J.S.C.
Jean Sousselier Conseil

ProFusion

Programme de Fusion de données

1. INTRODUCTION.

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.

2. METHODE.

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:

- 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.

3. LOGICIEL.

Le logiciel accepte comme données **des fichiers ASCII** (receveur et donneur) non délimités.

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, ou CoTab.

Exemple d'une telle description:

```
[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)
- i - j : plage de valeurs de la variable
- k - l : position de la variable dans l'enregistrement

Les variables communes doivent porter le même nom dans les fichiers donneur et receveur.

Les variables communes comme les variables spécifiques doivent être des variables Simples (S), c'est à dire nominales, dont les valeurs sont des nombres entiers allant de i à j . Toutes les valeurs extérieures à cet intervalle seront classées dans la catégorie conventionnelle "Rebut".

Enfin, les 2 fichiers doivent contenir une variable Identifiant, de type Littéral (NUMPAN dans l'exemple précédent) L(8) signifie que sa longueur est de 8 caractères ; cette variable doit être renseignée dans le fichier donneur ; le logiciel, après avoir trouvé le sosie pour un individu du fichier receveur, renseignera cette variable avec l'identifiant de son sosie.

La 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.

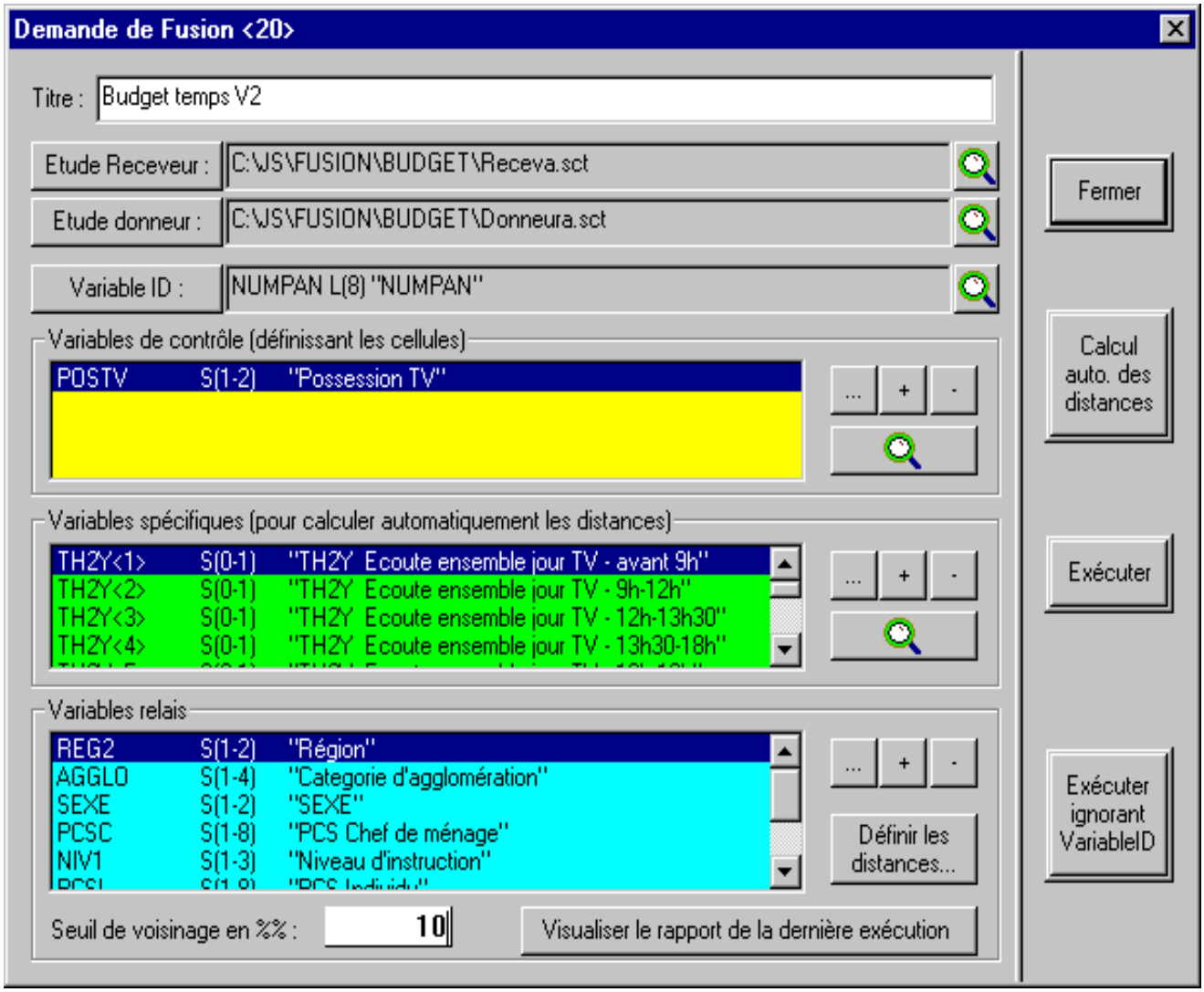
- l'utilisateur peut s'il le désire modifier manuellement cette matrice de distances.

