

Aplicação de Análise de Cluster em dados integrados de um estudo prospectivo: projeto epidoso como cenário

Application of Cluster Analysis in integrated data of a prospective study: project epidoso as a scenario

Aplicación del Análisis del Grupo en datos integrados de un estudio prospectivo: proyecto epidoso como escenario

Frederico Molina Cohrs¹, Fernando Sequeira Sousa¹, Josceli Maria Tenório¹, Luiz Roberto Ramos², Ivan Torres Pisa³

RESUMO

Descritores: Análise por conglomerado; Epidemiologia; Envelhecimento; Internet; Sistemas de gerenciamento de base de dados

Objetivo: Este estudo apresenta o reforço nas comunicações entre grupos de pesquisadores com o benefício de portais web, assim como buscou um olhar diferente na análise dos dados de um estudo de coorte possibilitando identificar grupos de indivíduos com características socioeconômicas e agravos ou vulnerabilidades semelhantes para um melhor desenho de promoção de saúde. **Método:** Estudo exploratório, de abordagem qualitativa e quantitativa, analítico, do tipo estudo de caso. **Resultados:** Os dados foram normalizados e transferidos para uma base de dados relacional. O portal web permitiu acesso mais seguro aos dados e com alta disponibilidade. Das 207 variáveis iniciais, apenas 56 foram selecionadas para análise de cluster, na qual foram identificados três grupos de indivíduos. **Conclusões:** Pesquisas longitudinais com vários grupos de pesquisa necessitam usar bases relacionais com alta disponibilidade e segurança. O acesso aos dados pode ser feito por meio da Internet com acesso controlado. A identificação de grupos permite uma promoção e comunicação em saúde com melhores resultados.

ABSTRACT

Keywords: Cluster analysis; Epidemiology; Aging; Internet; Database management systems

Objective: This study presents the strengthening of communications among groups of researchers with the benefit of web portals, and sought a different look in the analysis of data from a cohort study enabling to identify groups of individuals with similar socioeconomic, health problems or vulnerabilities to a better planning of health promotion. **Method:** This was an exploratory study of qualitative and quantitative approach, analytic, case study. **Results:** The data were normalized and transferred to a relational database. The web portal allowed more secure access to data, thus researchers now have safe access and high availability of data. In cluster analysis, it was identified three groups of subjects. **Conclusions:** Longitudinal research with some or several research groups need to use relational databases with high availability and security of data. Access to data can be done through the Internet with controlled access. Cluster analysis, especially the method TwoStep, identifies groups of individuals with socioeconomic and health problems or similar vulnerabilities without having to report the number of groups. This allows creation of groups in health promotion and communication with better results.

RESUMEN

Descriptores: Análisis por conglomerados; La epidemiología; El envejecimiento; Internet; Gestión de base de datos de los sistemas

Objetivo: Este estudio presenta el fortalecimiento de las comunicaciones entre grupos de investigadores con el beneficio de portales web, y buscó una mirada diferente en el análisis de los datos de un estudio de cohortes con la posibilidad de identificar grupos de individuos con socioeconómica y los mismos peligros o vulnerabilidades para un mejor diseño de promoción de la salud. **Método:** Un estudio exploratorio de abordaje cualitativo y cuantitativo, el estudio de caso analítico. **Resultados:** Los datos se normalizaron y se transfirió a una base de datos relacional. El portal de Internet ha permitido un acceso más seguro a los datos, ahora tenemos acceso a una alta disponibilidad. De las 207 variables iniciales, sólo 56 participaron en el análisis de conglomerados, que se identificaron tres grupos de personas. **Conclusiones:** La investigación longitudinal con diversos grupos de investigación que utilizar bases de datos relacionales con alta disponibilidad y seguridad. El acceso a datos puede hacerse a través de Internet con acceso controlado. El análisis de clusters, en particular el método TwoStep permite identificar grupos de individuos con riesgos socioeconómicos y similares o vulnerabilidades, sin tener que informar el número de grupos. Esto permite la creación de grupos de promoción de la salud y la comunicación, con mejores resultados.

