

# A implicação estatística usada como ferramenta em um exemplo de análise de dados multidimensionais

RÉGIS GRAS\* e SADDO AG ALMOULOU\*\*

## Resumo

O artigo apresenta os resultados de um questionário, cujo objetivo é analisar as opiniões dos participantes de um colóquio organizado pela PUC-SP sobre a implicação estatística de análise de dados multidimensionais. O colóquio foi realizado no intuito de estudar as potencialidades, a pertinência e a importância da realização de uma análise implicativa nas investigações das Ciências da Educação. A implicação estatística foi a principal ferramenta para analisar as informações envolvidas no questionário.

*Palavras-chave:* implicação estatística; metodologia de análise de dados; hierarquia de similaridade; hierarquia implicativa; avaliação.

## Resumé

*L'article présente les résultats d'un questionnaire dont l'objectif est l'étude des opinions des participants d'un colloque organisé à la Pontificia Universidade Católica de São Paulo, sur l'analyse implicative de données statistiques multidimensionnelles. Le colloque a pour objectif d'étudier les potentialités, la pertinence et l'importance de l'analyse implicative pour les recherches en sciences de l'éducation. L'implication statistique a été l'outil principal de traitement et d'analyse des données issues de ce questionnaire.*

*Mots-clés:* implication statistique; méthodologie d'analyse de données; hiérarchie de similarité; hiérarchie implicative; évaluation.

## Abstract

*The article presents the results of a questionnaire whose purpose is to analyse the opinions of the participants in a seminar organised by PUC-SP about the statistical implication of multidimensional data analysis. The seminar aimed at investigating the potentialities, the pertinence and importance of carrying out implicative analysis in the investigations in the area of Education Sciences. Statistical implication was the main tool to analyse the information contained in the questionnaire.*

*Key-words:* statistical implication; data analysis methodology; similarity hierarchy; implicative hierarchy; evaluation.

\* Professeur Emérite à l'École Polytechnique de l'Université de Nantes, La Chantrerie, BP 50609 44306 Nantes cedex 03. E-mail: regisgra@club-internet.fr

\*\* Professor do Programa de Estudos Pós-Graduados em Educação Matemática – PUC/SP. E-mail: saddoag@pucsp.br.

## Introdução

Este artigo discute os resultados de um questionário, cujo objetivo é avaliar o colóquio intitulado “O método estatístico implicativo utilizado em estudos qualitativos de regras de associação: contribuição à pesquisa em educação”, realizado em julho de 2003 no Programa de Estudos Pós-Graduados em Educação Matemática da PUC-SP.

O colóquio tem por objetivo principal realizar um estudo de refinamento sobre as pontecialidades, a pertinência, bem como a importância da realização de análises estatísticas de dados multidimensionais (análise hierárquica de similaridade, análise implicativa) nas investigações da Educação Matemática e em investigações em Educação de um modo mais abrangente. Foram previstos, para o desenvolvimento do trabalho, três momentos:

1. Realização de estudo das fases fundamentais de uma análise de dados multidimensionais (instrumentos de coleta de dados, organização e exploração, instrumentos de tratamentos, interpretação, levando em conta a questão e os objetivos da pesquisa);

2. Realização de um breve estudo de caráter teórico e intuitivo sobre os diferentes métodos de análise de dados multidimensionais e oferta de oficinas sobre os métodos;

3. Realização de estudo dos exemplos de dados encontrados em pesquisas desenvolvidas no Programa de Pós-graduação em Educação Matemática e em Educação: Currículo da PUC-SP;

4. Organização de um Colóquio sobre os métodos, aberto a distintos Programas de Pós-Graduação da PUC-SP, bem como de outras Universidades.

O colóquio contou com a participação ativa do professor emérito da Escola Politécnica da Universidade de Nantes: Régis Gras. O referido autor e seu grupo de pesquisa, desde 1979, procuraram, entre outros assuntos, colocar à disposição dos pesquisadores (em matemática, em psicologia, em biologia, em educação, etc.) ferramentas estatísticas (a análise implicativa, a hierarquia implicativa) que permitem evidenciar a dinâmica dos comportamentos de sujeitos (alunos, por exemplo) em situação de resolução de problemas, no caso da Educação Matemática. A análise implicativa (cf. o artigo de Gras nesta revista), como todos os

métodos de análise estatística de dados multidimensionais, permite visualizar, organizar, construir modelos e explicar fenômenos associados aos dados.

