

A história contada em fragmentos cerâmicos: visão temporal do fundo de habitação núcleo de solo antropogênico 2 - sítio arqueológico Piracanjuba - Piraju, SP

Clélia Franco*
Nilton Nobuhiro Imai**
Vilma Tachibana**
Neide Faccio Barrocá**

FRANCO, C.; IMAI, N.N.; TACHIBANA, V.; BARROCÁ, N.F. A história contada em fragmentos cerâmicos: visão temporal do fundo de habitação núcleo de solo antropogênico 2 - sítio arqueológico Piracanjuba - Piraju, SP. *Revista do Museu de Arqueologia e Etnologia*, São Paulo, 18: 69-88, 2008.

Resumo: Este artigo pretende, pela aplicação da metodologia da descoberta de conhecimento indireto aos atributos dos fragmentos cerâmicos coletados no antigo fundo de habitação Núcleo de Solo Antropogênico 2 (NSA2) do sítio arqueológico Piracanjuba - Piraju SP, prover aos arqueólogos uma visão temporal do comportamento dos atributos dos fragmentos cerâmicos capaz de auxiliá-los no conhecimento das populações pretéritas que ali habitaram.

Palavras-chave: Descoberta de conhecimento - Mineração de dados - Estatística multivariada - Arqueologia.

1. Introdução

Durante os últimos anos tem se verificado um crescimento substancial na quantidade de dados armazenados em meios magnéticos ou não em todas as áreas da atuação humana. Estes dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou

analisados por especialistas utilizando métodos manuais tradicionais (Piatetsky-Shapiro 1991), tais como planilhas de cálculos e relatórios informativos operacionais, onde o especialista testa sua hipótese contra a base de dados. Por outro lado, sabe-se que grandes quantidades de dados equivalem a um maior potencial de informação. Entretanto, as informações contidas nos dados não estão caracterizadas explicitamente, são dados crus e não interessam quando estudados individualmente. Diante deste cenário, surge a necessidade de se explorar estes dados para extrair informação, conhecimento implícito, e utilizá-la no âmbito do problema.

Essa conjuntura levou ao surgimento de uma área de pesquisa conhecida como Descoberta

(*) Universidade Estadual de Maringá - PR.
cfranco@din.uem.br,

(**) Universidade Estadual Paulista "Júlio de Mesquita Filho" - Presidente Prudente - SP. N.N.Imai <nnimai@fct.unesp.br>; V. Tachibana <vilma@fct.unesp.br>; N.F. Barrocá <nffaccio@fct.unesp.br>.

ta de Conhecimento em Banco de Dados – “*Knowledge Discovery Database*” (KDD), tendo sido definida por Fayyad *et al.* (1996) como “*um processo não trivial de identificação de padrões válidos, novos, úteis e implicitamente presentes em grandes volumes de dados*”, cujo cerne é composto pelas técnicas de mineração de dados – “*Data Mining*” (DM).

“O arqueólogo não anda a procura de “antiguidades”, mas de conhecimentos, de realizações inteligentes.” (Moberg 1986).

Artefatos cerâmicos, dentro da arqueologia, são considerados como elementos fundamentais na obtenção de informações e no entendimento de como diferentes populações pretéritas viviam (Munita *et al.* 2001, 2004). Este fato aliado a descobertas de sítios arqueológicos – segundo o Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) o Brasil possui atualmente 20.000 sítios arqueológicos, alguns com grande número de artefatos cerâmicos (milhares) –, propiciam excelente arsenal na investigação e conhecimento de populações que outrora habitaram o País.

Existem estudos recentes (Baxter *et al.* 2003; Munita *et al.* 2001, 2004), onde a análise dos artefatos cerâmicos é efetuada por meio de técnicas de estatística multivariada com enfoque em sua composição mineralógica, bem como aqueles que por meio da estatística descritiva realizam um levantamento considerando aspectos morfológicos, topológicos e decorativos (Lopes e Barrocá 2001). Todavia, análise de artefatos cerâmicos considerando aspectos morfológicos, topológicos, decorativos e temporais por meio de técnicas de mineração de dados é incipiente.

Neste contexto, este trabalho apresenta resultados obtidos em um processo de descoberta de conhecimento no Núcleo de Solo Antropogênico 2 (NSA2) do Sítio Arqueológico Piracanjuba localizado no Município de Piraju – SP.

A datação do setor NSA2 realizada no ano de 2004 aponta para o ano de 1355 d.C. e possui atualmente o maior número de níveis escavados, oito, dentre os núcleos já identificados e que compõem o Sítio, contando com 3.344 fragmentos cerâmicos coletados, limpos e registrados. O número de níveis escavados

justifica a escolha do NSA2 para análise.

As técnicas estatísticas de mineração de dados aplicadas ao conjunto de atributos morfológicos, topológicos e decorativos dos artefatos cerâmicos coligidos objetivam estudar o comportamento dos mesmos ao longo dos níveis escavados detectando alterações significativas, de forma a propiciar aos arqueólogos pesquisadores uma visão temporal do setor NSA2 e contribuir para os estudos dos povos indígenas Guaranis pretéritos que habitaram o Vale do Rio Paranapanema, iniciados em 1968 com a criação do Projeto Paranapanema (ProjPar).

2. Sítio arqueológico Piracanjuba

O Sítio Arqueológico Piracanjuba está localizado no Município de Piraju, São Paulo, no médio curso do Rio Paranapanema com coordenadas UTM 656.750 E e 7.438.600 N, conforme ilustra a Fig. 1.

Para a área desse sítio foram realizados levantamentos planialtimétricos, sondagens, perfis, trincheiras e decapagens. O sítio pode ser classificado como razoavelmente conservado, as alterações na superfície do terreno chegam a ser significativas, comprometendo parte do registro arqueológico. Os estratos arqueológicos chegam a ser afetados por sulcos de erosão, bioturbação e pela ação de implementos agrícolas ou outras atividades antrópicas.

O trabalho de pesquisa nesse sítio foi iniciado no ano 2000, e faz parte do Programa de Salvamento Arqueológico da UHE Piraju, coordenado por José Luiz de Moraes do Museu de Arqueologia e Etnologia (MAE) / Universidade de São Paulo (USP), tendo sido coletados até o momento 15.288 fragmentos cerâmicos em quatro fundos de habitações e entornos: Núcleo de Solo Antropogênico 1 (NSA1), Núcleo de Solo Antropogênico 2 (NSA2), Núcleo de Solo Antropogênico 3 (NSA3), Núcleo de Solo Antropogênico 4 (NSA4) além de coleta de superfície de Estacas e Leirias.

Segundo Moraes e Barrocá (2004) o material cerâmico recuperado pertence, de modo geral, à denominada tradição Tupiguarani, encontrando-se devidamente limpo, registrado e

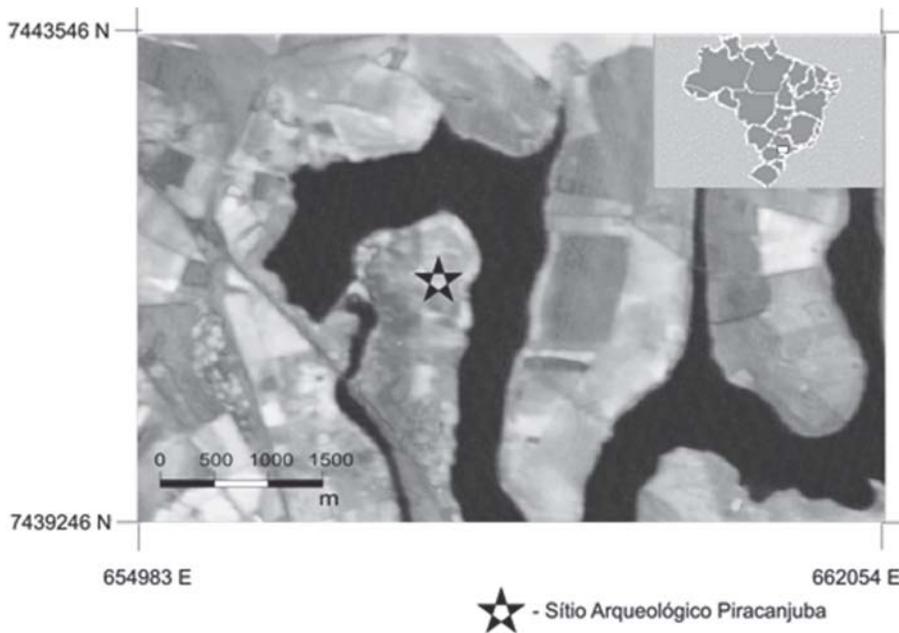


Fig. 1. Localização do Sítio Arqueológico Piracanjuba em imagem CBERS, bandas 2, 3, 4, realçada por Análise das Principais Componentes - 08/09/2004.

acervado no Centro de Pesquisas Mario Neme da Universidade de São Paulo.

