

SECRET STATISTIQUE : L'APPORT D'ARGUS

L. VIGLINO

INSEE, Direction de la coordination statistique
et des relations internationales

Argus est un logiciel de protection de l'anonymat lors de la diffusion de données statistiques.

Argus comprend deux modules : Tau pour les tableaux et Mu pour les fichiers. Ces deux modules suppriment des informations pour respecter les règles de secret statistique. Ils intègrent des algorithmes de minimisation des suppressions effectuées.

Les versions actuelles d'Argus (2.0 pour Tau et 3.0 pour Mu) ont été développées dans le cadre d'un projet européen du quatrième programme de travail.

Matteo Fischetti de l'université de Padoue, Alberto, Caprara de l'université de Bologne et Juan-Jodé Salazar de l'université de Ténériffe ont créé le sous-programme d'optimisation de τ -Argus. Serge Tiourine, de l'université d'Eindhoven s'est occupé de celui de μ -Argus.

Les versions ultérieures d'Argus, sous la supervision d'Anco Hundepool au CBS, sont prévues pour la fin de l'année 2002.

Les manuels d'Argus sont disponibles sur le site Web de CBS (cf. références).

La protection d'un tableau

Pour protéger l'anonymat dans un tableau, les statisticiens demandent qu'une case soit construite à partir d'un nombre minimal d'individus, et qu'aucun ne soit prépondérant dans la valeur diffusée.

La méthode classique pour satisfaire cette contrainte consiste à agréger les modalités des variables.

Argus détecte les cases du tableau qui ne respectent pas les règles de secret et les efface, ce sont les suppressions primaires. Ces suppressions ne sont pas suffisantes, car il serait alors souvent possible de recalculer les valeurs effacées par différence avec les marges du tableau.

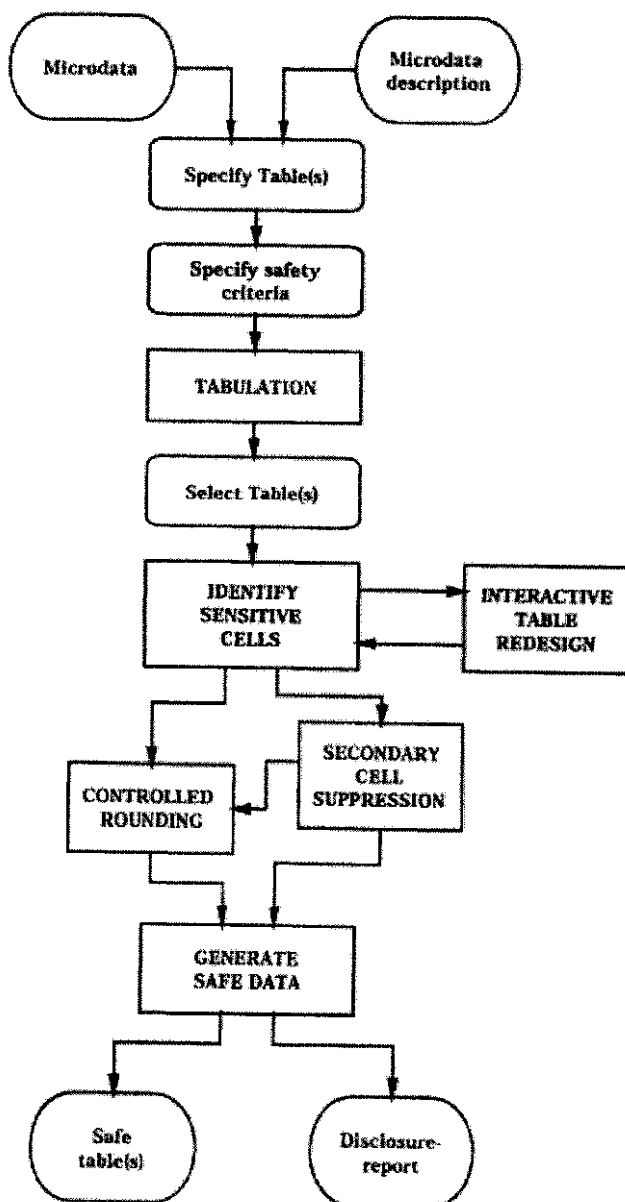
Argus efface d'autres cases pour éviter qu'un tel calcul soit possible. Ce sont les suppressions secondaires. Les suppressions secondaires sont optimisées pour minimiser la perte d'information.