## **Estrutura do questionário**

Para a constituição do corpo de informações a ser analisado foi aplicado um questionário-avaliação aos participantes do colóquio. O questionário (cf. Anexo 1), além de informações pessoais, solicitava que os participantes respondessem a quatro questões subjetivas, que objetivavam fornecer uma visão de como eles vivenciaram a programação e os conteúdos dos trabalhos no evento.

As questões subjetivas foram as seguintes:

- 1) *Suas expectativas foram satisfeitas?*  
( ) Sim      ( ) Parcialmente      ( ) Não
- 2) *Quais pontos da programação despertaram mais seu interesse?*
- 3) *Quais aspectos você desejaria aprofundar mais?*
- 4) *Na lista abaixo, escolha a(s) expressão(ões) que melhor expressa(m) seu sentimento, a propósito do colóquio.*  
( ) muito longo    ( ) muito curto    ( ) cansativo    ( ) estimulante  
( ) curiosidade    ( ) muito bom    ( ) inútil    ( ) abertura  
( ) difícil    ( ) gostaria fazer mais

As respostas foram codificadas (cf. Anexo 1) e as variáveis estatísticas que dizem respeito aos dados pessoais foram consideradas como variáveis suplementares (o código dessas variáveis é do tipo X s).

Analizamos as duas questões abertas (“Quais os pontos da programação despertaram mais seu interesse?” e “Quais aspectos você desejaria aprofundar mais?”), destacando as palavras mais significativas das respostas e essas palavras foram agrupadas por sinonímia. Obtivemos, assim, 6 variáveis para a primeira questão e 4, para a segunda, retomando, às vezes, as alegações que eram comparáveis de uma questão a outra.

Mesmo sabendo que o número de questionários esteja relativamente fraco, algumas estruturas interessantes são interpretáveis, sem que seu caráter explicativo seja de uma fidedignidade absoluta; por consequência, precisa ser confirmada por outro estudo. Nós nos contentaremos, então, em destacar as tendências e os pontos mais assegurados.

A tabela do Anexo 2 apresenta as ocorrências, as médias e os desvios-padrão das variáveis estatísticas estudadas. Observa-se que as variáveis que tiveram maiores ocorrências são: XSS (expectativas satisfeitas), EST (o trabalho foi estimulante) e IMA (manipulação de CHIC). Essas três variáveis expressam o grau de satisfação dos participantes e a pertinência da metodologia de trabalho adotada.

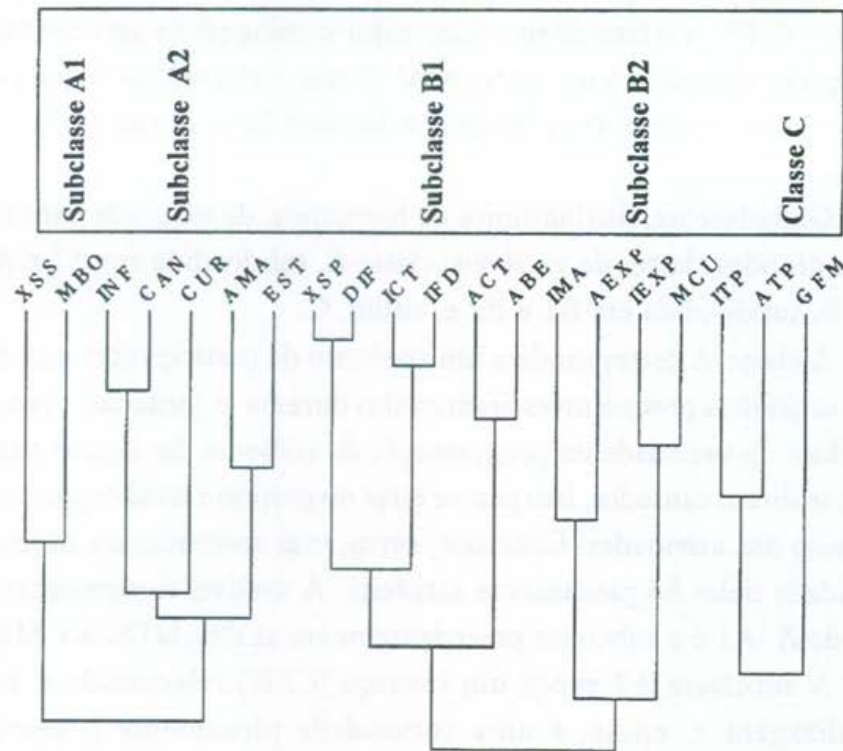
### **Análise de similaridades segundo I. C. Lerman**