¹ Doutorando pelo Departamento de Informática em Saúde da Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.

² Professor. Departamento de Medicina Preventiva da Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.

³ Professor. Departamento de Informática em Saúde da Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.

INTRODUÇÃO

Este artigo tem como cenário o Projeto Epidoso⁽¹⁾, um estudo longitudinal iniciado em 1991 e conduzido na UNIFESP como projeto temático com foco em idosos residentes na área da Vila Clementino em São Paulo visando identificar qualidade de vida, perdas funcionais, condições de saúde física sob a óptica da cardiologia, saúde mental e saúde bucal. O projeto original considerou quatro períodos, denominados ondas, de entrevistas com questionários estruturados para coletar informações demográficas e clínicas. A primeira onda iniciou-se em 1991 e contou com informações sobre 1.667 idosos, e foram armazenadas em um programa computacional próprio. Esta foi a onda utilizada para este estudo.

O trabalho aqui relatado considerou a organização dos dados do Projeto Epidoso para facilitar o acesso geográfico aos dados coletados e identificar uma visão conjunta, integrada, considerando o sujeito, o idoso, com suas características socioeconômicas, de agravos, e vulnerabilidades, colaborando ainda com a integração dos pesquisadores de diferentes especialidades e pontos geográficos. Desta forma, os autores optaram pela criação de um portal web que possibilitasse uma gestão integralizada dos dados. Para o gerenciamento desses dados coletados ao longo de 15 anos foi construída uma infoestrutura⁽²⁾ provendo uma estrutura de apoio tecnológico para o projeto; software necessário para acessar, manipular, organizar e analisar as informações; apoio na governança, gestão e uso da informação e treinamento de pessoas e organizações envolvidas na criação da informação. Neste trabalho a questão norteadora para a análise dos dados foi considerar grupos de idosos com características socioeconômicas e agravos ou vulnerabilidades semelhantes, a partir de técnicas de clusterização, para criação de ações educativas e preventivas com maior e melhor foco. Esta técnica também está presente em Sistema de Apoio à Decisão Clínica (SADC), alvo e estudos durante as últimas décadas na área de informática em saúde, são projetados para ajudar profissionais de saúde na tomada de decisão⁽³⁾.

MÉTODOS

Foi realizado um estudo exploratório, de abordagem qualitativa e quantitativa, analítico, do tipo estudo de caso⁽⁴⁾. O Projeto Epidoso foi o cenário escolhido para este estudo de caso na aplicação de uma análise de cluster em estudos epidemiológicos. Os sujeitos estão no conjunto de dados coletados ao longo dos dois primeiros anos, considerado a primeira onda de entrevistas e avaliações clínicas de idosos. Por utilizar dados secundários que abasteceram a base para o portal considerou-se como local de realização a Vila Clementino, São Paulo. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da UNIFESP (CEP #1890/08). Os autores declaram não haver conflito de interesse quanto aos resultados obtidos. Os dados utilizados para a análise não permitem a identificação dos pacientes.

As etapas para criação do Portal Web e a análise de

cluster são descritas nos itens seguintes.

Base de Dados do Estudo

Para a etapa da construção da base de dados, foi feito um confronto entre o arquivo-base e os questionários aplicados quando foi detectada a necessidade de entrevistar os funcionários e pesquisadores participantes do Projeto Epidoso para identificar outros arquivos existentes. Esta ação culminou na localização de 41 arquivos nos formatos SAV, DBF, XLS e MDB com dados redundantes. Um arquivo-base (*epidoso.sav*) contendo todos os dados unificados de todas as ondas, com 929 variáveis e 1.666 registros também foi analisado. Para compreender o conteúdo das variáveis foi construído um dicionário de dados com nome, significado e tipo (categórico ou numérico) de cada variável e onda a qual estava vinculada, para posterior normalização.