Il faut noter que le logiciel **inscrit pour chaque receveur l'identifiant de son sosie**, sans 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.

Le logiciel est d'une **utilisation facile et conviviale**: choix des variables spécifiques, relais, de contrôle, etc, permettant de faire de nombreux essais si nécessaire sur les mêmes fichiers. Des exemples d'écrans sont donnés ci-après.

Ses limites sont les suivantes :

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


Demande de Fusion <20>
Titre : Budget temps V2
Etude Receveur : C:\JS\FUSION\BUDGET\Receva.sct
Etude donneur : C:\JS\FUSION\BUDGET\Donneura.sct
Variable ID : NUMPAN L(8) "NUMPAN"
Variables de contrôle (définissant les cellules)
POSTV S(1-2) "Possession TV"
Variables spécifiques (pour calculer automatiquement les distances)
TH2Y<1> S(0-1) "TH2Y Ecoute ensemble jour TV - avant 9h"
TH2Y<2> S(0-1) "TH2Y Ecoute ensemble jour TV - 9h-12h"
TH2Y<3> S(0-1) "TH2Y Ecoute ensemble jour TV - 12h-13h30"
TH2Y<4> S(0-1) "TH2Y Ecoute ensemble jour TV - 13h30-18h"
Variables relais
REG2 S(1-2) "Région"
AGGLO S(1-4) "Categorie d'agglomération"
SEXE S(1-2) "SEXE"
PCSC S(1-8) "PCS Chef de ménage"
NIV1 S(1-3) "Niveau d'instruction"
Définir les distances...
Seuil de voisinage en %% : 10
Visualiser le rapport de la dernière exécution
Fermer
Calcul auto. des distances
Exécuter
Exécuter ignorant VariableID

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 "TI 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