Como em todos os métodos de classificação, procuramos constituir, em um conjunto  $V$  de variáveis, partições de  $V$  cada vez menos finas, construídas de maneira ascendente. Essas partições encaixadas são representadas por uma árvore construída usando um critério de similaridade ou de semelhança estatística entre variáveis. A similaridade se define a partir do cruzamento do conjunto  $V$  das variáveis com um conjunto  $E$  de sujeitos (ou de objetos). Esse tipo de análise permite ao usuário estudar e interpretar, em termos de tipologia e de semelhança (e não semelhança) decrescente, classes de variáveis constituídas significativamente a certos níveis da árvore e se opondo a outras nesses mesmos níveis.

O critério de similaridade se exprime da maneira seguinte, nos casos das variáveis binárias (presença – ausência, verdadeiro – falso, sim – não, etc.): 2 variáveis  $a$  e  $b$ , satisfeitas respectivamente por dois subconjuntos  $A$  e  $B$  de  $E$ , são muito semelhantes quando o número  $k$  dos sujeitos de  $A \cap B$  é importante de um lado, pelo que teria sido no caso da ausência de ligação entre  $a$  e  $b$ , e por outro lado, com relação aos cardinais de  $E$ ,  $A$  e  $B$ . Medimos essa semelhança pela probabilidade de que  $k$  seja superior ao número aleatório esperado nesta situação.

A modelagem probabilista da variável aleatória, cujo  $k$  é a realização, pode ser uma distribuição *binomial* ou de *Poisson*, à escolha do usuário. A segunda distribuição supõe que  $E$  seja uma amostra de uma população grande, o que a primeira não supõe. Se  $E$  não tem nenhuma razão estatística, *a priori*, de ser representativo, é preferível usar o modelo binomial que analisa a estrutura de  $E$  enquanto tal. Quando os parâmetros o permitem, uma aproximação gaussiana dessas duas leis é efetuada.

O índice de similaridade entre variáveis é usado para definir um índice de similaridade entre duas classes de variáveis, segundo o princípio de comparação entre a observação e o que seria dado pelo acaso. Um índice, dito de coesão, permite reagrupar as classes de variáveis.



Árvore de similaridade

Assim, para construir uma árvore de similaridade (cf. árvore acima), reunimos em uma classe de primeiro nível, primeiramente, as 2 variáveis que são mais similares, no sentido do índice de similaridade, depois 2 outras variáveis ou uma variável e a classe já formada, no sentido do índice da classe e, depois, outras variáveis ou classes de variáveis.

### Interpretação sucinta

No primeiro nível aparece uma relação de similaridade entre as únicas duas variáveis, mais ou menos restritivas: "expectativas parcial-

mente satisfeitas” e “difícil”. O ritmo do colóquio pôde explicar o cansaço, tanto mais que a carga teórica era muito importante. Identificamos aqui, talvez, o desequilíbrio entre aportes teóricos e ilustrações. Mas a representação desse fenômeno é muito minoritária.

Sublinhamos, por outro lado e ao contrário, que a relação entre satisfação (84%) e o fato de reconhecer que o colóquio foi satisfatório, de modo geral, aparece em um outro nível. É esse caráter (MBO) que consideramos como indício de satisfação. A variável EF é a mais típica dessa relação.