Estudos que contemplam a vasilha enquanto objeto de análise, bem como a análise dos fragmentos através de técnicas da estatística descritiva constam do trabalho intitulado “O trabalho do barro no sítio Arqueológico Piracanjuba - Piraju, SP.” (Morais e Barrocá, 2004).

3. Descoberta de conhecimento em banco de dados

O processo capaz de descobrir conhecimento em banco de dados chama-se KDD (Knowledge Discovery Database). O processo de KDD foi proposto em 1989 para referir-se às etapas que produzem conhecimento a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que transforma dados em informações.

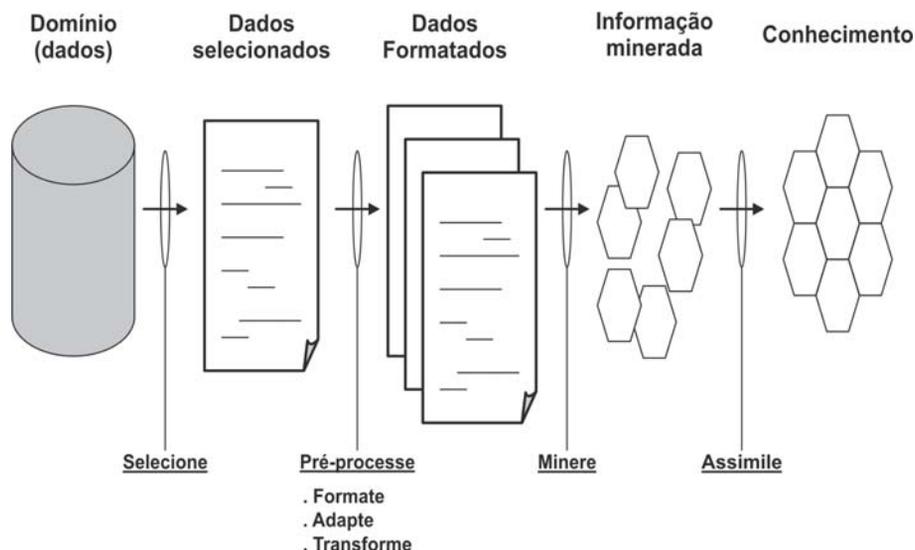
Descoberta de Conhecimento em Banco de Dados é um processo interativo e iterativo, podendo ser tipificado em quatro passos: seleção de dados, pré-processamento dos dados, extra-

ção de informações e a assimilação da informação minerada conforme ilustra a Fig. 2.

Dados crus coletados de um ou mais repositórios de dados dificilmente se encontram prontos para a mineração devendo ser preparados para tal, o que envolve as fases de pré-processamento e tratamento dos dados. Algumas operações comuns desta fase são: escolha dos dados relevantes, padronização, eliminação de redundância, simplificação da estrutura e a limpeza de dados de maneira a se obter uma base representativa no formato atributo/valor (Bigolin 2000).

A mineração de dados é o passo onde são aplicadas técnicas voltadas para atingir objetivos específicos dentro da área de atuação humana da aplicação (Groebel e Gruenwald 1999).

A escolha das técnicas de mineração de dados é função da tarefa específica a ser executada e dos dados disponíveis para a análise, sendo que a familiaridade com as técnicas é indispensável para a melhor abordagem do problema apresentado. Desta forma, em um processo completo de descoberta de conhecimento podem ser utilizadas diversas técnicas de mineração de



Fonte: Mendonça Neto, 2001.

Fig. 2. O Processo da Descoberta de Conhecimento.

dados, especificadas em função do nível de conhecimento a ser extraído e da área de conhecimento humano da aplicação. As técnicas dão suporte a um conjunto de tarefas: associações, padrões seqüenciais, séries temporais similares, classificação e regressão e agrupamentos, que diferem entre si pelo tipo de problema que são capazes de resolver (IBM 1996).

O último passo da descoberta do conhecimento envolve a interpretação e assimilação da informação minerada expressa por padrões ou modelos.

Convém ressaltar que a apresentação visual dos resultados obtidos pelas técnicas de mineração de dados possibilita melhor interpretação e assimilação dos resultados (Mendonça Neto e Almeida 2001) porque padrões minerados geralmente são complexos e variam em muitas dimensões, o que dificulta a sua interpretação quando apresentados na forma de tabelas ou textos.

4. Materiais e métodos

4.1 Metodologia

A metodologia proposta encontra-se descrita em Berry e Linoff (1997) e é conhecida

como descoberta de conhecimento indireto. Os passos que a integram são descritos sumariamente abaixo:

- identificar os repositórios de dados disponíveis;
- preparar os dados para análise;
- selecionar as técnicas apropriadas com base nas características dos dados e na tarefa de mineração de dados;
- usar os resultados para descobrir estruturas escondidas nos dados;
- identificar alvos em potencial para a descoberta de conhecimento indireto;
- gerar hipóteses a serem testadas

4.2 Materiais

4.2.1 Dados arqueológicos do sítio Piracanjuba-Piraju-SP, contendo 15.288 registros de fragmentos cerâmicos em planilha eletrônica, com atributos registrados em grupos quanto à natureza das informações, ilustrados a seguir: quanto à identificação, Quadro 1; quanto à proveniência, Quadro 2; quanto ao material constituinte, Quadro 3; quanto à fabricação, Quadro 4; quanto à origem (parte do vaso) e forma do fragmento, Quadro 5; e quanto ao uso e conservação, Quadro 6.

Quadro 1

Atributos registrados quanto à identificação		
Atributo	Número da Peça	Número do Vaso
Abrev.	PEÇA	VASO
Tipo	nomi.	nomi.

Quadro 2

Atributos registrados quanto à proveniência										
Atributo	Setor	Estaca	Leiria	Quadra	Quadrícula	Trincheira	Corte de Verificação	Área de Decapagem	Perfil	Nível
Abrev.	Setor	Estaca	Leiria	QDA	QDI	T	CV	D	P	NI
Tipo	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.

Quadro 3

Atributos registrados quanto ao material constituinte						
Atributo	Tipo do Antiplástico	Material mineral (mm)	Cariapé (mm)	Caco Moído (mm)	Carvão (mm)	Concha Moída (mm)
Abrev.	TP	MN	CR	CM	CV	CC
Tipo	nomi.	ordinal	ordinal	ordinal	ordinal	Ordinal

Quadro 4

Atributos registrados quanto à fabricação								
Atributo	Espessura da parede (cm)	Queima	Dureza	Cor da Argila	Tratamento Superficial	Decoração Interna	Decoração Externa	Técnica de manufatura
Abrev.	ES.PAR	QUEIMA	DUREZ	COR.ARG	TRA.SUP.	DECORI	DECORE	TEC.MAN.
Tipo	ordinal	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.

Quadro 5

Atributos registrados quanto à origem (parte do vaso) e forma do fragmento							
Atributo	Classe	Ângulo da Parede	Bolota de Argila	Forma de Lábio	Forma da Borda	Forma da Vasilha	Contorno do recipiente
Abrev.	CLASSE	ANG. PAR	BOLOT.ARG.	LÁBIO	BORDA.	FORMA	CONTOR
Tipo	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.	nomi.
Atributo	Diâmetro da Boca (mm)	Altura do Vaso (mm)	Largura da Garganta (mm)	Volume do Vaso (cm ³)	Tipo de Base	Diâmetro da Base (mm)	Ângulo da Base (graus)
Abrev.	D. BOCA.	ALT. VASO	LAR. GAR	V. VASO	BASE	D. BASE	AG. BASE
Tipo	ordinal	ordinal	ordinal.	Ordinal	nomi.	ordinal	ordinal

Quadro 6

Atributos registrados quanto ao uso e conservação		
Atributo	Marcas de Uso	Estado de Conservação
Abrev.	MAR.USO	EST.CONS.
Tipo	nomi.	nomi.