Posteriormente, o conteúdo do arquivo base foi transferido para um gerenciador de banco de dados MySQL pela facilidade de manipulação e segurança no acesso. O dicionário de dados também foi inserido no banco de dados. A tabela base foi particionada em quatro outras tabelas, uma para onda. Os dados de mortalidade foram importados para a base. A construção da estrutura de banco de dados com a descrição das variáveis possibilitou definir uma infoestrutura do projeto.

Técnicas de Clusterização

Em epidemiologia a clusterização pode ser utilizada para analisar uma série de eventos bem agrupados ou casos de doença ou fenômeno de saúde relacionados com outros padrões de distribuição bem definidos em relação ao tempo ou lugar, ou ambos. Os métodos de clusterização mais conhecidos podem ser classificados como hierárquicos, de particionamento, baseados em densidade, baseados em grade, baseados em modelos, clusterização de dados ultradimensionais ou clusterização baseada em parâmetros e restrições⁽⁵⁾. O mais comumente utilizado é o método o hierárquico⁽⁶⁾ que aglomera ou particiona uma hierarquia de clusters. Neste trabalho optou-se pela utilização de uma clusterização do tipo particionamento cujos algoritmos disponíveis mais conhecidos são k-means, k-medoides e TwoStep.

O método TwoStep⁽⁷⁾ possibilita uma análise escalonável apropriada para grandes conjuntos de dados que possuam uma mistura de variáveis contínuas e categóricas. Seu funcionamento é baseado em dois passos: no primeiro passo, o algoritmo forma os pré-clusters, que são clusters dos casos originais, substituindo-os com objetivo de se ter um menor número de casos para a próxima etapa, reduzindo o tamanho da matriz que contém distâncias entre todos os casos pareados possíveis. Essa estratégia representa uma abordagem de agrupamento sequencial. O algoritmo faz uma varredura dos registros um a um e decide se o registro atual deve fundir-se com os clusters formados anteriormente ou iniciar um novo cluster com base no critério da distância⁽⁸⁾. Após a finalização da pré-clusterização todos os casos em um mesmo pré-cluster são tratados como uma única entidade com uma técnica baseada em hierarquização. As

suas características são utilizadas como novos casos. O tamanho da matriz de distâncias não é mais dependente do número de casos, mas do número de pré-clusters⁽⁹⁾.

O Quadro 01 apresenta o script utilizado no processo de clusterização com a indicação de todas as variáveis utilizadas e as configurações para o algoritmo TwoStep.

Quadro 1 - Script do SPSS utilizado no processo de clusterização dos dados.

```

/CATEGORICAL VARIABLES=educ1 civil1 xfilht1 xfilhv1
trab1 ocupat1 ocupant1 geracl1 razapo1 whoaju$1 whoaju1
sub$1 sub$rou1 nectrab1 casa1 tipcas1 submor1 subsau1
sauout1 mental1 mmtot1 reu1 ha1 var1 iu1 pv1 ins1 cat1 col1
out1 mov1 reab1 oculo1 difvis1 aud1 audicv1 xdent1 dente1
prot1 mast1 engol1 qd1 xqd1 xsex1 prz1 intl1 rem1 rempro1
satisf1 atfis1 atfistp1 cami1 camifor1 conv1 segur1 publ1
mdpar1 q98sau1 q98vis1 xinter1 whocuid1 satfam1 conflit1
relfam1 xviz1 xpar1 xami1 satxfam1 satxami1 confid1 lazrad1
laze1 lazman1 lazand1 xvaj1 satlaz1 sateco1 satsau1 satsoc1
probl1 valtran1 probtral1 casoul1 xpress1
/CONTINUOUS VARIABLES=agecal1 tpsp1 tpmora1
civilt1 renido$1 tpbene1 idapos1 mmse1 adltot1 ndoenc1
renfam$1
/DISTANCE LIKELIHOOD
/NUMCLUSTERS AUTO 20 BIC
/HANDLENOISE 25
/MEMALLOCATE 512
/CRITERIA INTTHRESHOLD(0) MXBRANCH(8)
MXLEVEL(3)
/VIEWMODEL DISPLAY=YES.