Le véritable intérêt d'Argus réside dans cette optimisation. En effet, la résolution de ce programme d'optimisation nécessite en général un nombre d'opérations qui augmente comme la factoriel du nombre de cases du tableau. Argus recherche une solution approchée par une méthode de relaxation des contraintes.

Néanmoins, Argus ne peut traiter que des tableaux d'au plus quatre dimensions. Argus ne sait pas prendre en compte les contraintes supplémentaires créées par d'éventuels liens entre les tables ou les variables. Ces fonctionnalités ne seront disponibles que dans la version future.

La meilleure façon d'utiliser τ -Argus consiste à trouver un point d'équilibre entre, d'une part, les regroupements des modalités de variables, et d'autre part, les suppressions de cases.

Functional design of r-ARGUS



La première étape consiste à décrire les variables

Callout boxes:

- Ligne ou colonne
- Variable présentée dans les cases.
- Le code APE de l'activité peut être tronqué.
- Le logiciel peut rechercher la liste des modalités d'une variable.

Le tableau à protéger et les règles de secret statistique à suivre sont définis ensuite

Callout boxes:

- Règle de dominance, si la plus grande valeur apporte plus de 85 % du total, alors la case doit être effacée.
- Au moins trois individus par case.
- Pour calculer la perte d'information pendant les suppressions secondaires.
- Pour la règle de dominance des suppressions primaires.

Tableau du chiffre d'affaires (CA) selon l'activité et les tranches d'effectifs

Un bilan des cases ne respectant pas les règles de confidentialité est fourni à l'utilisateur. Ce bilan est présenté par variable.

TAU-ARGUS 2.0 - C:\v_arch\job\Argus\user\Secret.asc

Selected table: activite_xtrancha_ellip1 - CA

variable name: activite

Variable	dim1	dim2	Code	Label	Effectif	diff	diff2
activite	3	87	181Z		137	0	1
tranch...	0	87	182A		147	0	1
			182C		2177	0	2
			182D		1186	0	1
			182E		3494	0	2
			182G		508	0	2
			182J		1014	0	1
			183Z		334	0	2
			191Z		203	0	2
			192Z		1412	0	2
			193Z		587	0	0
			251A		20	0	4
			251C		127	0	3
			251E		434	0	1
			292A		116	0	1
			292C		401	0	3
			292D		988	0	2
			292F		2305	0	1
			292H		184	0	1
			292J		110	0	2
			292K		837	0	1
			296A		22	0	5
			296B		134	0	4
			351A		11	0	5
			351B		153	0	3
			351C		378	0	1
			351E		1103	0	2
			353A		59	0	5
			353B		292	0	4
			353C		6	1	4
			354A		70	0	3
			354C		101	0	2
			354E		11	1	2
			355Z		31	0	4
			364Z		516	0	3
			365Z		545	0	1
			366A		687	0	2
			366C		116	0	2
			366E		1789	0	0

Deux cases du tableau dans la ligne 182C ne respectent pas les règles de confidentialité.

Le total de l'activité 351A ne peut être diffusé.

La statistique de la ligne provient de 1103 individus.

Le tableau est édité. Des indications sur le respect des règles de confidentialité par la cellule pointée sont présentées dans le pavé « Cell information ».

Les suppressions primaires sont visualisées en rouge. (En gris dans ce document).

Dans l'exemple, le total du chiffre d'affaires de l'activité 351A ne respecte la règle de dominance.

activité x tranche_effectif -> CA (dominance rule)