4.2.2 Software estatístico para processamento, geração de resultados e gráficos.

5. Resultados e discussão

5.1 Identificação do repositório de dados

Uma das bases da descoberta do conhecimento é a premissa de que dados necessários à mineração de dados em determinado domínio estão disponíveis em algum repositório de dados, o analista deve ter garantias quanto à integridade e a confiabilidade dos mesmos, realizando então uma pré-seleção dos dados a serem utilizados pelo algoritmo de mineração.

O repositório de dados foi fornecido pela equipe de arqueólogos do ProjPar, planilha eletrônica contendo os registros de fragmentos cerâmicos do sítio Piracanjuba - Piraju, SP. Do mesmo foram selecionados os registros pertencentes ao NAS2 sendo posteriormente excluídos os atributos que não possuíam valores registrados

bem como aqueles cuja distribuição de frequência apresenta mais de 95% pertencentes à categoria desconhecida, perfazendo então um total de 3307 registros de fragmentos cerâmicos.

5.2 Seleção dos atributos para análise

A seleção teve por base a metodologia desenvolvida por Robrahn (1996), posteriormente adaptada por Faccio (1998) em atendimento às especificidades da área do ProjPar e à consistência dos dados registrados para cada atributo.

Segundo Faccio (1998), "A análise do material cerâmico compreende a verificação de classes de atributos tecnológicos, estilísticos e morfológicos, além das marcas de uso e do estado de conservação". Portanto, no sentido de integrar este estudo aos realizados na área do ProjPar, foram considerados como atributos relevantes e suas classes os definidos por Faccio (1998).

Os atributos selecionados constam no Quadro 7 e a descrição das categorias no Quadro 8.

Quadro 7

Atributos selecionados para o processo da Descoberta de Conhecimento

Especificação	Atributo	Abrev.	Tipo	Categorias
Proveniência	Nível	NI	Qualitativo categórico	0 - 8
Origem (parte do vaso)	Classe	CL	Qualitativo categórico	1-6, 16, 19, 21, 23
Material Constituinte	Antiplástico	TP	Qualitativo categórico	1 e 13
	Material Mineral	MN	Quantitativo contínuo	(mm)
	Caco Moido	CM	Quantitativo contínuo	(mm)
Fabricação	Espessura da Parede	EP	Quantitativo contínuo	(mm)
	Queima	Q	Qualitativo categórico	0 - 6
	Tratamento Superficial	TS	Qualitativo categórico	3 - 7, 9-12, 15, 16, 18, 19, 20
	Decoração Interna	DI	Qualitativo categórico	0, 1, 4, 9 - 12, 14-16, 21, 31, 32, 34
	Decoração Externa	DE	Qualitativo categórico	0, 1-5, 10, 11, 14-16, 19 - 21, 25, 26, 30, 34, 36
	Técnica de Manufatura	MAN	Qualitativo categórico	1,3

Quadro 8

Descrição das categorias dos atributos_categorias selecionadas para análise			
Atributo	Abrev.	Categorias	Descrição
Proveniência	NI	0	Do solo
		1	0,1 - 9,99 cm
		2	10 - 19,99 cm
		3	20 - 29,99 cm
		4	30 - 39,99 cm
		5	40 - 49,99 cm
		6	50 - 59,99 cm
		7	60 - 69,99 cm
		8	70 - 70,99 cm
Classe	CL	1	Parede
		2	Borda
		3	Base
		4	Base, parede, borda
		5	Parede angular
		6	Parede e base
		16	Ombro
		19	Parede com furo de suspensão
21	Borda c/ suporte para tampa		
23	Borda/parede angular		
Antiplástico	TP	1	Mineral
		13	Mineral e caco moído
Material Minera	MN	–	Quantitativo
Caco Moído	CM	–	Quantitativo
Espessura da Parede	EP	–	Quantitativo
Queima	Q	0	Não identificada
		1	Seção transversal sem presença de núcleos, com cor uniforme, variando laranja-tijolo ao amarelo
		2	Seção transversal sem presença de núcleos, com cor uniforme variando cinza-claro ao pardo
		3	Seção transversal com presença de núcleo central escuro e uma camada interna e externa clara
		4	Seção transversal sem presença de núcleos, com cor uniforme variando do cinza-escuro ao preto
		5	Seção transversal com uma camada clara na parte externa e uma camada escura na interna
		6	Seção transversal com uma camada clara na parte interna e uma camada escura na externa
Tratamento Superficial	TS	3	Alisamento interno sem alisamento externo
		4	Alisamento interno e externo
		5	Polimento interno/alisamento externo
		6	Polimento externo/alisamento interno

Quadro 8 (cont.)

Descrição das categorias dos atributos_categorias selecionadas para análise			
Atributo	Abrev.	Categorias	Descrição
Tratamento Superficial	TS	7	Polimento interno e externo
		9	Lustro externo/ alisamento interno
		10	Lustro interno/alisamento externo
		11 e 12	Sem descrição
		15	Lustro interno/ sem tratamento externo
		16	Polimento externo/ sem tratamento interno
		18, 19 e 20	Sem descrição
		0	Não identificada
		1	Liso
		2	Entalhado
Decoração Interna e Decoração Externa	DI e DE	3	Ungulado
		4	Inciso
		5	Corrugado
		6	Escovado
		7	Ponteadado
		9	Engobo preto
		10	Engobo vermelho
		11	Engobo branco
		12	Engobo preto/vermelho
		14	Engobo vermelho/branco
		15	Engobo laranja
		16	Pintado
		19	Digitado
		20	Serrungulado
		21	Engobo branco/inciso no contorno da garganta
25	Engobo branco associado ao inciso		
26	Pintado associado ao inciso		
30	Roletado		
31	Engobo branco/laranja		
32	Engobo vermelho/pintado		
34, 36	Sem descrição		
Técnica de Manufatura	MAN	1	Roletado (acordelado)
		3	Modelado à mão

A codificação para os atributos e as categorias tem a forma abreviatura do atributo - categoria.

5.3 Pré-processamento

Antes de escolher a técnica e o nível de mineração de dados são necessárias a análise e compreensão dos mesmos, para tanto a estatística

descritiva (média, moda, desvio padrão, a distribuição dos dados) e as ferramentas de visualização, onde entende-se visualização como o uso de computadores para representação gráfica de dados (Santos *et al.* 2000), constituem-se em poderosas armas pois possibilitam conhecimento *a priori* do conjunto de dados.

Dentre as ferramentas de visualização bastante utilizadas encontram-se os gráficos de

coluna, os de distribuição de frequência (histogramas) e os de dispersão em duas ou três dimensões de diferentes atributos que mostram claramente os valores omissos, os valores excepcionais (*outliers*) e os relacionamentos dos atributos sob análise.

A análise preliminar por meio de gráficos de coluna com o objetivo de estudar a distribuição, a variabilidade e determinar o senso comum (moda) dos atributos mostrou que:

- os fragmentos em sua maioria, 97%, encontram-se em bom estado de conservação, tendo como material constituinte preponderante o material mineral com granulometria entre 0,1 e 0,26mm + caco moído com granulometria menor que 0,1 mm, sendo que a espessura da parede em 37% dos fragmentos varia entre 0,5 e 1,0 cm;

- as queimas mais comuns, encontradas em 77% dos fragmentos, são as que apresentam seção transversal com presença do núcleo central escuro e uma camada interna e externa clara (preponderante) e a que apresenta seção transversal com uma camada clara na parte externa e uma camada escura na interna;

- a decoração lisa é a mais encontrada tanto interna quanto externamente, sendo que dentre das demais categorias de decoração a de engobo branco é a mais presente. O tratamento de superfície preponderante é o alisado interna e externamente, e a técnica de manufatura em 99,5% das peças é a do roletado;

- quanto aos atributos pode-se dizer que sua distribuição não se ajusta à curva normal, sendo em sua grande maioria categorizados.