```

Análise dos Dados

Das 207 variáveis da primeira onda, 137 delas foram selecionadas para participar da análise de cluster. A exclusão imediata de 70 variáveis deu-se porque o conteúdo delas era extremamente sensível, com variação muito rápida (em poucas horas como glicemia e pressão venosa central por exemplo). Após essa exclusão foi executada a análise de cluster e observado o nível de clusterização.

O método TwoStep pontua a qualidade da clusterização com valores entre 0 e 1: valor adimensional. Quanto mais próximo a 1, melhor é a clusterização, com menos casos sendo deixados à margem dos clusters identificados. Caso a pontuação da clusterização resultasse inferior a 0,1, considerado ruim, era identificada a variável com menor significância no processo de clusterização cujos dados eram iguais em todos os clusters. Essa variável era então excluída, e a clusterização era novamente executada. Este processo iterativo foi repetido até que a pontuação da clusterização tivesse seu valor acima de 0,1. Com isso, das 137 variáveis iniciais, apenas 56 foram usadas na análise de cluster final.

RESULTADOS

Os dados do Projeto Epidoso foram importados dos arquivos-base para um banco de dados com tabelas normalizadas segundo as formas normais de CODD⁽¹⁰⁾. O Portal Epidoso foi construído em ambiente web, com acesso e gerenciamento dos dados possíveis exclusivamente para usuários autorizados pelo gestor do projeto, com controle de acesso de acordo com o perfil de usuário.

Tabela 1 - Clusters identificados com 77,7% de inclusão dos sujeitos.

Cluster #	Número de pacientes	%
1	349	27,0
2	424	32,8
3	521	40,3
Total	1.294	77,7

Após a execução da rotina de análise de clusters foram identificados 3 (três) grupos de idosos com características socioeconômicas, de agravos e vulnerabilidades semelhantes, não balanceados em termos de quantidade

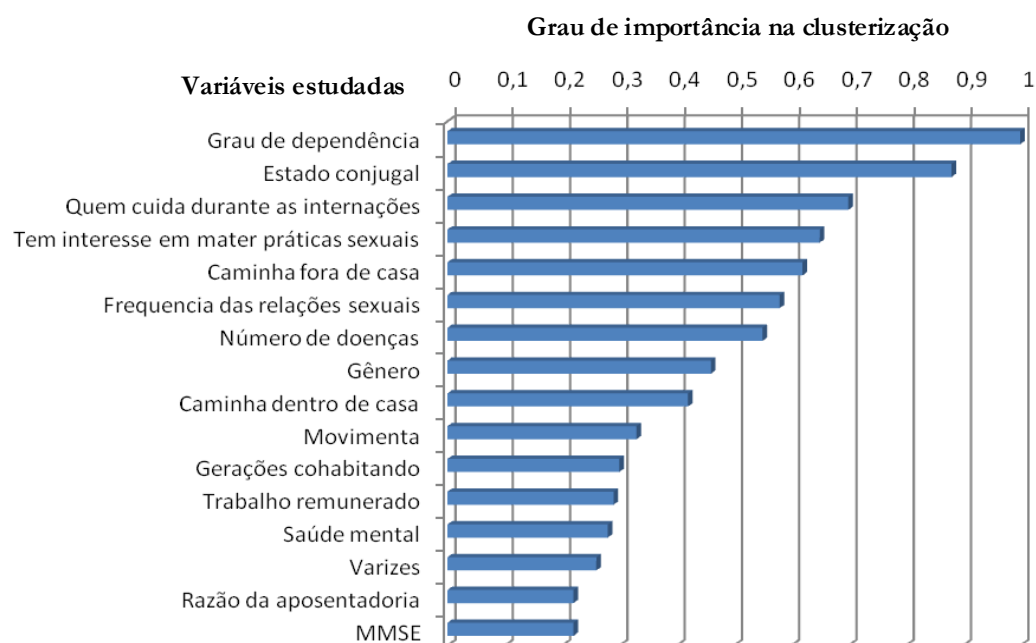


Figura 1 - Variáveis de maior importância na participação da formação geral dos clusters.

de indivíduos, o que é próprio do método. A cobertura máxima alcançada foi de 77% dos 1.667 participantes, conforme indicado na Tabela 1, com identificação de três clusters usando 56 variáveis.