activité tranche_effec

	Total	00	13	25	35	45	55	65	75	85
Total	535949781	35663756	17218207	49311559	33824845	45773743	58233952	41382282	114289344	14081
181Z	745223	190441	76033	410392	68357					
182A	2466965	151279	309234	566127	359012	592122	509091			
182C	3560386	796349	177116	212040	366126	610833	518688	679314		
182D	13671140	2288915	965027	2354125	1717647	3042940	2339257	902229		
182E	34414691	6146585	3992102	11037079	2573394	4056412	4946530	609045	1150739	
182G	12415515	431283	376998	2380271	1367112	1824056	2248992	1498171	2288632	
182J	11799704	1332491	922213	2494975	1529554	1714482	497934	3317655		
183Z	895926	334288	90047	326373	63184		82834			
191Z	2701471	383985	406398	771474	364724	426191	336699			
192Z	9895375	1287154	478548	1500558	465786	891177	1268398	379344	3634400	
193Z	16516138	509485	277820	2013610	1735518	2861709	6446147	2671049		
251A	39756341	4293	13959	76125		726474	511936		4666345	325
251C	1332890	254750	75563	413217	39108	88272		469980		
251E	22463495	590163	521843	1582665	1861520	1949054	4747694	2920324	3562330	43
292A	4183520	214182	272297	767995	321155	1151790	1456161			
292C	13266314	513868	352952	793279	422352	386732	1000866	2115554	2171117	55
292D	26457061	1551445	1425469	3537868	2970698	4466879	4600886	2182159	5722458	
292F	32682086	3658995	2745697	5819933	2923184	5401018	5952845	3482756	2697858	
292H	4326485	350932	241840	1023929	1190722	873381	645681			
292J	1744842	168961	215270	258865	366195	478847	324703			
292K	14428142	1237403	744532	2300269	1705228	1955196	2296179	3633648	556687	
296A	12538149	1886794		26295	235721		5255049	433222		55
296B	720772	143882	53668	77414	50364	161980	233464			
351A	5500753	1297	4783	33432	59471	4906942		494828		

Cell Information

Cell item: resp var

Value: 535949781

Status: Safe

contributions: 22825

Top n of shadow: 73931286

place Variable

response CA

shadow CA

cost CA

Record

Round Undo Round

Suppress Undo Suppress

Suppress Group Undo Suppress

Enfin « Suppress » lance l'algorithme des suppressions secondaires. L'utilisateur doit paramétrer une autre règle de confidentialité qui interdit une estimation trop précise des valeurs supprimées qui serait possibles en prenant en compte les contraintes de positivité. Un exemple de l'impact des contraintes de positivité est disponible dans la note N°24/D250 de L. Viglino. Il est possible de choisir pour bornes, 130% et 70% de la valeur supprimée.

L'image suivante montre le tableau protégé. Les suppressions secondaires sont normalement éditées en bleu. Dans ce document, les suppressions secondaires ont été soulignées.

activité x tranche_effectif -> CA (dominance rule)											
activité	tranche	effectif									
	Total	00	13	25	35	45	55	65	75	85	
Total	535949781	35063756	17218207	49311559	33824845	45773743	58233552	41382282	114289344	1408	
181Z	745223	190441	76933	410392	68357						
182A	2486865	151279	305234	566127	359012	592122	509091				
182C	9560386	795349	177116	212040	366126	610633	516508	879314			
182D	13611140	2288915	968027	2354125	1717647	3042940	2339257	902229			
182E	34414691	6149589	3992102	11037079	2573394	4056412	4848530	609846	1150739		
182G	12415515	431283	376998	2380271	1367112	1824056	2248992	1498171	2288632		
182J	11799704	1332491	922213	2494975	1529954	1714482	497934	3317855			
183Z	895326	334288	90847	326373	63184		82034				
191Z	2701471	393985	406398	771474	364724	426191	338693				
132Z	9895375	1287154	478548	1500558	485796	891177	1258398	379344	3634408		
133Z	16516138	509485	277820	2013610	1735518	2961709	6446147	2671849			
251A	38756341	4293	13990	76125		726474	511936		4866945	325	
251C	1332890	254750	75563	413217	39108	88272		463988			
251E	22463455	590163	521843	1582665	1881520	1949054	4747694	2920324	3962330	43	
292A	4183520	214162	272257	767995	321155	1151790	1456161				
292C	13266914	513868	352952	793279	422352	366732	1000066	2115554	2171117	55	
292D	26457061	1551445	1425469	3537869	2970698	4465879	4600886	2182159	5722458		
292F	32682886	3658995	2745697	5819933	2923184	5401018	5952645	3482756	2697858		
292H	4326485	350932	241840	1023929	1190722	873381	645681				
292J	1744842	160961	215270	258866	306195	478847	324783				
292K	14428142	1237403	744532	2300269	1705228	1955196	2296179	3633648	555687		
296A	12538149	1886794		26295	235721		5255049	433222		55	
296B	720772	143882	53688	77414	50364	161980	233454				
351A	9500783	1297	4783	33432	93471	4906842		494828			

Cell Information

CellItem: resp var
 Value: 535949781
 Status: Scale
 # contributions: 22825
 Top n of shadow: 73931286

place	Variable
response	CA
shadow	CA
cost	CA

Encode

Round: Hide Prev

Success: Undo Supp

Suppress Group: Hide Data

Dans le tableau de diffusion, les suppressions, primaires et secondaires, sont indiquées par un « s ».