A diversidade nos tipos de atributos e as restrições da técnica de mineração a ser aplicada conduziram a discretização dos atributos quantitativos material mineral, caco moído e espessura da parede. Foram utilizados para tanto os quartis 1 (Q1) e 3 (Q3) e a mediana (Md), assim, para valores iguais ou menores que Q1 foi atribuída a categoria 1, maior que Q1 e menor ou igual à Md categoria 2, maior que Md e menor ou igual a Q3

categoria 3 e maior que Q3 categoria 4, além da categoria 0 que indica ausência.

5.4 Mineração de dados

As técnicas estatísticas estão dentre as mais utilizadas na mineração de dados, pois é uma área do conhecimento já sedimentada e amplamente estudada, disponibilizando aos usuários grande número de algoritmos já implementados e otimizados (Koperski *et al.* 1998).

Segundo Von Wangenheim (2004), do ponto de vista da estatística, as técnicas úteis para reconhecimento e descoberta em ambientes onde os fenômenos são descritos/baseados em uma grande variedade de dados são conhecidas como Análise Exploratória de Dados (AED) ou estatística exploratória, dentre elas encontramos a análise dos resíduos padronizados.

Os resíduos padronizados apresentam os resíduos (diferenças entre o valor observado e o esperado) em uma forma padronizada expressos em unidades de desvio padrão, podendo instruir pontos de corte para um nível de significância de excesso ou falta de ocorrências. (Haberman 1978).

Segundo Pereira (2004) “os resíduos padronizados representam valores de relação biunívoca com probabilidades de ocorrência, isto é, valores maiores que 1,96 ou menores que -1,96 têm pequenas chances de ocorrência ($\pm 2,5\%$)”, e podem instruir pontos de corte para a falta ou excesso de ocorrências, permitindo a distinção entre as **casuais** e as **causais** que são a causa das associações/relações detectadas pelo teste do χ^2 .

Dessa forma, a mineração de dados pretende: traçar o perfil dos atributos_categorias considerados significativos (que apresentam excesso/falta de ocorrência) ao longo dos níveis escavados e caracterizar os níveis em termos da concentração/falta de fragmentos dos diversos atributos categorias;

5.4.1 Resíduos padronizados

O ponto de partida da análise da tabela dos resíduos padronizados é uma tabela de contingência como a ilustrada na Fig. 3.

X/Y	y ₁	...	y _j	...	y _c	FREQÜÊNCIA MARGINAL DE X
x ₁	n ₁₁	n _{1c}	n _{1.}
...
x _i	n _{i1}	...	n _{ij}	...	n _{ic}	n _{i.}
...
x _r	n _{r1}	...	n _{rj}	...	n _{rc}	n _{r.}
FREQÜÊNCIA MARGINAL DE Y	n _{.1}	...	n _{.j}	...	n _{.c}	n _{..}

Fig. 3. Tabela de contingência.

Essas tabelas são constituídas de dados relativos ao número de elementos existentes nos atributos combinados de X e Y, sendo:

X - conjunto de atributos {x₁, ..., x_r};

Y - conjunto de atributos {y₁, ..., y_c};

n_{i.} - total de unidades de x_i em Y;

n_{.j} - total de unidades de y_j em X;

Freqüência marginal de X - $n_{i.} = \sum_j n_{ij}$;

Freqüência marginal de Y - $n_{.j} = \sum_i n_{ij}$;

Freqüência total - $n_{..} = \sum_i \sum_j n_{ij}$

Não são considerados os valores absolutos, mas as correspondências entre as freqüências relativas, isto é, os valores relativos cuja soma é igual a 1, ou seja, os valores originais são transformados de modo a poder ser interpretados como probabilidades condicionais.

Dessa forma, uma primeira caracterização das distribuições de linhas e colunas pode ser obtida pelos:

1. perfis que descrevem as distribuições condicionadas, sendo:

perfis linha: descrevem as distribuições condicionadas de uma linha segundo as colunas, como em a.

$$\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \dots, \frac{n_{ij}}{n_{i.}}, \dots, \frac{n_{ic}}{n_{i.}} \quad (02)$$

b. perfis coluna: descrevem as distribuições condicionadas de uma coluna segundo as linhas

$$\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \dots, \frac{n_{ij}}{n_{.j}}, \dots, \frac{n_{rj}}{n_{.j}} \quad (03)$$

2. perfis marginais (massa) que descrevem a distribuição marginal:

a. das linhas condicionadas pelo total geral da tabela

$$\frac{n_{1.}}{n_{..}}, \frac{n_{2.}}{n_{..}}, \dots, \frac{n_{i.}}{n_{..}}, \dots, \frac{n_{r.}}{n_{..}} \quad (04)$$

b. das colunas condicionadas pelo total geral da tabela

$$\frac{n_{.1}}{n_{..}}, \frac{n_{.2}}{n_{..}}, \dots, \frac{n_{.j}}{n_{..}}, \dots, \frac{n_{.c}}{n_{..}} \quad (05)$$

A validação da análise dos resíduos padronizados em uma tabela de contingência é feita por meio do teste do chi-quadrado de Pearson, das comparações dos perfis linha e coluna com os perfis marginais correspondentes, expressão (06), determinando a probabilidade da associação global entre as linhas e colunas.

Para o teste: a hipótese nula H0 é: X e Y são independentes, e a hipótese alternativa H1 é: X e Y são dependentes. Convém observar que se H0 é certa todos os perfis linha/coluna são iguais entre si e iguais ao perfil marginal das linhas/colunas,

$$G^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^r \sum_{j=1}^c \frac{n_i \left(\frac{n_{ij}}{n_i} - \frac{n_j}{n_{..}} \right)^2}{\frac{n_j}{n_{..}}} = \sum_{i=1}^r \sum_{j=1}^c \frac{n_j \left(\frac{n_{ij}}{n_j} - \frac{n_i}{n_{..}} \right)^2}{\frac{n_i}{n_{..}}} \quad (06)$$

sendo:

$$e_{ij} = E[n_{ij} | H_0 \text{ é certa}] = \frac{n_i n_j}{n_{..}}$$

Valores pequenos de G^2 significam que os valores de n_{ij} e e_{ij} estão próximos e, portanto, que H_0 é certa. Pelo contrário, valores grandes evidenciam que H_0 é falsa. O valor p do teste é dado por

$$p[\chi^2_{(r-1)(c-1)} \geq G^2] \quad (07)$$

Para um nível de significância $0 < \alpha < 1$, rejeita H_0 se o valor de p for menor ou igual a α .

Se a hipótese nula for rejeitada, então a natureza das associações pode ser explorada examinando-se as porcentagens linhas e/ou colunas das tabelas de contingência. (Bartholomew *et al.* 2002).

O teste do χ^2 é hábil em determinar a probabilidade da associação global entre as linhas e colunas, mas não fornece nenhuma informação a respeito de quais são as associações individuais significativas entre pares de linhas/colunas da matriz de dados. Portanto, quando o teste do χ^2 detecta a existência de relações/associações em uma tabela de contingência é interessante a investigação dos resíduos padronizados dado por:

$$P_{resij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}} \sqrt{\left(1 - \frac{n_i}{n_{..}}\right) \left(1 - \frac{n_j}{n_{..}}\right)}} \quad (08)$$

sendo:

n_{ij} = freqüência observada;
 e_{ij} = freqüência esperada;
 n_i = total da coluna;
 n_j = total da linha;
 $n_{..}$ = total geral da tabela,

5.4.2 Estudos preliminares

A análise preliminar da tabela de contingência mostrou a baixa freqüência de algumas categorias de atributo (> 5), o que segundo Pereira (2004) desautoriza a aplicação do teste

do χ^2 e por conseqüência o seu estudo pelos resíduos padronizados.

Como conseqüência optou-se pela geração de duas tabelas de contingência analisadas em etapas distintas. A etapa 1 contempla a tabela designada por comum gerada pelos registros de fragmentos cujas categorias de atributos possuem no mínimo seis ocorrências. A etapa 2, a tabela designada por raríssimos gerada pelos registros de fragmentos que possuem ao menos uma categoria de atributo com freqüência inferior a 6. O Quadro 9 apresenta as características das tabelas.

O procedimento adotado em contrapartida ao simples descarte dos registros com categorias de atributos de baixa freqüência deve-se à importância que os mesmos possuem em pesquisas arqueológicas.