Globalmente, distinguimos na hierarquia, da esquerda para direita, três grandes classes de variáveis: classe A, subdividida em A1 e A2, a classe B, subdividida em B1 e B2 e, enfim, C.

A classe A corresponde a um conjunto de participantes que estão muito satisfeitos porque investiram muito durante as jornadas, apesar da extensão e da variedade da programação do colóquio. Se alguns participantes se dizem cansados, isto parece estar no próprio e total engajamento do sujeito nas atividades. Cansados, certo, mas apreenderam muito e a curiosidade deles foi plenamente satisfeita. A variável suplementar G é típica de A. A1 é a subclasse precedentemente analisada (XSS e MBO).

A subclasse A2 evoca um cansaço (CAN) relacionado a novas aprendizagens e, então, a uma curiosidade plenamente preenchida (tipicidade G).

A classe B contém as nuances explicando a percepção da dificuldade e, então, da satisfação parcial:

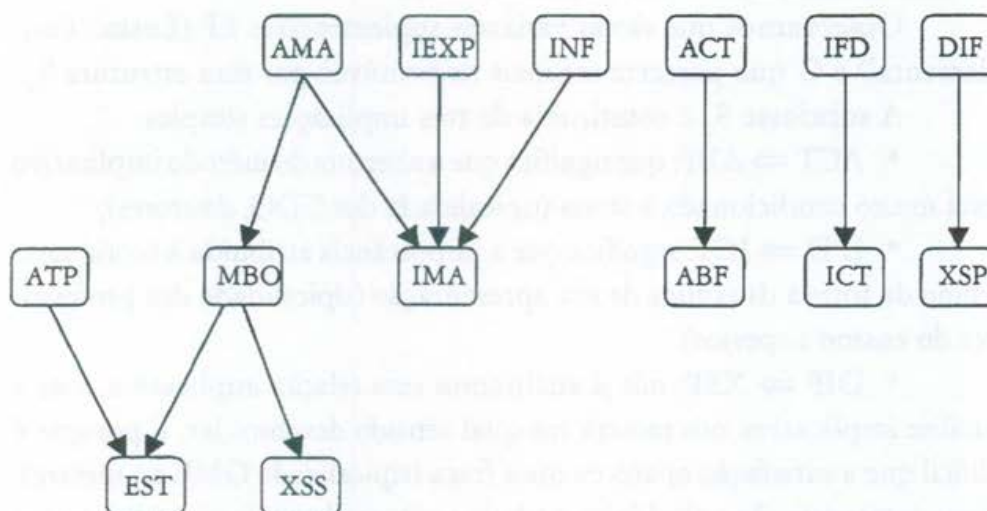
- a subclasse B1 é justamente a já analisada (XSP e DIF). GME é a variável a mais típica desta subclasse e, globalmente, de B.
- a subclasse B2 explica que a origem da restrição vem da complexidade teórica, o que não impede que haja demanda posterior (ACT) de mais atividades sobre o método, em razão das possibilidades (ABE) que a teoria permite. A variável M é a mais típica.

A classe C revela a satisfação e uma nova demanda (GFM), de uma programação na qual a teoria e prática (ITP e ATP) seriam intimamente associadas. GMA é a variável a mais típica.

## Análise Implicativa

### Análise do grafo implicativo obtido

Consideramos a construção do grafo implicativo no valor mínimo de 0,70 que permite apresentar uma estrutura interessante e, ainda, significativa do ponto de vista estatístico. É possível diminuir esse valor, a fim de obter mais relações implicativas, relações essas se tornando ainda mais fracas. Mas pode-se igualmente aumentar esse número para conservar as implicações mais fortes.



Grafo implicativo

### Interpretação sucinta

#### 1ª Hierarquia coesitiva observada

Duas subestruturas aparecem, sem relação entre si.

A subclasse S1 corresponde à satisfação total (XSS), a subclasse S2 à satisfação parcial (XSP). Essas duas variáveis são colocadas nas extremidades dos caminhos implicativos, significando, portanto, que elas são mais as conseqüências dos sentimentos expressos.

A subclasse S1 reúne dois caminhos:

O primeiro corresponderia, antes de tudo, a futuras expectativas; o segundo corresponderia à satisfações de expectativas *a priori*.