4. 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 :

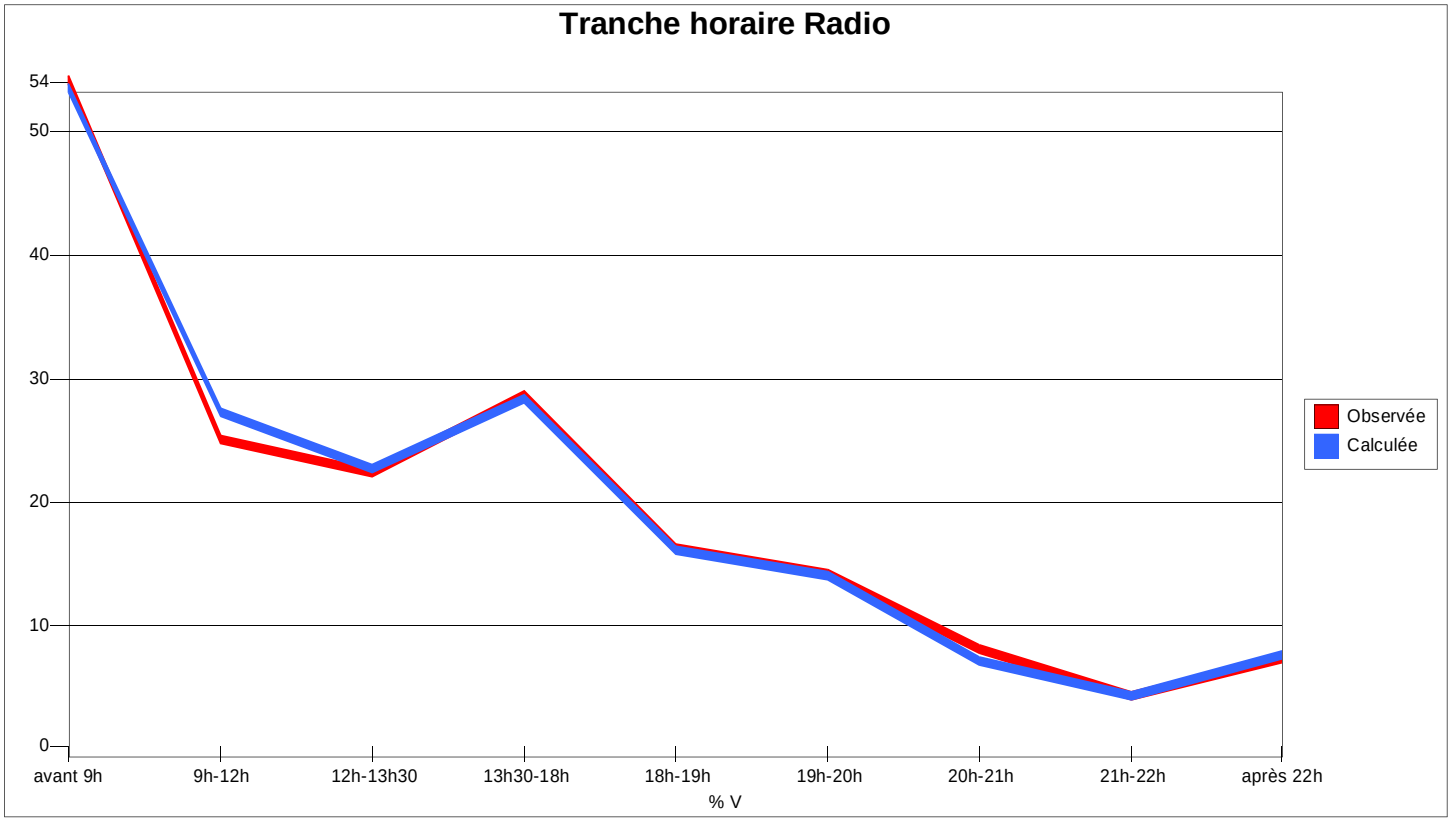
- méthode SEXE-REG-AGE en prenant comme variables relais uniquement SEXE, REGION et AGE

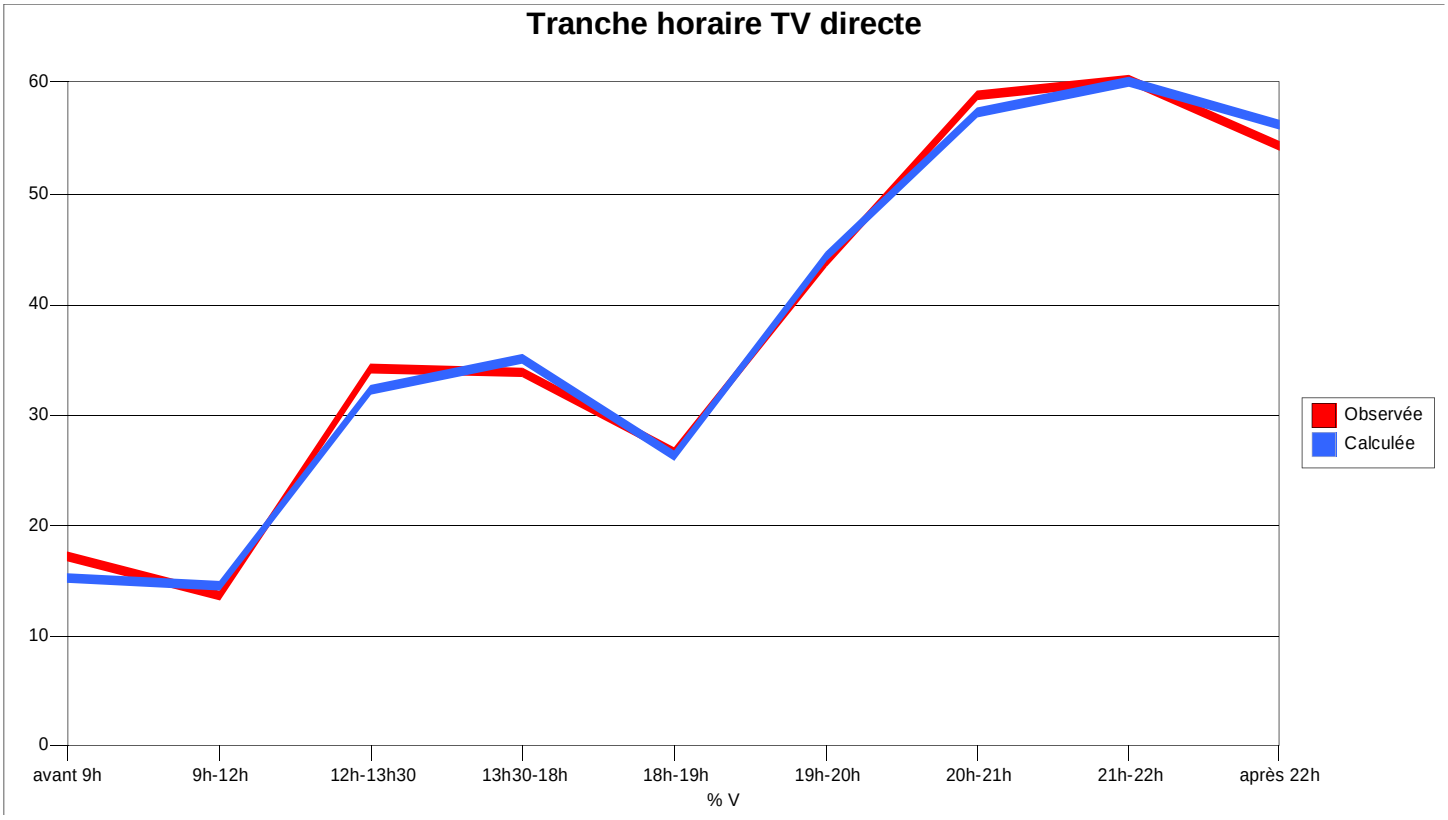
- méthode HASARD : neutralisation complète des variables relais, tous les donneurs étant équidistants de chaque receveur.

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

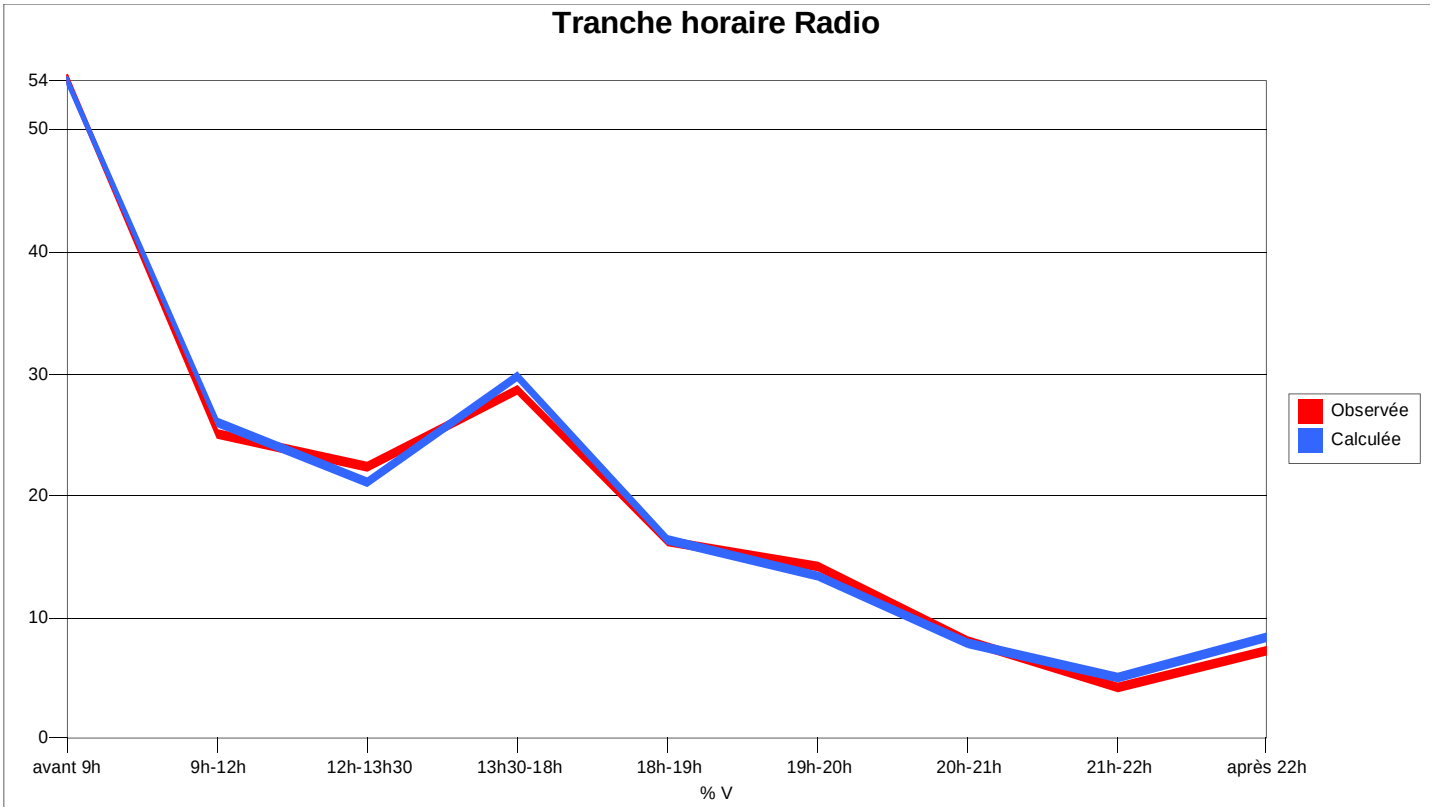
Il s'agit des écoutes cumulées par tranches horaires, pour la Radio et la TV en général.

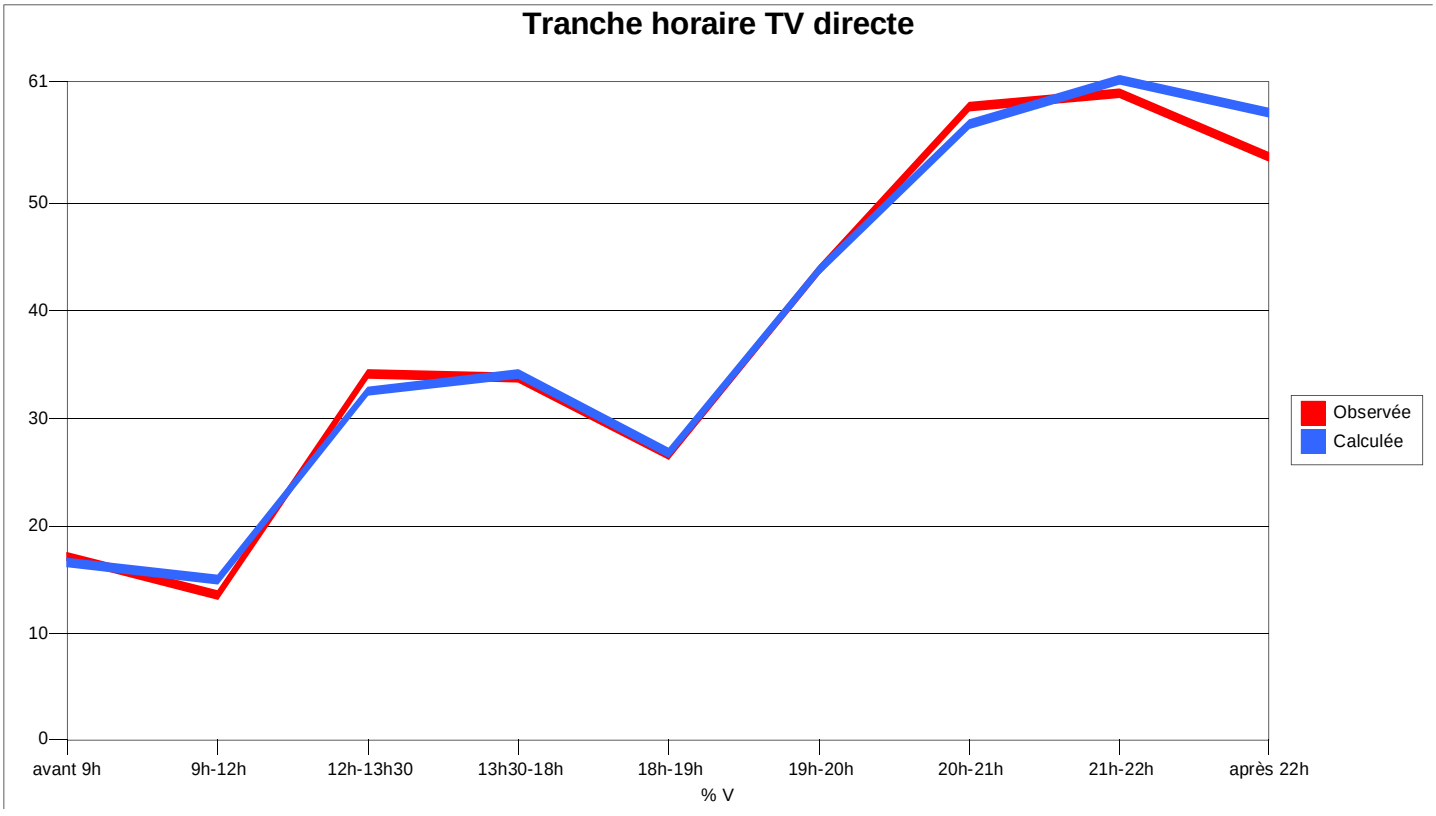
METHODE COMPLETE :





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 égales 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.

5. 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.
