



**HAL**  
open science

## Le logiciel SODAS : avancées récentes Un outil pour analyser et visualiser des données symboliques

Myriam Touati, Mohamed Rahal, Filipe Afonso, Edwin Diday

### ► To cite this version:

Myriam Touati, Mohamed Rahal, Filipe Afonso, Edwin Diday. Le logiciel SODAS : avancées récentes Un outil pour analyser et visualiser des données symboliques. Extraction et gestion des connaissances EGC 2008, Jan 2008, Sophia Antipolis, France. pp.239,240. hal-00461547

**HAL Id: hal-00461547**

**<https://hal.science/hal-00461547>**

Submitted on 7 Mar 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## **Le logiciel SODAS : avancées récentes**

### **Un outil pour analyser et visualiser des données symboliques**

Myriam Touati<sup>(1,2)</sup>, Mohamed Rahal<sup>(2,1)</sup>, Filipe Afonso<sup>(2,1)</sup>, Edwin Diday<sup>(2,1)</sup>

(1) CNRS, UMR7534, F-75016 Paris, France

(2) Université Paris-Dauphine, CEREMADE, F-75016 Paris, France

(touati, rahal, afonso, diday) @ceremade.dauphine.fr

<http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm>

Le logiciel public SODAS, issu de deux projets européens (9 pays participants) soutenu par EUROSTAT: SODAS et ASSO, est un logiciel d'Analyse de Données Symboliques qui permet de construire puis d'exploiter des unités statistiques à un niveau de généralité supérieur à celui des individus usuels en les représentant par des variables dites « symboliques » prenant en compte leur variation interne. Ainsi ces variables sont à valeur numériques ou qualitatives mais plus généralement à valeur intervalle, histogrammes, suite de valeurs, etc. Ce logiciel est sans cesse amélioré et de nouvelles fonctions y sont ajoutées au fur et à mesure de nos recherches. Nous nous proposons donc de vous exposer l'amélioration des méthodes de classification (SCLUST), d'interprétation et de caractérisation de classes (DSTAT) et de visualisation et classification pyramidale (HIPYR, PYR2D et PYR3D). Ces nouveaux modules ont été développés dans le cadre de l'ANR SEVEN pilotée par EDF (Clamart).

SCLUST est une méthode de classification par Nuées Dynamiques étendue aux données symboliques intervalles et histogrammes. La dernière version de ce module, améliore l'exécution dans le cas d'un nombre important de variables de type histogramme (plus de 100). D'autre part, en plus des fichiers attribuant à chaque individu sa classe et décrivant les classes obtenues, des fichiers sur les inerties inter-classes, intra-classes et totales sont fournis en sortie. Ces fichiers permettent l'étude et la visualisation de la qualité et de la caractérisation des classes obtenues ainsi que la sélection des variables discriminantes.

Le module DSTAT est formé d'un ensemble de méthodes de statistiques descriptives spécifiquement adaptées à des données «symboliques ». Elles permettent d'interpréter et de décrire graphiquement ces données. Une option de ce module permet d'afficher la variation des fréquences de chaque modalité d'une variable à valeur histogramme donnée (fig. 1). Une autre option permet de caractériser un concept par les modalités les plus caractéristiques (fig. 2). L'option *BIPLOT* du module DSTAT permet de visualiser le tableau croisant deux variables histogrammes (fig. 3).

Le module de classification ascendante hiérarchique et pyramidale HIPYR (appellation SODAS) construit une pyramide (resp. hiérarchie) sur un ensemble de données symboliques et/ou numériques, il permet de caractériser les classes résultantes en les organisant sous forme de paliers et offre donc une représentation en classes recouvrantes et empiétantes permettant de découvrir des ordres et sous ordres dans une population. Dans le cadre du projet ANR SEVEN, C. Jacquemin et F. Vernier de l'équipe AMI (CNRS-LIMSI) ont coordonné la réalisation d'une interface intuitive de visualisation et d'accès aux données, de sélection et d'annotation de classes, et de report des informations numériques pertinentes des visualisation des pyramides 2D et 3D (voir fig. 4 un premier aperçu).