A Figura 1 apresenta as 16 variáveis mais importantes utilizadas no processo de identificação dos clusters, ordenadas por importância na clusterização. Cada cluster, no entanto, apresenta importância diferente para cada variável com variação entre 0 e 1.

A variável “Grau de dependência” foi a mais importante no processo de clusterização. Ela indica o quanto o indivíduo consegue fazer atividades diárias sem o auxílio de outra pessoa.

Fixando a observação no Grau de dependência, foi possível identificar no terceiro grupo (terceiro cluster), A menor idade média e o menor número de doenças (comorbidades). Essas informações apontam uma possível diferença no cuidar dos indivíduos contidos no

cluster 3 e um diferente planejamento de promoção da saúde dos idosos entrevistados.

Para o cluster 1 apenas a dificuldade de movimentar pernas e braços mostra significância. Para o cluster 2, foi significativo o fato de os indivíduos relatarem não ter nenhuma dificuldade de movimentação de braços e pernas, e não serem casos de saúde mental. O cluster 3 foi o que apresentou mais variáveis com significância a começar pela movimentação. Neste cluster, os indivíduos relatam não terem nenhuma dificuldade na movimentação de braços e pernas. Os indivíduos deste cluster cohabitam com uma geração, com remuneração vinda da própria aposentadoria, não são casos de saúde mental, e não apresentam varizes.

Como várias das variáveis utilizadas na análise são de natureza categórica, estas apresentam grande dificuldade de serem utilizadas em outros métodos que não o escolhido para este trabalho. Em outros métodos, ou essas variáveis seriam desprezadas ou teriam seu tipo alterado para numérico, possibilitando uma visão equivocada dos dados armazenados.

O método TwoStep aplicado possibilita medir, em cada cluster, uma importância interna, relativa, das variáveis considerando-se apenas o cluster em questão. A Figura 2 mostra as variáveis na ordem de importância de cada cluster e a barra interna exhibe a importância da variável na identificação do cluster.

A variável, aqui caracterizada por uma característica, mais importante para o cluster 1 é “caminhar fora de casa”, seguido por “caminhar em casa” e o “grau de dependência”, em ordem. Para este cluster, a limitação de caminhar fora de casa ou caminhar dentro de casa é muito importante ao considerar a grande limitação de movimentos que essas pessoas têm, conduzindo a privações do convívio social, independência para a realização de suas atividades pessoais até mesmo básicas como a alimentação e higiene pessoal, chegando até a diminuição do lazer.

Para o cluster 2 a variável mais importante foi “estado conjugal”, seguida por “quem cuida durante as internações” e “tem interesse em manter relações sexuais”. As três variáveis de maior importância para este cluster são também categóricas e conduzem à condição mais isolada deste segmento, constituído por viúvas que moram sozinhas. Pessoas, supostamente, mais independentes fisicamente.

O cluster 3 foi formado por pessoas que moram com suas famílias, considerando descendentes em primeiro nível (filhos) e segundo nível (netos). A co-habitação possibilita que outras pessoas contribuam para que o idoso tenha uma maior aderência aos tratamentos.

A Figura 3 mostra a integração dos três clusters, juntamente com a importância da participação das variáveis no processo de clusterização. Foi possível identificar como uma variável é importante para um grupo de pessoas, mas pouco importante para outros grupos. Os picos em diferentes pontos mostram que variáveis importantes na clusterização não foram importantes para a visão, para a compreensão de todos os clusters. Como exemplo tem-se a variável “grau de dependência”, que foi a variável



Figura 2 - Distribuição das variáveis mais importantes em cada cluster identificado.