Na etapa 1:

- será verificada a viabilidade do estudo dos resíduos padronizados dos atributos pelo teste do χ^2 ;
- os resíduos padronizados significativos serão apresentados em gráficos de linha (unidades de desvio padrão X níveis escavados) possibilitando o estudo dos perfis dos atributos categoriais ao longo dos níveis, bem como a identificação dos atributos categoriais "causais", ou seja, aqueles responsáveis pelas associações globais detectados pelo teste do χ^2 . O nível de significância adotado é de 2,5%, isto é, pontos de corte 1,96 para excesso e -1,96 para falta.

A etapa 2 contempla a tabela raríssimos onde o comportamento dos atributos categoriais de baixa freqüência (< 5) será analisado visualmente por meio de gráficos de barra.

5.4.3 Etapa 1 - tabela comum

Os valores esperados e calculados do χ^2 para os atributos a um nível de significância de 2,5%

Quadro 9

Características das tabelas para análise					
Nome da tabela	Número de fragmentos	Níveis de ocorrência	Atributos	Abreviatura do Atributo	Categorias
Comum	3206	8	Classe	CL	1, 2, 3, 5
			Antiplástico	TP	1, 13
			Mat. Mineral	MNC	1 - 3
			Caco Moído	CMC	0 - 3
			Esp. da Parede	EPC	1 - 4
			Queima	Q	0 - 6
			Deco. Interna	DI	0, 1, 10, 11, 14, 15, 16
			Deco. Externa	DE	0, 1, 4, 10, 11, 15, 16
			Trat. Superficial	TS	4, 5, 6, 7, 11, 12
			Téc. de Man.	MAN	1, 3
Raríssimos	100	8	Classe	CL	1 - 6, 16, 19, 21, 23
			Antiplástico	TP	1, 13, 19
			Mat. Mineral	MNC	1 - 3
			Caco Moído	CMC	0 - 3
			Esp. da Parede	EPC	1 - 4
			Queima	Q	1 - 6
			Deco. Interna	DI	0, 1, 4, 9 - 12, 15, 16, 21, 31, 32, 34
			Deco. Externa	DE	0 - 5, 10, 11, 14 -16, 19 -21, 25, 26, 30 - 34, 36
			Trat. Superficial	TS	0, 1, 3 -7, 9 - 14, 16 -18, 20, 21
			Téc. de Man.	MAN	1, 3

são apresentados na Tabela 1; pode-se concluir que a distribuição das categorias dos atributos queima e decoração interna nos níveis escavados guarda certa homogeneidade desabilitando o estudo dos resíduos padronizados. Para os demais atributos o estudo é válido.

Os gráficos de linha dos atributos considerados significantes (possuem ao menos uma categoria apresentando excesso/falta) expresso em unidades de desvio padrão nos níveis escavados são apresentados a seguir.

O Gráfico 1 apresenta o atributo classe onde se observa a concentração de fragmentos de parede angular no nível 5, de fragmentos de base no nível 6 e de fragmentos de borda no nível 7,

outrossim, cabe ressaltar que a concentração dos fragmentos pertencentes a essas categorias ocorrem nos níveis inferiores.

Tabela 1

Atributos, valor esperado e calculado do χ^2 e significância - tabela comum

Atributo	χ^2 esperado (p=0,025)	χ^2 calculado	significativo
Classe	39,364	49,511	sim
Material mineral	28,845	34,726	sim
Caco moído	39,364	74,692	sim
Espessura da Parede	28,845	41,268	sim
Queima	69,022	61,699	não
Tratamento superficial	59,341	722,313	sim
Decoração Interna	69,022	60,528	não
Decoração Externa	69,022	86,408	sim
Técnica de manufatura	17,534	19,392	sim

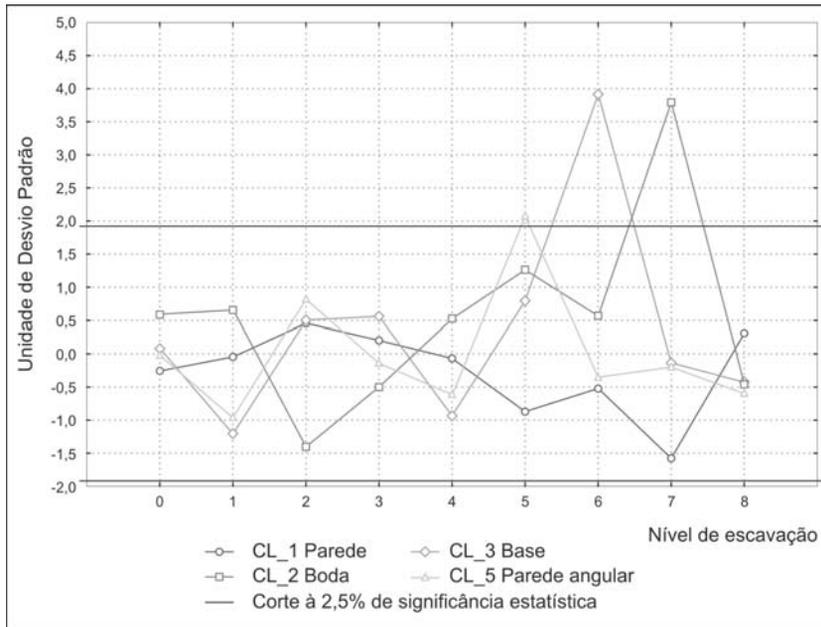


Gráfico 1. Gráfico de linha dos resíduos padronizados do atributo classe - tabela comum.

O Gráfico 2 ilustra o atributo granulometria do material mineral, onde se observa a concentração de fragmentos com material mineral de granulometria média no nível 3 e a sua escassez no nível 1. Os fragmentos com material mineral

de granulometria fina ou grossa apresentam homogeneidade nos níveis escavados. Interessante observar a contraposição da linha do material mineral de granulometria média e o de granulometria fina.

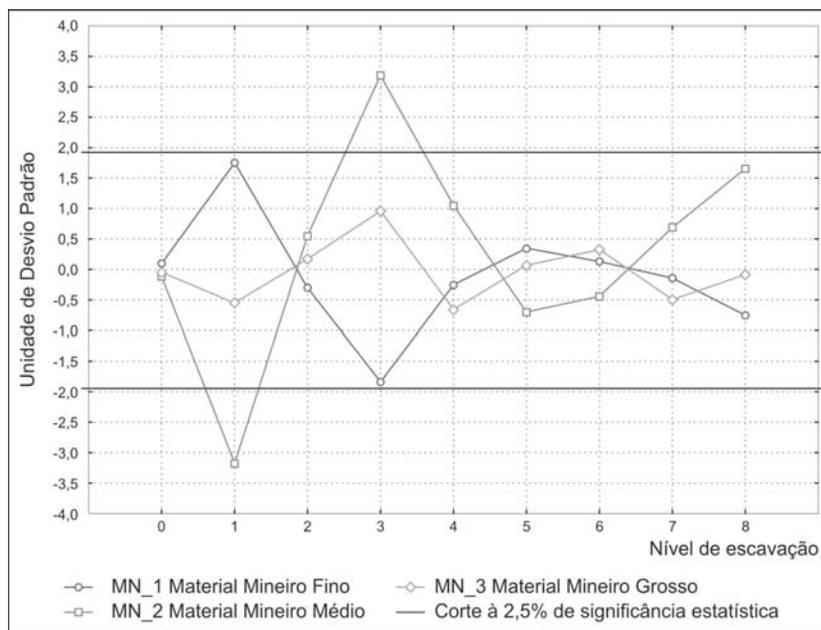


Gráfico 2. Gráfico de linha dos resíduos padronizados do atributo material mineral - tabela comum.

Quanto ao atributo caco moído ilustrado no Gráfico 3, podemos dizer que existe concentração de fragmentos com caco moído grosso no nível 5 e escassez no nível 1. O nível 5 apresenta ainda escassez de caco moído de granulometria fina. Salientamos o desenvolvimento da linha do caco moído grosso decrescente em relação aos níveis mais superficiais e contraposta à de caco moído fino.