Les améliorations du logiciel SODAS ont permis d'enrichir l'analyse de données symboliques aussi bien du point de vue des interprétations statistiques que visuelles. Ces nouveaux modules vont être intégrés dans la plateforme interactive SEVEN.

Le logiciel SODAS : avancées récentes



FIG. 1 – Variation des histogrammes de la variable « DIVERS »

probA/mean	probA	category	variable	opposite individual	range
3.425911	0.898371	1	vague	13	0.831271
2.987727	0.952028	NR	sa_suit	3	0.807888
2.438132	0.427673	tarif - consommation	Divers	13	0.329578
3.491417	0.255372	NR	s_Soc	11	0.244955
1.373186	0.425977	ARoDnt	segment_client	6	0.168091
1.194586	0.440374	NR	Faibles	4	0.154438
1.296363	0.416667	Q	giz	6	0.152299
1.177148	0.801677	Rubst OK	Image	11	0.135010
2.288297	0.171608	NR	sa_benef	6	0.134029
2.22413	0.159136	Moy OK	sa_coursel	16	0.118342
1.591803	0.176101	Rubst OK	sa_coursel	4	0.100761
1.176920	0.358491	Region Parisienne	region_age	6	0.100915
1.804786	0.129000	Rubst OK	sa_services	11	0.096223
2.618123	0.078652	NSP	nat_17	9	0.077096
1.623014	0.358071	NR	sa_courier	4	0.059081
2.607660	0.016568	NR	s_groupe	3	0.016068
2.980629	0.016772	NR	sa_groupe	16	0.016772
18.000000	0.004193	NR	Image	2	0.004193

FIG. 2 – Caractérisation de la classe 7

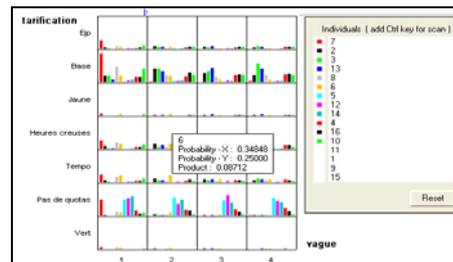


FIG. 3 – Représentation croisée des variables histogramme tarification et vague

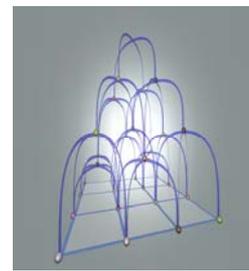
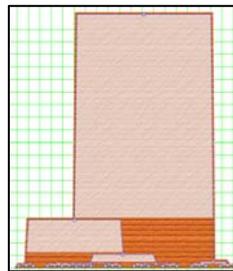
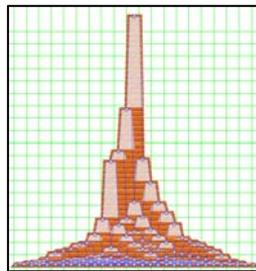


FIG. 4 – Une nouvelle visualisation 2D et 3D pour les pyramides et hiérarchies

## Références

- Billard L., E. Diday (2006). *Symbolic Data Analysis: conceptual statistics and data mining*. Wiley series in computational statistics. Wiley. ISBN 0-470-09016-2.
- Bock H.H., Diday E. editors (2000). *Analysis of Symbolic Data for extracting statistical information from complex data*. Heidelberg, Springer Verlag, ISBN 3-540-66619-2.
- Diday E., M. Noirhomme editors (2007). *Symbolic Data Analysis and the SODAS software*. Livre à paraître cette année, Wiley.
- Pak K., M.C. Rahal et E. Diday (2005). *Élagage et aide à l'interprétation symbolique et graphique d'une pyramide*. Congrès d'extraction et gestion des connaissances, EGC 18-21 Janvier 2005, Paris, Cepadues