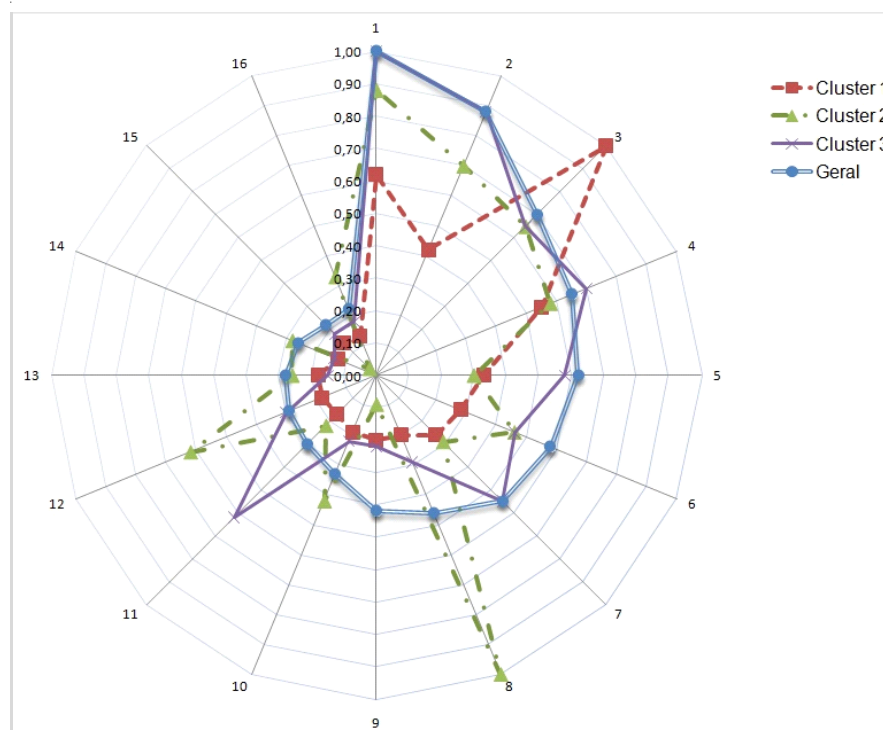


Figura 3 - Gráfico em radar exibindo a importância de cada variável com sobreposição dos três clusters e avaliação geral.

mais importante na definição dos clusters. Para o cluster 1, esta variável está em terceiro lugar de importância, enquanto que para o cluster 2, aparece em oitavo lugar. Já no cluster 3 esta variável está em primeiro lugar de importância na definição do cluster. Essas variações possibilitam compreender a variabilidade dos grupos.

Apesar de o método TwoStep indicar um valor de grau de importância para a identificação dos clusters, esses valores devem ser compreendidos apenas como identificação no processo de clusterização.

DISCUSSÃO

O Portal Web Epidoso contemplou funcionalidades para manipulação (seleção e filtragem) de dados e exportação dos dados para uso de softwares estatísticos. O nível segurança ao acesso aos dados foi considerado adequado. A infoestrutura criada é de fácil manutenção e há a possibilidade de realização de novas implementações em projetos futuros.

O uso da técnica de clusterização dos dados da primeira onda do Projeto Epidoso resultou em três clusters com variáveis apresentando diferença na participação da criação dos grupos.