O atributo tratamento superficial ilustrado no Gráfico 4, mostra a forte concentração de fragmentos cerâmicos com tratamento superficial polimento externo/interno e com polimento externo/alisamento interno no nível 6 bem como a concentração de fragmentos com tratamento superficial polimento interno/externo, polimento interno/alisamento externo e polimento externo/alisamento interno no nível 8. Os níveis 6 e 8 apresentam ainda a escassez de fragmentos com alisamento interno/externo.

O nível 4 apresenta concentração de fragmentos com tratamento superficial polimento interno/alisamento externo e o nível 1 da categoria 11 - sem descrição.

Interessante observar que os fragmentos que possuem como tratamento superficial o polimen-

to em uma ou ambas as faces encontram-se concentrados nos níveis inferiores.

Quanto ao atributo espessura de parede, ilustrado no Gráfico 5 cabe ressaltar o excesso de fragmentos de espessura grossa no nível 1 e a sua falta no nível 0.

O atributo técnica de manufatura ilustrado no Gráfico 6 tem como relevante a concentração de fragmentos modelados à mão no nível 6.

Do atributo decoração externa ilustrado no Gráfico 7, destacamos a decoração incisa que se encontra concentrada nos níveis 0, 5 e 7 apresentando escassez no nível 1.

5.4.4 Etapa 2 - tabela raríssimos

O Gráfico de barra 8 (a) e 8 (b) apresenta a distribuição dos fragmentos que possuem ao menos um atributo_categoria considerado raríssimo quanto à classe e ao tratamento superficial.

Quanto à classe observa-se que os fragmentos:

- de borda/parede angular (CL_23) apresentam o maior número de ocorrências contíguas a partir do nível 4 em direção aos níveis superiores;

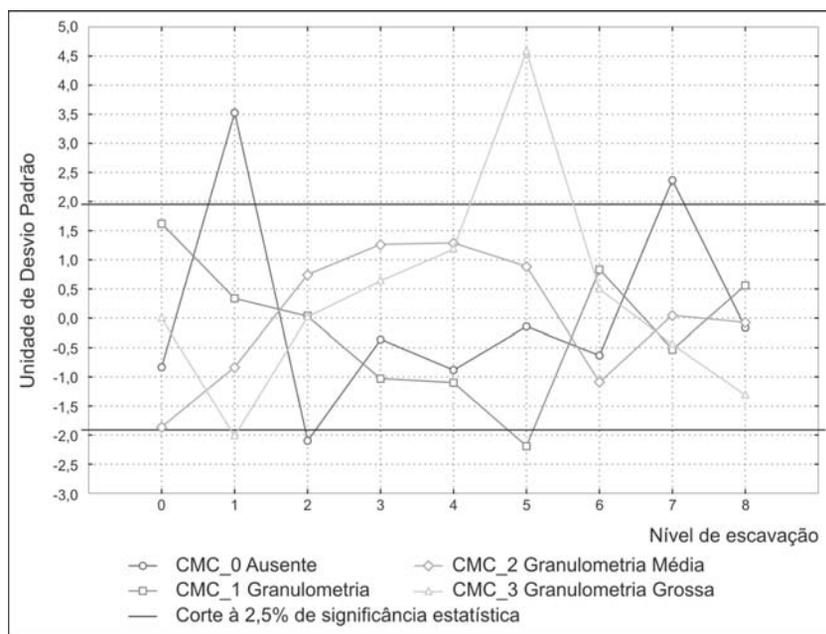


Gráfico 3. Gráfico de linha dos resíduos padronizados do atributo caco moído - tabela comum.

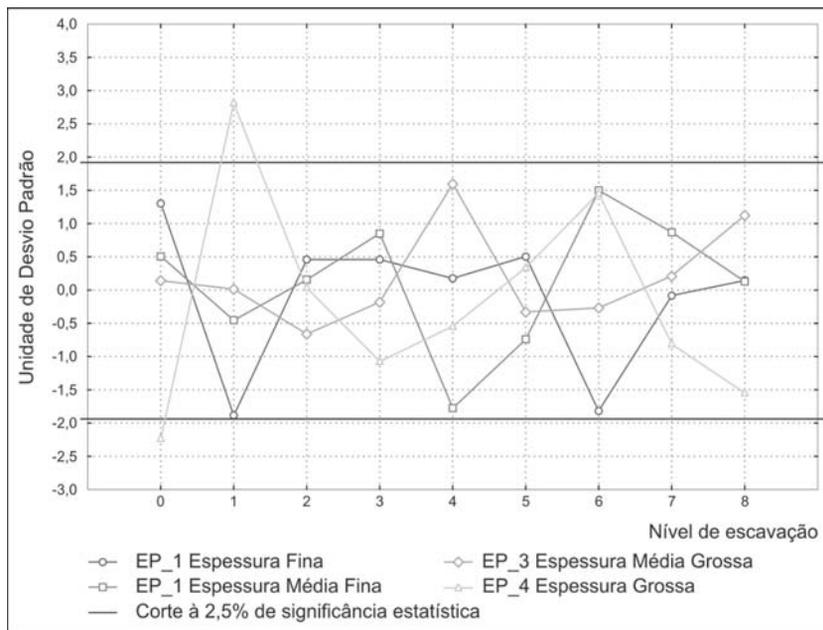


Gráfico 4. Gráfico de linha dos resíduos padronizados do atributo tratamento superficial.

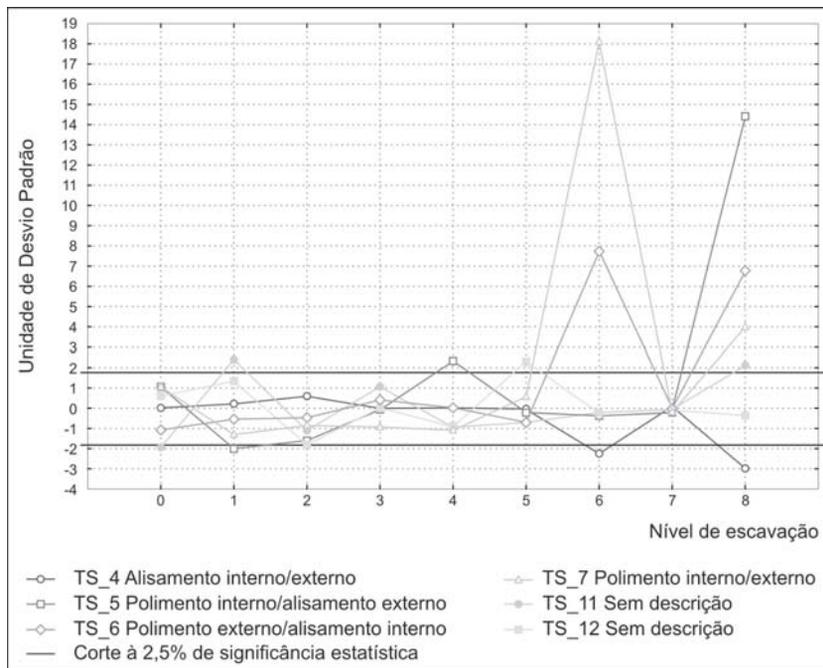


Gráfico 5. Gráfico de linha dos resíduos padronizados do atributo espessura da parede-tabela comum.

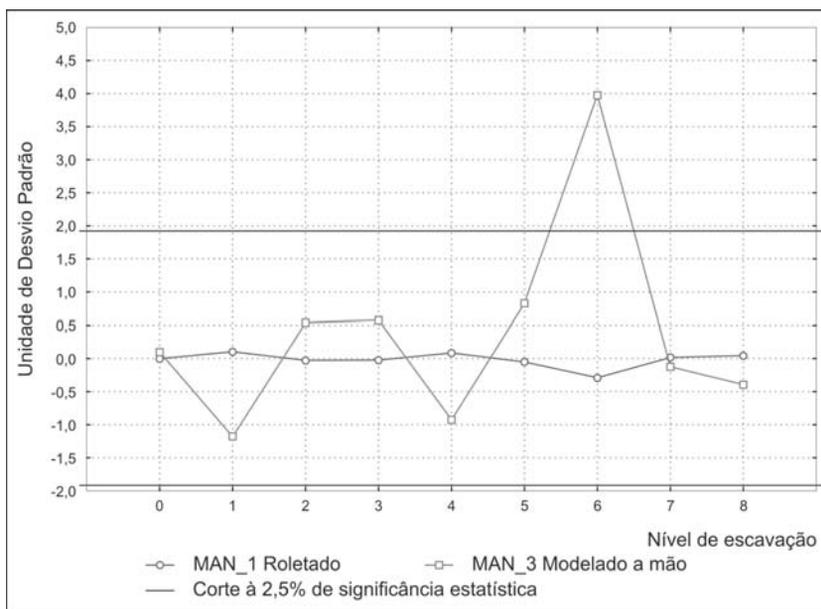


Gráfico 6. Gráfico de linha dos resíduos padronizados do atributo técnica de manufatura - tabela comum.