Chiffre d'affaires par activité et tranche d'effectifs

Tranche d'effectifs	Total	0	13	25	35
Activité					
Total	535949792	35063756	17218208	49311560	3382844
181Z	745223	190441	s	410392	s
182A	2486865	151279	309234	566127	359012
182C	3560386	796349	177116	212040	366126
182D	13611140	2288915	966027	2354125	1717647
182E	34414692	6148589	3992102	11037079	2573394
182G	12415515	431283	376998	2380271	1367112
182J	11799704	1332491	922213	2484975	1529954
183Z	895926	334288	90047	326373	s
191Z	2701471	393985	406398	77174	364724
192Z	9895375	1287154	478548	1500558	465796
193Z	16516138	509485	277820	2013610	1735518
251A	38756340	4293	s	s	0
251C	1332890	254750	75563	413217	s
251E	22463456	590163	521843	1582665	1881520
292A	4183520	214162	272257	767995	s
292C	13266914	513868	352952	793279	422352
292D	26457060	1551445	1425469	3537868	2970698
292F	32682086	3658995	2745697	5819933	2923184
292H	4326485	350932	241840	1023929	1190722
292J	1744842	160961	215270	258866	s
292K	14428142	1237403	744532	2300269	1705228
296A	12538149	s	0	s	s
292B	720772	143882	s	77414	s
351A	s	1297	s	s	s
351B	6846476	206113	90521	398767	0
351C	2357617	476824	305488	570476	s
351E	6041720	1588500	313732	636746	472403
353A	32912344	53854	s	120455	s
353B	155774624	s	213525	743397	s
353C	s	s	s	0	s
354A	3962803	115089	54271	57650	s
354C	2680916	100111	64050	496380	341257
354E	s	17724	0	s	0
355Z	1190773	8747	s	260334	s
364Z	7289389	614239	250647	1434856	600034
365Z	5006444	420737	253104	904974	791314
366A	1342549	505358	140154	295680	s
366C	2677988	147019	138312	107088	743935
366E	13233733	2872532	610347	2178737	1366778

(suite)

Tranche d'effectifs	45	55	65	75	85
Activité					
Total	45773744	58233552	41382280	114289344	140852496
181Z	0	0	0	0	0
182A	s	s	0	0	0
182C	610833	s	s	0	0
182D	3042940	s	s	0	0
182E	4056412	4846530	s	s	0
182G	1824056	2248992	s	s	0
182J	1714482	s	s	0	0
183Z	0	s	0	0	0
191Z	s	s	0	0	0
192Z	891177	1258398	s	s	0
193Z	2861709	6446147	2671849	0	0
251A	726474	s	0	s	32556578
251C	s	0	s	0	0
251E	1949054	4747694	2920324	s	s
292A	1151790	s	0	0	0
292C	386732	1000066	s	s	s
292D	4466879	4600086	s	s	0
292F	5401018	5952645	s	s	0
292H	s	s	0	0	0
292J	478847	s	0	0	0
292K	1955196	2296179	s	s	0
296A	0	5255049	s	0	s
292B	s	s	0	0	0
351A	s	0	s	0	0
351B	596627	s	s	0	s
351C	494511	s	0	0	0
351E	590465	s	s	0	0
353A	s	1943308	4212102	s	s
353B	942414	3553238	3547655	s	s
353C	0	0	0	0	s
354A	s	0	0	s	0
354C	s	1066280	s	0	0
354E	0	s	0	0	0
355Z	s	0	s	0	0
364Z	636528	s	s	s	0
365Z	919519	s	s	0	0
366A	164497	s	0	0	0
366C	s	482473	s	0	0
366E	2045969	2014236	2145134	0	0

Argus archive un rapport avec les options retenues, et un bilan quantitatif des caves effacées.

The input file with metadata is C:\lv_arch\job\Argus\suse\Secret_tau.rda (20 :17 :24 on 23/11/2000)

The input file with microdata is C:\lv_arch\job\Argus\suse\Secret.asc (17 :44 :20 on 17/11/2000)

The table was saved in C:\lv_arch\job\Argus\suse\secret_tau.ttb as follows :

Table :

activité x tranche_effectif -> CA

Cost variable for cells ; CA

The dominance rule with $N=1$ and $K=85\%$ on variable CA and the cell frequency limit 3 were applied ;

the safety range for each cell was [70%,130%].