Dos indivíduos presentes no cluster 1, 57% caminhavam fora de casa sem dificuldade ou com pouca dificuldade, 88% caminhavam em casa sem ou com pouca dificuldade; o grupo possui grau médio de dependência ($7,1 \pm 3,7$), detectado como o maior entre os clusters; alto número médio de doenças ($5,4 \pm 1,9$); 51% relataram dificuldade de movimentação de braços ou pernas; 45% apresentam casos de saúde mental; 42% relataram que possuem varizes e isso interfere nas atividades diárias; 19% referiam ter incontinência urinária com interferência no dia-a-dia; 59% referem que os

problemas na compra de remédios estão relacionados com o custo; 51% apresentavam problemas na coluna; 29% relataram ter prisão de ventre; 40% relataram ter reumatismo; 72% não realizavam caminhadas; 26% relataram ter hipertensão arterial; 36% relataram ter insônia, e todas as doenças relatadas acima interferiam nas atividades diárias; 54% apresentavam dificuldade de conversar.

Uma visão facilitada do cluster 1 é vê-lo como um conjunto de mulheres que caminhavam dentro e fora de casa, com alguma dificuldade, com razoável dependência para realização de atividades diárias e um considerável número de comorbidades, cujas doenças relatadas interferiam nas atividades diárias. Em poucas palavras: mulheres com alguma dependência física, algum comprometimento cognitivo, com doenças interferindo nas atividades diárias.

Já os indivíduos presentes no cluster 2 eram predominantemente viúvos (74,8%), cuidadas predominantemente pela filha (58,3%) ou parentes (24,8%) quando estavam internados; sem interesse em manter contato sexual (97,4%); sem relato de práticas sexuais (99,5%); a minoria (30,7%) morando só; do sexo feminino (91%); a maioria (72,7%) vivendo com a aposentadoria ou pensão, sem um trabalho remunerado; um baixo grau (2,9) de dependência; 64,2% não realizavam atividades manuais; caminhavam em casa e sem dificuldade (99,5%); 56,4% não se aposentaram; caminhavam fora de casa e sem qualquer dificuldade (85,8%); 55,9% não possuíam varizes; 89,4% não apresentaram casos de saúde mental; 80,7% possuíam filhos, e 80,4% não relataram nenhum problema de movimentação.

De forma sintética, o cluster 2 é composto por mulheres, viúvas, cuidadas predominantemente pela filha ou parentes quando estão internados, sem interesse em

manter contato sexual, sem relato de práticas sexuais, algumas morando só, vivendo com a aposentadoria ou pensão, sem nenhum trabalho remunerado, independentes na realização de atividades diárias, não realizam atividades manuais, caminham em casa e sem dificuldade, boa parte não se aposentou, caminham fora de casa e sem qualquer dificuldade, não possuem varizes, não são caso de saúde mental, possuem filhos, não relatam nenhum problema de movimentação. Ou seja, mulheres, viúvas, fisicamente independentes, convivendo com poucos familiares, sem realizarem atividades manuais, vivendo de aposentadoria ou pensão, caminhando, e são mentalmente capazes.

Os indivíduos reunidos no cluster 3 possuíam baixo grau de dependência (2,1 d.p. 1,5); 94% eram casados; 70% relataram ter interesse em manter contato sexual; 78% eram cuidados pelo cônjuge durante as internações; 62% relataram ter relação sexual pelo menos uma vez por mês; 66% são do sexo masculino; possuem uma média de 2,5 (d.p. 1,7) doenças; 55% vivem só com os recursos da aposentadoria; atingiram uma pontuação média de 26,4 (d.p. 3,1) no exame de estado mental; 54% aposentaram por tempo de serviço; 99% caminham fora de casa sem dificuldade ou com alguma dificuldade; 99% conviviam com pelo menos mais uma geração; apresentam idade média de 72,6 (d.p. 5,4) anos; 81% possuíam casa própria; 43% tinham o ensino médio ou superior; 71% relataram realizar caminhadas. Em síntese, são indivíduos fisicamente independentes, casados, com interesse em manter prática sexual, sendo cuidados pelo cônjuge quando internados, predominando o sexo masculino, com baixo número de comorbidades, vivendo apenas com o dinheiro da aposentadoria, apresentando uma boa capacidade cognitiva, caminhando bem fora de casa, convivendo com gerações em casa própria, relatando ainda terem cursado o ensino médio ou já o superior.