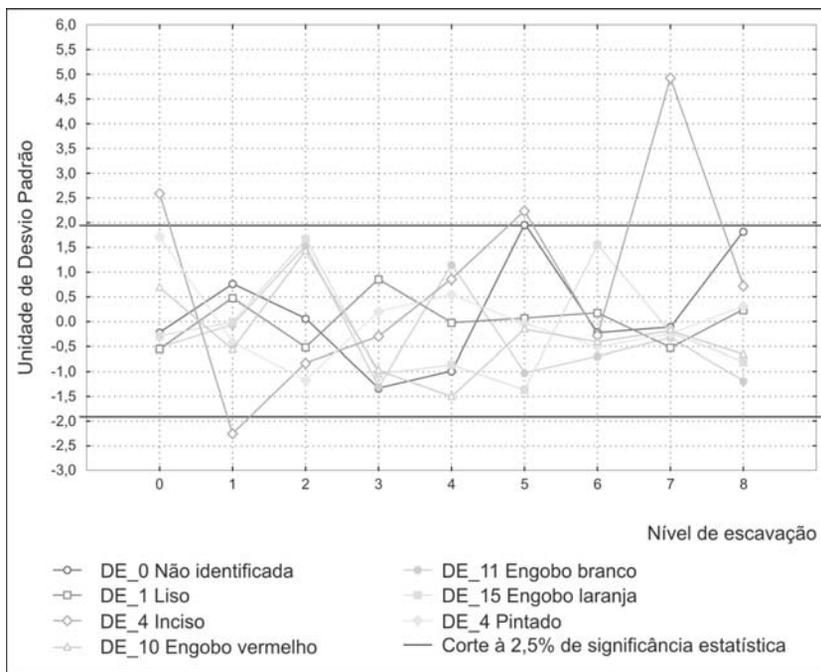


Gráfico 7. Gráfico de linha dos resíduos padronizados do atributo decoração externa.

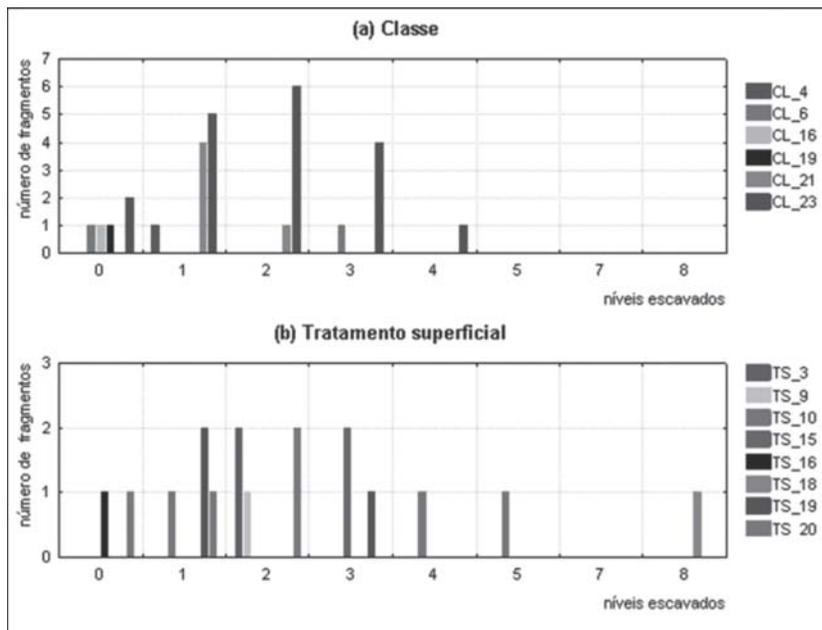


Gráfico 8. Gráfico de barras dos atributos - categorias - (a) classe (b) tratamento superficial - tabela raríssimos.

- de base, parede, borda (CL_4), ombro (CL_16) e parede com furo de suspensão (CL_19) têm apenas uma ocorrência cada registrada nos níveis superiores 0 e 1;

- de parede angular (CL_5) e parede e base (CL_6) possuem dois fragmentos registrados;

- de borda com suporte para tampa (CL_21) ocorrem em dois níveis contíguos com um total de cinco fragmentos.

Quanto ao tratamento superficial, os fragmentos:

- da categoria 20 (TS_20) – sem descrição apresentam contigüidade nos níveis 0, 1 e 2, sendo o tipo com o maior número de ocorrências;

- com lustro interno/alisamento externo (TS_10) e categoria 19 (TS_19) – sem descrição ocorrem em níveis esparsos;

- da categoria 18 (TS_18) – sem descrição possui uma ocorrência no nível 8;

- com alisamento interno sem alisamento externo (TS_3) e com lustro externo/alisamento interno (TS_9) ocorrem apenas no nível 2;

- com lustro interno sem tratamento externo (TS_15) possui duas ocorrências no nível 3;

- com polimento externo sem tratamento interno (TS_16) possui uma ocorrência no nível 0.

No Gráfico de barra 9 (a) e 9 (b) está ilustrada a distribuição dos fragmentos que possuem ao menos um atributo_categoria considerado raríssimo quanto à decoração interna e decoração externa respectivamente. Onde se observa que:

- quanto à decoração interna apenas a de engobo vermelho/pintado (DI_32) apresenta mais de uma ocorrência em níveis esparsos;

- quanto à decoração externa é grande a diversidade de categorias que ocorrem uma única vez, sendo que a decoração externa entalhada (DE_2) e ungulada (DE_3) apresentam o maior número de ocorrências e contigüidade nos níveis superiores. A decoração externa de engobo branco associado ao inciso (DE_25) apresenta contigüidade nos níveis inferiores.

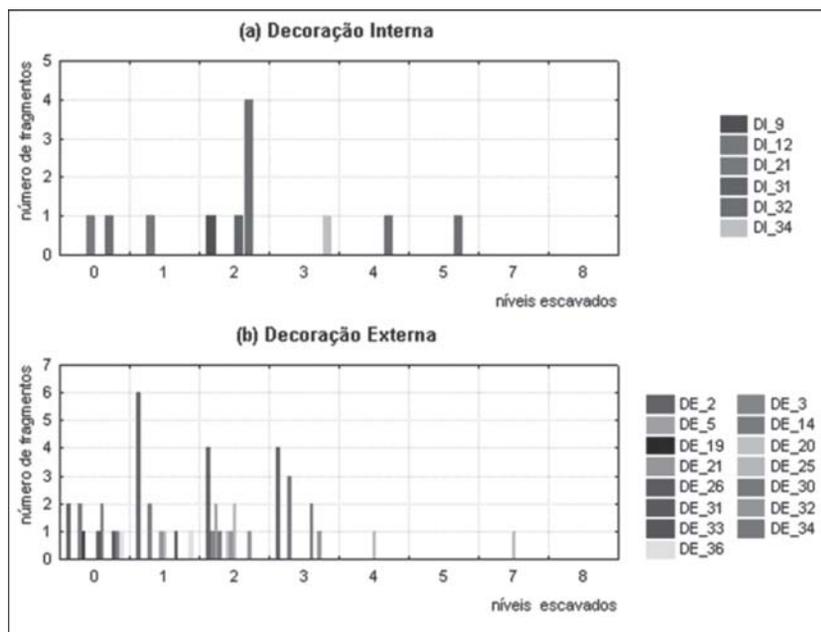


Gráfico 9. Gráfico de barras dos atributos - categorias - (a) decoração interna e (b) decoração externa - tabela raríssimos.