There are 87 primary and 11 secondary suppressions in the elementary cells.

There are 3 primary and 0 secondary suppressions in the 1-dimensional marginals.

The general total was not suppressed

La protection d'un fichier

Pour protéger l'anonymat dans un fichier, il faut rechercher et exclure les clés d'identification. Une clé d'identification est une combinaison unique, de caractéristiques propres à un individu, et présentes dans le fichier.

Par exemple, même en l'absence des noms, prénoms et adresses, de nombreux anciens élèves de l'Ensaë pourraient être identifiées sans ambiguïté dans un fichier de salariés comportant le diplôme et la raison sociale de l'entreprise.

L'utilisation de clés d'identification dépend des informations externes disponibles. Une fois l'individu identifié, l'utilisateur malveillant du fichier pourrait alors en extraire les informations qu'il ne possède pas déjà.

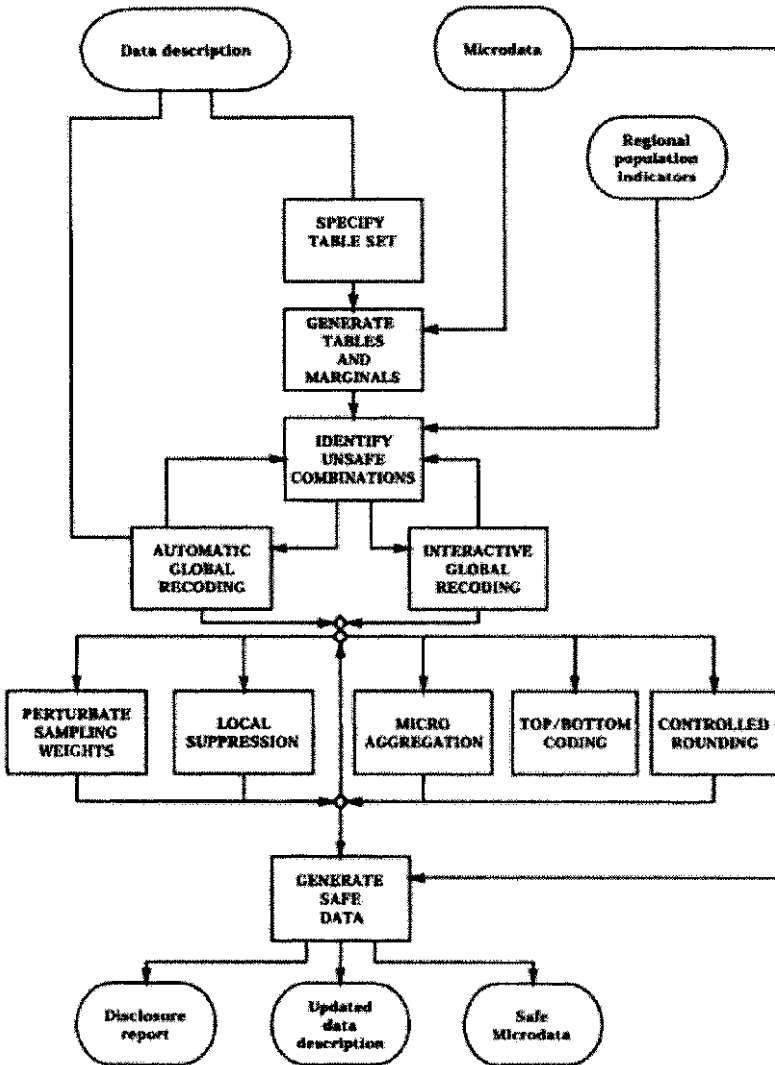
Argus recherche les clés d'identification et les supprime en effaçant les valeurs d'une ou plusieurs variables de la clé. Si une variable est commune à plusieurs clés pour un individu, sa valeur est effacée en priorité.

En pratique, il est judicieux de commencer par une phase de transformation globale des variables, avant d'utiliser les suppressions locales, et d'arbitrer entre le nombre de suppressions et la finesse des variables.

L'utilisateur indique au logiciel la nature des variables et leur niveau d'identification estimé. Les clés d'identification sont recherchées parmi des croisements de variables, soit générés par le logiciel, soit indiqués par l'utilisateur. μ -Argus peut effectuer de lui-même les modifications globales des variables, ou laisser l'utilisateur maître de cette étape.