Este cluster é formado, assim, por homens fisicamente independentes, com boa capacidade cognitiva, aposentados, convivendo com gerações em casa própria.

Considerando que uma comunicação em saúde eficaz busca primordialmente identificar o público-alvo⁽¹¹⁾, o uso da clusterização para identificação segmentos permite o planejamento otimizado da comunicação com direcionamento para cada segmento, ou selecionar um dos segmentos para uma comunicação ainda mais direta.

CONCLUSÃO

Bancos de dados bem concebidos e projetados assim como portais web devem fazer parte dos projetos de grupos de pesquisa com direcionamento translacional que estão se tornando realidade nas universidades. Essa tecnologia possibilita acesso aos dados com autenticação para controle e rastreamento de ações, oferecendo maior segurança e também acesso distribuído aos dados.

No cenário do Projeto Epidoso, uma análise de cluster mostrou-se interessante para identificar grupo de pessoas com características socioeconômicas e agravos ou vulnerabilidades comuns, o que conduz a uma concepção mais eficaz de promoção da saúde. Cada cluster identificado possui características próprias e é composto por grupos distintos de indivíduos, que trazem a possibilidade de uma nova visão para os dados de um estudo longitudinal. Mais do que uma visão epidemiológica, uma visão estatística com criação de segmentos pode ser construída, mostrando que há várias facetas para se compreender problemas, e também há várias formas para a resolução de um mesmo problema. O método TwoStep apresentou-se como uma opção aceitável para a identificação de grupos com características semelhantes, uma vez que não é necessário informar a quantidade ótima de grupos.

REFERÊNCIAS

- Ramos LR, Rosa TEC, Oliveira ZM, Medina MCG, Santos FRG. Perfil do idoso em área metropolitana na região sudeste do Brasil: resultados de inquérito domiciliar. *Rev. Saúde Pública*. 1993;27(2):87-94.
- Office of Health CA. The Canadian Health Info-Structure: a conceptual overview: background paper. In: National Conference on Health Info-Structure; 1998 fev 8 -10; Toronto, CA. Available from: http://www.hc-sc.gc.ca/hcs-sss/alt_formats/pacrb-dgapcr/pdf/pubs/ehealth-esante/1997-chiics-overview-survol/1997-chiics-overview-survol-eng.pdf
- Tenório JM, Hummel AD, Sdepanian VL, Pisa IT, Marin HF. Experiências internacionais da aplicação de sistemas de apoio à decisão clínica em gastroenterologia. *J. Health Inform*. 2011;3(1): 27-31.
- Sampieri RH, Collado CF, Lucio PB. Definição da pesquisa a ser realizada: exploratória, descritiva, correlacional ou explicativa. 3a. ed. São Paulo: McGraw-Hill; 2006.
- Han J, Kamber M. Cluster analysis. In: *Data mining: concepts and techniques*. 2a. ed. São Francisco: Morgan Kaufmann; 2006.
- Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: an empirical comparison. *Data Knowl Eng*. 2007;63(1):155-66.
- Theodoridis S, Koutroumbas K. *Pattern recognition*. 4a. ed. USA: Elsevier; 2009.
- SPSS I. 1122644952_The SPSS TwoStep Cluster Component.pdf (objeto application/pdf): SPSS Inc.; 2001. Available from: http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf
- Bacher J, Brand R, Bender S. SPSS TwoStep Cluster – A first evaluation. [cited 2010 dec 17]. Available from: http://www.opus.ub.uni-erlangen.de/opus/volltexte/2004/81/pdf/a_04-02.pdf
- Codd EF. A relational model of data for large shared data banks. *Commun ACM*. 1970;13(6):377-87.
- Kotler P, Shalowitz J, Stevens RJ. Planejamento e gerenciamento integrado de comunicação de marketing. In: *Marketing estratégico para a área da saúde*. São Paulo: Bookman; 2010. p. 395-448.