes cumulées par tranches horaires, pour la Radio et la TV en général.

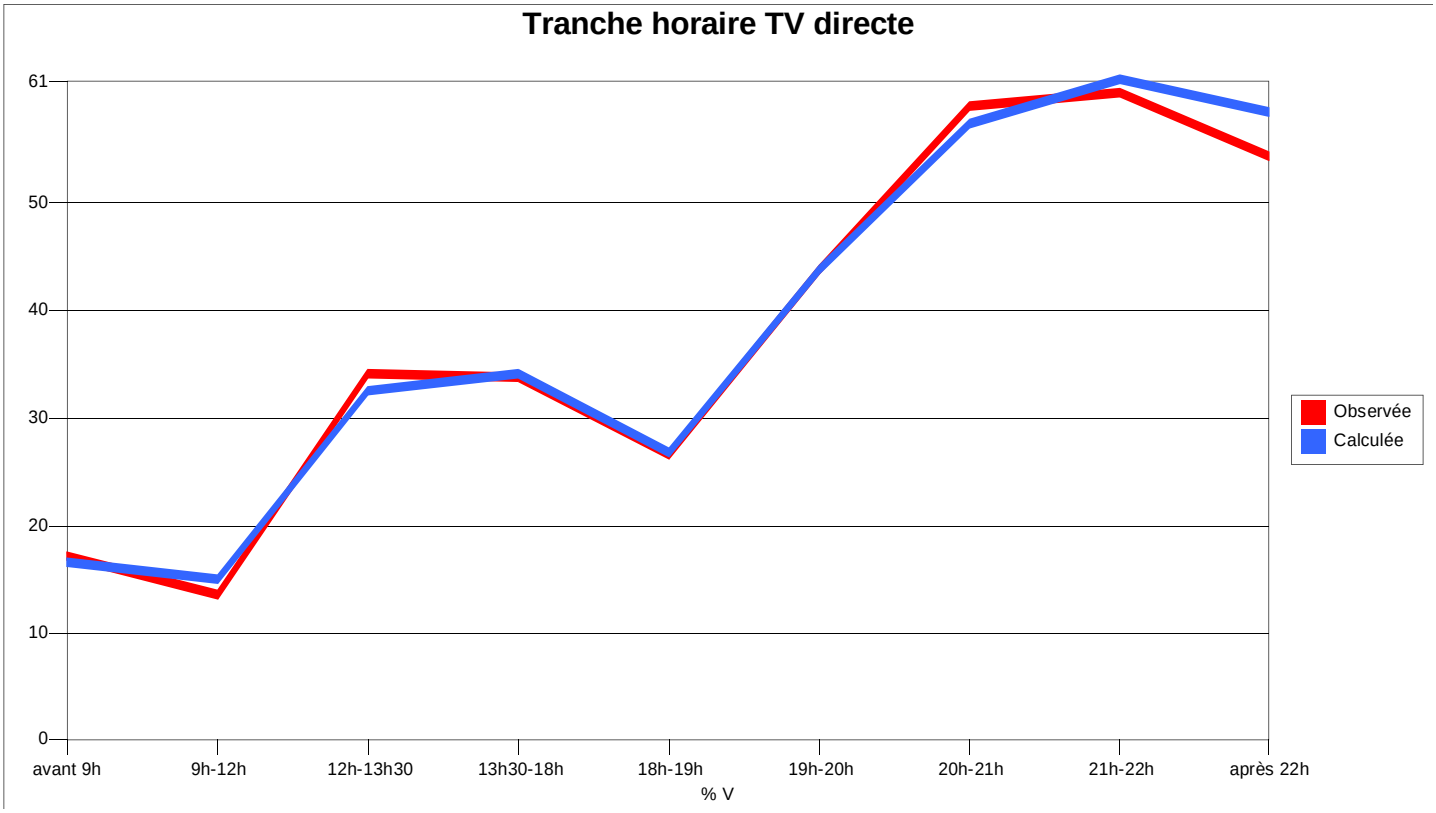
Méthode COMPLETE :





Méthode HASARD:





Les résultats sont satisfaisants, mais on constate qu'ils le sont aussi bien pour la méthode au hasard que pour la méthode complète ! En fait, cela n'est pas surprenant: la méthode au hasard revient en fait à reproduire au hasard la quasi totalité des donneurs, compte- tenu de ce que l'on recherche à chaque fois le voisin ayant le moins servi.

Pour apprécier plus précisément la qualité des reconstitutions des 3 méthodes, il ne suffit pas d'examiner les **résultats marginaux**, ni même les **corrélations entre les variables transférées**, puisque la méthode des sosies sauvegarde ces corrélations.

En revanche, on peut essayer d'appréhender les **bien classés individuellement**.

Pour ce faire on a construit l'indicateur ci-après: pourcentage du nombre de tranches horaires communes (observées et calculées) par rapport au nombre de tranches horaires écoutées (observées).

Cet indicateur témoigne de la similitude individuelle des écoutes observées et calculées. On constate alors (tableau ci-dessous), que ces similitudes ne sont **à peu près correctes qu'avec la méthode complète.**

Répartition des individus selon leur similitudes entre écoutes TV par tranches horaires observées et calculées (en %).

	Non auditeurs observés et calculés	Auditeurs calculés, non observés	Auditeurs observés, non calculés	1-25% de valeurs communes	26-50% de valeurs communes	51-75% de valeurs communes	76-100% de valeurs communes
Méthode complète	**12**	7	6	5	9	8	**53**
Sexe, reg et age	**5**	13	13	15	17	15	**22**
Hasard	**4**	14	13	16	17	13	**22**

Les bien classés correspondent à la première et à la dernière colonnes :

- la première colonne contient les non auditeurs réels et reconstruits.
- la dernière colonne contient les receveurs pour lesquels plus des trois quarts des variables transférées sont égaux aux variables réelles.

On voit donc que le pourcentage de correctement classés est de **65%** pour la méthode complète, et de **27 ou 26%** seulement pour les autres méthodes.

6 CONCLUSION.

Avec **ProFusion,** le chercheur ou le chargé d'études dispose d'un logiciel simple à apprendre, facile à utiliser, et performant en terme de résultats.

En présence d'une recherche se prêtant bien à une fusion comme décrit ci-dessus, il peut donc tester plusieurs hypothèses, faire plusieurs essais, afin de construire "le meilleur" fichier possible : choix des variables, des distances, etc..

Il sera assuré d'une bonne fiabilité des résultats, comme le montre l'exemple ci-dessus.