Ces techniques de μ -Argus ne sont pas très efficaces pour les fichiers d'entreprises où de nombreuses clés existent. La version ultérieure d'Argus développera des méthodes d'addition d'aléas à la valeur prise par les variables

Functional design of μ -ARGUS



μ -ARGUS (3.0) user's manual

La première étape consiste à décrire les variables du fichier.

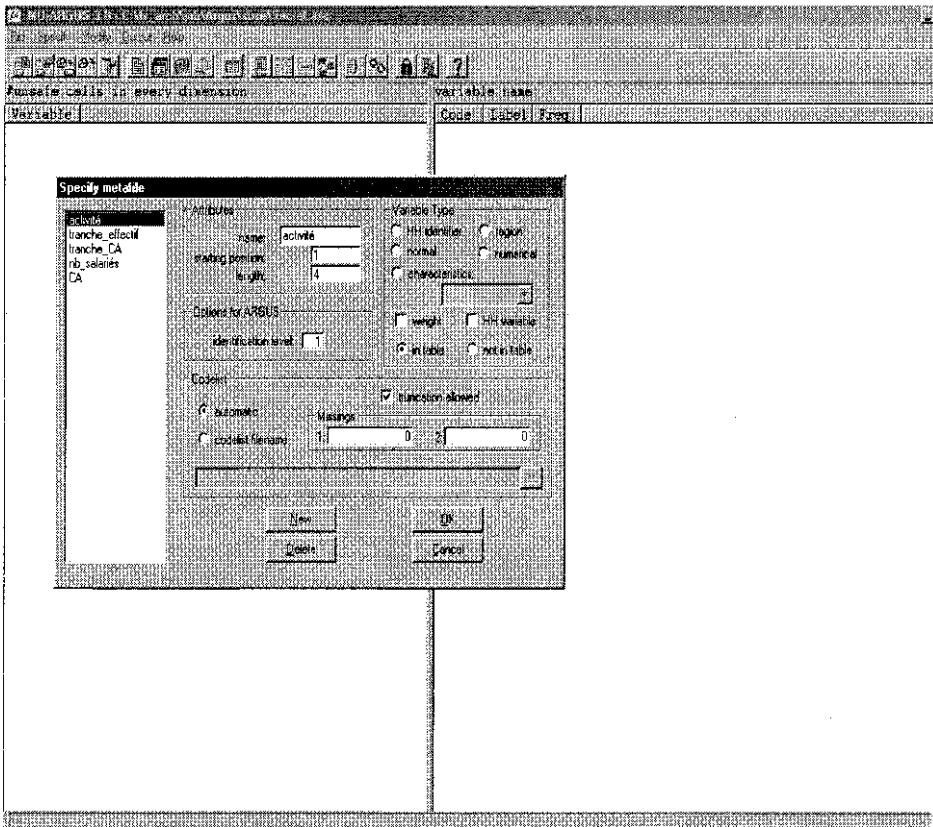
Les variables « HH identifiant » et « HH variable » servent à décrire un fichier hiérarchisé. Par exemple, « HH variable » sert à indiquer qu'une valeur de cette variable est commune à tout un ménage. Si le logiciel doit supprimer cette valeur pour un membre du ménage, alors il l'effacera aussi pour chaque membre.

Les variables numériques peuvent être arrondies et queues de distribution regroupées.

« Idlevel » indique le niveau d'identification a priori d'une variable. μ -Argus utilise cette information dans l'une des options de recherche des clés d'identification.

Région sert à indiquer toute variable géographique, même la commune.

L'option « not in table » permet d'exclure une variable du processus de recherche des clés d'identification. Cette option est utile pour éviter une recherche automatique, inutilement longue, des modalités de la variable.



Le tableau suivant présente l'étape où les clés d'identification sont recherchées parmi les croisements prédéfinis par l'utilisateur.

The screenshot shows a software window titled "Specify combinations". On the left, a list of variables includes "Echelle", "tranche_effecit", and "tranche_CA". The central area contains a "Variable" field with "3" entered, a "Generate table" button, and a "Work" button. On the right, a table displays the resulting combinations:

Triaxibus	Var 1	Var 2	Var 3
3	activité	tranche_eff...	tranche...