A satisfação dessas últimas provém da manipulação de CHIC (IMA) e, particularmente, da descoberta de suas novas funcionalidades (INF) (EM variável típica), mas, igualmente, das aplicações e experiências associadas (IEXP). São essas relações que deram sentido à manipulação (IEXP  $\Rightarrow$  IMA) (G é a variável mais típica).

Notamos, portanto, que é a variável AMA que articula os dois caminhos. Pode-se interpretá-la de modo contraposto, assim: “se não estivesse satisfeito com as manipulações, então não desejaria ter outras, na ocasião de um próximo colóquio”.

O segundo mostra que a satisfação total (MBO) está muito ligada ao caráter estimulante das atividades (EST).

Observamos que são as variáveis suplementares EF (Ensino Fundamental) e G que parecem ser mais responsáveis por essa estrutura  $S_1$ .

A subclasse  $S_2$  é constituída de três implicações simples:

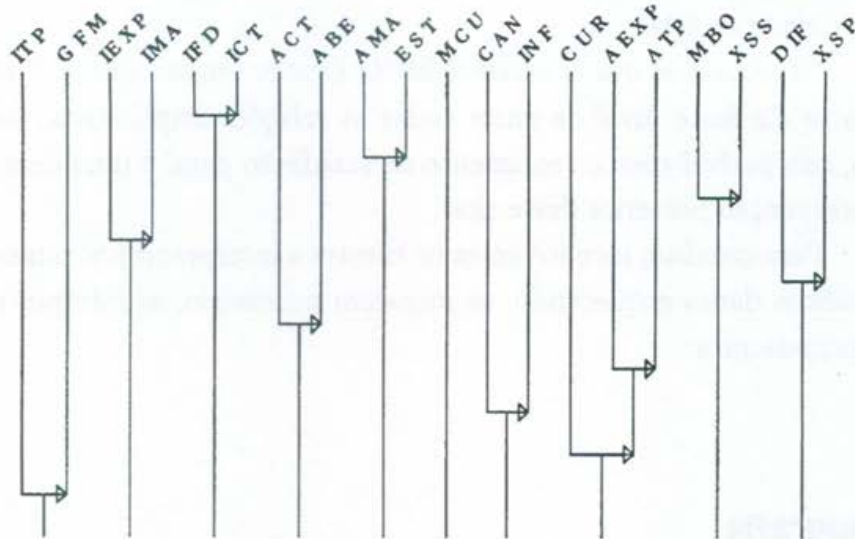
- ACT  $\Rightarrow$  ABF: que significa que a abertura do método implicativo está muito condicionada à teoria (tipicalidade dos GDO, doutores);
- IFD  $\Rightarrow$  ICT: significa que a importância atribuída à teoria teria vindo da forma dinâmica de sua apresentação (tipicalidade dos professores do ensino superior).
- DIF  $\Rightarrow$  XSP: nós já analisamos esta relação implicativa, mas a análise implicativa nos mostra em qual sentido devemos ler. É por que é difícil que a satisfação apareceu mais fraca (tipicalidade GME ou mestre). Uma interpretação psicológica poderia ser tentada aqui: os participantes que começaram a se engajar na pesquisa parecem ter dificuldade no caminho a percorrer, pois eles devem perceber os aportes teóricos como uma fronteira a ultrapassar para acessar ao estatuto de pesquisador. Este não é o caso dos mais jovens, não ainda implicados no uso do método, nem os mais avançados ou já iniciados.

## 2ª Hierarquia coesitiva observada

Observa-se que poucas grandes classes se formaram. Todas as classes, exceto uma, são constituídas de dois elementos. Esse fato é devido muito ao número relativamente importante de variáveis em relação ao número de sujeitos. Analisaremos, então, brevemente, cada uma dessas classes, respeitando a sua ordem de qualidade de coesão decrescente.



- IFD  $\Rightarrow$  ICT: a forma dinâmica e eficiente é fonte de interesse no que tange ao aporte teórico. Esta observação enfatiza a prudência com a qual um curso magistral deve ser pensado. Um curso não acompanhado de exemplo, nos quais os sujeitos encontram um domínio já conhecido e manipulado se torna desestimulante e sem grande efeito sobre a aprendizagem (tipicalidade de ES).