6. Conclusões

O estudo das associações globais e dos perfis dos atributos forneceu amplo arsenal de pesquisa, permitindo aos arqueólogos levantar inúmeras questões a partir da análise desenvolvida.

Podemos citar a título de exemplo:

- o comportamento do atributo caco moído, apresentando concentração de fragmentos com granulometria mais grossa no nível 5 decrescente em direção aos níveis mais superficiais em contraposição aos fragmentos com caco moído de granulometria mais fina apresentando baixa concentração no nível 5 crescente em direção aos níveis mais superficiais, pode ser encarado como uma evolução na fabricação de objetos cerâmicos ou até mesmo no ferramental utilizado na fabricação cerâmica?

- da mesma forma, observando o comportamento da granulometria do material mineral onde existe concentração de material mineral de granulometria média no

nível 3 decrescente em direção aos níveis mais superficiais em contraposição ao material mineral de granulometria mais fina, pergunta-se, esse comportamento pode ser explicado pelo esgotamento e substituição do depósito de material de origem ou talvez pela manifestação de preferência da oleira?

- há justificativa para a concentração de fragmentos com decoração externa incisa nos níveis inferiores?

- a concentração de fragmentos com tratamento superficial-polimento em uma ou ambas as faces nos níveis 6 7 e 8 tem algum significado?

- o que justificaria a diversidade e a concentração de fragmentos com decoração externa considerada raríssima no nível 2?

Raciocínios similares com hipóteses a serem testadas podem ser desenvolvidos tendo por base os resultados encontrados e o conhecimento *a priori* dos arqueólogos.

Portanto, a metodologia eleita para esse estudo, descoberta de conhecimento indireto, bem como a análise dos perfis dos atributos em suas diversas categorias traçados em gráficos de

linha – unidades de desvio padrão X níveis escavados – mostraram-se eficazes em fornecer uma visão temporal do fundo de habitação

NSA2 do sítio arqueológico Piracanjuba, contribuindo dessa forma para o conhecimento das populações pretéritas que habitaram o local.

FRANCO, C.; IMAI, N.N.; TACHIBANA, V.; BARROCÁ, N.F. The history told in ceramic fragments: a temporal vision of the anthropogenic soil nucleus 2 (ASN2) – Piracanjuba archaeological site, Piraju, SP, Brazil. *Revista do Museu de Arqueologia e Etnologia*, São Paulo, 18: 69-88, 2008.

Abstract: This paper aims, by applying indirect knowledge discovery methodology to attributes from ceramic fragments collected from the old bottom Anthropogenic Soil Nucleus 2 (ASN2) at the archaeological site in Piracanjuba – Piraju, SP, Brazil, at providing to archaeologists a temporal vision of the attribute behaviors of ceramic fragments, in order to assist them in the acknowledgement of the past people that inhabited there.

Keywords: Knowledge Discovery – Data mining – Multivariate statistics – Archaeology.

Referências bibliográficas

- BARTHOLOMEW, D.J.; STEELE, F.; MOUSTAKI, I.; GALBRAITH, J.I.
2002 *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Florida: Chapman & Hall.
- BAXTER, M.J.; BEARDAH, C.C.; COOL, H.E. M.; JACKSON, C.M.
2003 Compositional data analysis in archaeometry. Presented at CoDAWork 2003, Girona, Spain . Disponível em <http://ima.udg.es/Activitats/CoDaWork03/paper_baxter_Beardah1.pdf>. Acesso em junho de 2004.
- BERRY, M.; LINOFF, G.
1997 *Data Mining Techniques: For marketing, sales and customer support*. Wiley Computer Publishing, USA.
- BIGOLIN, N.M.
2000 *Data Mining: Conceitos e Técnicas*, Escola Regional de Informática (ERI 2000), Foz de Iguaçu, Tubarão, Maio 2000: 233-250.
- FACCIO, N.B.
1998 *Arqueologia dos cenários das ocupações horticultoras da Capivara, Baixo Paranaíba*.
- Tese de Doutorado em Arqueologia - FFLCH, Universidade de São Paulo, São Paulo.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.
1996 From data mining to knowledge discovery: an overview. In: Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Eds.) *Advances in knowledge discovery*. Cambridge, MIT Press: 1- 36.
- GROEBEL, M.; GRUENWALD, L.
1999 A survey of data mining and knowledge discovery software tools. *SIGKDD Explorations*, 1: 20- 33.
- HABERMAN, S.J.
1978 *Analysis of Qualitative Data: Introductory Topics*. New York: Academic Press.
- IBM
1996 *IBM's Data Mining Technology*. White Paper.
- IPHAN
2004 Instituto do Patrimônio Histórico e Artístico Nacional. Disponível em <<http://iphan.gov.br>>. Acesso em agosto de 2004.

- KOPERSKI, K.; HAN, J.; ADHIKARY, J.
1998 *Mining Knowledge in Geographical Data*, accepted by IEEE Computer. Disponível em <[ftp://ftp.faz.sfu.ca/pub/cs/han/pdf/geo_survey98.pdf](http://ftp.faz.sfu.ca/pub/cs/han/pdf/geo_survey98.pdf)>. Acesso março 2003.
- LOPES, P.R.; BARROCA, N.F.
2001 *Análise da Cerâmica Guarani do Sítio Arqueológico Lagoa Seca II*. Disponível em <http://www.cibergeo.org/agbnacional/VICBG-2004/Eixo5/e5%20219.htm#_ednref1>. Acesso setembro 2004.
- MENDONÇA NETO, M.G., ALMEIDA, M.
2001 O Uso de Interfaces Abundantes em Informação para Exploração Visual de Dados. In: *Brazilian Workshop on Human Factors in Computer Systems (IHC 2001)*. Florianópolis, v. 1: 256-268.
- MOBERG, C.
1986 *Introdução à Arqueologia*. Lisboa: Edições 70.
- MORAIS, J.L.; BARROCA, N.F.
2004 O trabalho do Barro no Sítio Arqueológico Piracanjuba- Piraju, São Paulo. Relatório de Pesquisa do ProjPar. São Paulo, Universidade de São Paulo.
- MUNITA, C.S.; PAIVA, R.P.; ALVES, M.A.; OLIVEIRA, P.M.S.; MOMOSE, E.F.
2001 Characterization of Brazilian prehistoric ceramics by neutron activation analysis. Progress Report 2000-2001 Research Reactor Center. Disponível em <<http://www.ipen.br/biblioteca/progress/pdfs/CRPQ/>> Acesso em julho 2004.
- MUNITA, C.S.; NASCIMENTO, A.; SCHREIBER, S.B.; LUNA, S.; OLIVEIRA, P.M.S.
2004 Chemical study of some ceramics from Brazilian Northeast. *Journal of Radioanalytical and Nuclear Chemistry*, 259 (2). Kluwer Academic Publishers: 305-309(5). Disponível em <<http://www.ingenta.com/isis/browsing/TOC/ingenta?issue=pubinfobike://klu/jrnc/2004/00000259/00000002>>. Acesso julho 2004.
- PEREIRA, J.C.R.
2004 *Análise de Dados Qualitativos: Estratégias Metodológicas para as Ciências da Saúde, Humanas e Sociais*. 3a. ed., São Paulo: EDUSP.
- PIATETSKY-SHAPIRO, G.
1991 Knowledge discovery in real databases: A report on the IJCAL- 89 Workshop. *AI Magazine*, 11 (5), Jan., Special issue: 68- 70.
- ROBRAHN GONZÁLEZ, E. M.
1996 *A ocupação ceramista pré-colonial do Brasil Central*. Tese de Doutorado em Arqueologia - FFLCH, Universidade de São Paulo, São Paulo.
- SANTOS, A.J.F.; ALMEIDA, M.O.; ALMEIDA, R.G.M.; MENDONÇA NETO, M.G.
2000 Uso de interfaces abundantes em informação para mineração visual de dados. Relatório técnico RT-NUPERC-2000-4/p. (Relatório técnico do estado da arte), Universidade de Salvador, Salvador.
- VON WANGENHEIM, A.
2004 *Técnicas Estatísticas para Reconhecimento de Padrões*. Disponível em <<http://www.inf.ufsc.br/%7Eawangenh/RP/estatisticas.html>> Acesso setembro 2004.

Recebido para publicação em 25 de maio de 2008.