Three callout boxes provide the following text:

- La règle d'anonymat est un minimum d'au moins trois individus avec la même combinaison
- Le logiciel peut examiner les croisements selon les niveaux d'identification définis précédemment.
- Utilisation, du ou des, croisements de variables directement imposés par l'utilisateur

Le tableau suivant est un extrait du fichier qui a servi pour les exemples. Le fichier qui a servi d'exemple est un extrait de 22 825 entreprises du registre Suse.

Fichier à protéger

182E	00	30	6	3111
296B	45	70	189	161980
182E	25	50	23	48290
182C	35	60	76	53826
182E	00	40	8	7907
182G	00	50	9	17977
366E	13	40	12	7410
182J	25	40	23	7029
292D	13	30	11	3539
292D	13	50	14	13590
182D	00	20	7	1860
251E	13	50	14	12282
292F	13	50	12	11774
192Z	25	40	25	7696
193Z	25	50	34	17801
182E	25	40	20	9170
292K	00	50	9	10952
292D	55	70	205	273277
354A	00	30	4	3038
182E	25	40	35	7483
182E	25	50	25	40927
182E	25	50	46	32286

Fichier protégé

182E	00	30	6.000000	3111.0000
0	45	70	189.000000	161980.00
182E	25	50	23.000000	48290.000
182C	35	60	76.000000	53826.000
182E	00	40	8.000000	7907.0000
182G	00	50	9.000000	17977.000
366E	13	40	12.000000	7410.0000
182J	25	40	23.000000	7029.0000
292D	13	30	11.000000	3539.0000
292D	13	50	14.000000	13590.000
182D	00	20	7.000000	1860.0000
251E	13	50	14.000000	12282.000
292F	13	50	12.000000	11774.000
192Z	25	40	25.000000	7696.0000
193Z	25	50	34.000000	17801.000
182E	25	40	20.000000	9170.0000
292K	00	50	9.000000	10952.000
292D	55	70	205.000000	273277.00
354A	00	30	4.000000	3038.0000
182E	25	40	35.000000	7483.0000
182E	25	50	25.000000	40927.000
182E	25	50	46.000000	32286.000

Argus a enregistré le rapport suivant :

Produced 21:36:2 on 23/11/2000

The input file with metadata is C:\lv_arch\job\Argus\suse\Secret_tau.rda (20:17:24 on 23/11/2000)

The input file with microdata is C:\lv_arch\job\Argus\suse\Secret.asc (17:44:20 on 17/11/2000)

The output file with metadata is C:\lv_arch\job\Argus\suse\safefile_mu.rds

The output file with microdata is C:\lv_arch\job\Argus\suse\safefile_mu.saf

Variables:

Suppressions:

"activité" has been suppressed in 60 records.

"tranche_effectif" has been suppressed in 433 records.

"tranche_CA" has been suppressed in 64 records.

The total number of suppressions is 557.

Références.

Manuel d'argus sur le site Web du CBS: www.cbs.nl/sdc/argus.htm.

Actes du congrès ECE/Eurostat de mars 1999 à Salonique sur la confidentialité: www.unece.org/stats/documents/1999.03.confidentiality.htm.

Eurostat (1993), Manual on disclosure control methods.

Fischetti, M. and Salazar J.J. (1998) Modeling and solving the cell suppression problem for linearly constrained tabular data. Actes de la conférence d'Eurostat: Statistical data protection, du 25-27 mars à Lisbonne.

Giessing, S. (1999), A survey on software packages for automated secondary cell suppression, Working paper 12 des actes du congrès ECE/Eurostat 1999.

James, A.M. et Lang, G. (1998), Règles de secret statistique applicables à la diffusion et à la cession des données et textes de référence. Document de travail n° D9802 bis de l'Insee.

Takemura, A. (1999), Local recoding by optimal matching algorithm for disclosure control of microdata sets. Institut International de Statistique, Helsinki, 10-18 août 1999.

Viglino, L. (1999), Travail engagé sur le secret statistique et les techniques de mises en oeuvre des règles de confidentialité, note interne Insee n°24/D250, 27 juillet 1999.

Viglino, L. (1999), Argus, Gardien du secret statistique. Courrier des statistiques, n° 91-92, décembre 1999.

Pohl, R. (2000), Outils de protection de la confidentialité des données statistiques, fichiers de données individuelles, logiciel Muargus, Note interne Insee, n° 002/H005, 27 octobre 2000.