Árvore coesita

- AMA  $\Rightarrow$  EST: as futuras manipulações de CHIC poderiam ser estimulantes (tipicalidade de EM).
- MBO  $\Rightarrow$  XSS: a significação vem sobretudo do caráter total, completo (teoria – prática), adequado do colóquio (tipicalidade de EF).
- IEXP  $\Rightarrow$  IMA: as manipulações se justificam por que elas foram acompanhadas da experimentação não formal (tipicalidade de G).
- DIF  $\Rightarrow$  XSP: já foi analisada.
- ACT  $\Rightarrow$  ABE: já foi analisada.
- CAN  $\Rightarrow$  INF: lemos esta implicação de modo contraposto, pois ela é mais significativa nesse sentido: “se CHIC não tivesse apresentado novas funcionalidades, então não teríamos terminado tão cansados”. Sabe-se bem que são os desequilíbrios dos conhecimentos que são dolorosamente ressentidos (tipicalidade de GDA).

- $CUR \Rightarrow (AEXP \Rightarrow ATP)$ : esta relação é estatisticamente significativa. Assim, a curiosidade será tão satisfeita quanto a regra: “as experiências permitem esclarecer a relação teórico-prática” também será. O sujeito é curioso de ver como a teoria é utilizada como ferramenta para a prática (tipicalidade de GDA).

- $ITP \Rightarrow GFM$ : as demandas posteriores são principalmente induzidas pelo interesse em trabalhar mais a relação teórico-prática (tipicalidade de EM).

Observamos que a variável MCU (tempo muito curto) é neutra como se ela fosse dividida entre todas as relações implicativas. Isto é, claro, compatível com o sentimento de satisfação geral e uma demanda de intervenção posterior desse tipo.

Para concluir, incentivamos os leitores a interpretar por si mesmos os gráficos dados esquecendo, se julgaram necessário, as interpretações que propusemos.

## Bibliografia

- ALMOULOU, S. A. (1992). *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques*. Thèse. Université de Rennes 1.
- BODIN, A. (1997). Modèles sous-jacents à l'analyse implicative et outils complémentaires. *Prépublication IRMAR*, n. 97-32.
- BODIN, A. et GRAS, R. (1999). Analyse du préquestionnaire enseignants avant EVAPM-Terminales. *Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public*, n. 425, pp. 772-786.
- COUTURIER, R. et GRAS, R. (1999). Introduction de variables supplémentaires dans une hiérarchie de classes et application à CHIC. *Actes des 7èmes Rencontres de la Société Francophone de Classification*, Nancy, 15-17 septembre.
- GRAS, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Thèse d'Etat. Université de Rennes I.

- GRAS, R. et LARHER, A. (1992). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématique, Informatique et Sciences Humaines*, n. 20, pp. 5-31.
- GRAS, R. et RATSIMBA-RAJOHN, H. (1996). Analyse non symétrique de données par l'implication statistique. *RAIRO-Recherche Opérationnelle*, 30-3, AFCET, pp. 217-232.
- GRAS, R.; BRIAND, H. et PETER, P. (1996). Structuration sets with implication intensity. In: DIDAY, E.; CHEVALLIER, Y. et OPITZ, O. (eds.). *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis – OSDA 95*, Paris, Springer, pp. 147-156.
- GRAS R. et alii (1996). *L'implication Statistique*. Collection Associée à "Recherches en Didactique des Mathématiques". Grenoble, La Pensée Sauvage.
- GRAS R.; BRIAND H.; PETER P. et PHILIPPE J. (1997). Implicative statistical analysis. *Proceedings of International Congress I.F.C.S.*, 96. Kobé, Tokyo, Springer-Verlag, pp. 412-419.
- GRAS R.; KUNTZ P.; COUTURIER R. et GUILLET F. (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage (ECA)*, v. 1, n. 1-2, pp. 69-80.
- GRAS, R.; KUNTZ, P. et BRIAND, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines*, n. 154-155, pp. 9-29.
- GRAS, R.; DIDAY, E.; KUNTZ, P. et COUTURIER, R. (2001). Variables sur intervalles et variables-intervalles en analyse statistique implicative. *Actes du 8<sup>ème</sup> Congrès de la Société Francophone de Classification*. Université des Antilles-Guyane, Pointe-à-Pitre, 17-21 décembre, pp. 166-173.
- GRAS, R.; GUILLET, F.; GRAS, R. et PHILIPPE, J. (2002). Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables. *Extraction des connaissances et apprentissage, Hermès*, v. 1, n. 4, pp. 197-202.
- GRAS, R.; KUNTZ, P. et BRIAND, H. (2003). Hiérarchie orientée de règles généralisées en analyse implicative. *Extraction des Connaissances et apprentissage, Hermès*, pp. 145-157.

LERMAN, I. C.; GRAS, R. et ROSTAM, H. (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II. *Mathématiques et sciences Humaines*, n. 75, pp. 5-47.

Recebido em mar./2002; aprovado em abr./2002

## **Anexos**

### **Anexo 1**

#### **1ª Parte: Identificação (variáveis suplementares)**

Data: \_\_\_\_\_

- 1) Sexo:     Masculino M s     Feminino F s
- 2) Em que grau(s) de ensino leciona neste ano?  
 Ensino Fundamental EF s  
 Ensino Médio EM s  
 Ensino Superior ES s  
 Pós-Graduação PG s
- 3) Qual a sua formação acadêmica?  
 Graduado G s  
 Pós-Graduado – Mestrando GMA s  
 Pós-Graduado – Mestre GME s  
 Pós-Graduado – Doutorando GDA s  
 Pós-Graduado – Doutor GDO s

#### **2ª Parte: Opinião (variáveis principais)**

- 4) Suas expectativas foram satisfeitas:  
 Sim XSS     Parcialmente XSP     Não XSN
- 5) Quais os pontos da programação despertaram mais seu interesse?  
 teoria, conceitual ICT     novas funcionalidades de CHIC INF  
 manipulação de CHIC IMA     relação teoria-prática ITP  
 aplicações, experiências IEXP     forma dinâmica eficaz IFD
- 6) Quais aspectos você desejaria aprofundar mais?  
 teoria, conceitual ACT     manipulação de CHIC AMA  
 relação teoria-prática ATP     aplicações, experiências AEXP
- 7) Na lista abaixo, escolha a(s) expressão(ões) que melhor expressa(m) seus sentimentos a propósito do colóquio.  
 muito longo MLO     muito curto MCU     cansativo CAN  
 estimulante EST     curiosidade CUR     muito bom MBO  
 inútil INU     abertura ABE     difícil DIF  
 gostaria fazer mais GFM

Anexo 2

	Ocorrência	Média	Desvio padrão
XSS	21.00	0.84	0.37
XSP	4.00	0.16	0.37
ICT	10.00	0.40	0.49
INF	5.00	0.20	0.40
IMA	17.00	0.68	0.47
ITP	3.00	0.12	0.32
IEXP	11.00	0.44	0.50
IFD	4.00	0.16	0.37
ACT	9.00	0.36	0.48
AMA	8.00	0.32	0.47
ATP	16.00	0.64	0.48
AEXP	3.00	0.12	0.32
MLO	0.00	0.00	0.00
MCU	4.00	0.16	0.37
CAN	1.00	0.04	0.20
EST	18.00	0.72	0.45
CUR	11.00	0.44	0.50
MBO	13.00	0.52	0.50
INU	0.00	0.00	0.00
ABE	10.00	0.40	0.49
DIF	2.00	0.08	0.27
GFM	19.00	0.76	0.43
M s	6.00	0.24	0.43
F s	19.00	0.76	0.43
EF s	4.00	0.16	0.37
EM s	10.00	0.40	0.49
ES s	15.00	0.60	0.49
PG s	4.00	0.16	0.37
G s	4.00	0.16	0.37
GMA s	9.00	0.36	0.48
GME s	4.00	0.16	0.37
GDA s	8.00	0.32	0.47
GDO s	1.00	0.04	0.20