Análise de dados do DataViva utilizando técnicas de projeção multidimensional

Alternative Title: DataViva data analysis using multidimensional projection techniques

Charles Mendes Lima
Universidade Federal de Uberlândia
Uberlândia-MG, Brazil
charlesmendeslima@hotmail.com

Jose Gustavo S. Paiva
Universidade Federal de Uberlândia
Uberlândia-MG, Brazil
gustavo@ufu.br

RESUMO

A visualização de informação cria representações gráficas de coleções de dados para comunicar melhor seu conteúdo informacional ao usuário, revelando tendências e padrões que propiciam uma melhor tomada de decisão. O DataViva é uma plataforma computacional que disponibiliza um conjunto de ferramentas de análise visual aplicados a dados socioeconômicos, educacionais e de comércio internacional de localidades brasileiras, para direcionar a criação de políticas públicas que contribuam para o desenvolvimento nessas localidades. Técnicas de visualização baseadas em projeção multidimensional apresentam potencial para ressaltar a estrutura global da coleção de dados, bem como a seleção/exploração de grupos de interesse. A coleção é organizada baseada na similaridade entre as instâncias, realçando os relacionamentos entre elas. Nenhuma das ferramentas de análise fornecidas pelo DataViva atualmente oferece essa perspectiva dos dados de forma explícita. Nesse sentido, este artigo apresenta uma aplicação de duas técnicas de projeção multidimensional aos dados oferecidos por essa plataforma. Os resultados apresentados demonstram o potencial dessa classe de técnicas em revelar grupos de localidades com perfis semelhantes, além de destacar o perfil de localidades com comportamento peculiar, representando uma ferramenta com potencial para permitir a compreensão do comportamento das localidades brasileiras, bem como o relacionamento entre elas.

Palavras-Chave

DataViva; Visualização de Informação; Projeção Multidimensional.

ABSTRACT

Information Visualization creates graphical representations for data collections to better communicate its informational content to the user, revealing trends and patterns that allow a better decision making. DataViva is a computational platform that provides a set of visual analysis tools applied on socioeconomic, educational and international trade information from Brazilian localities, to direct the creation of public policies that contribute with the development in these localities. Multidimensional projection techniques present potential to highlight the collection global structure, as well as the selection/exploration of interest groups. The collection is organized based on instances similarity, emphasizing the relationships among them. None of the analysis tools currently provided by DataViva offers this data perspective explicitely. In this sense, this paper presents an application of two multidimensional projection techniques to the data provided by this platform. The results show the potential of this class of techniques in revealing groups of localities with similar profiles, and highlighting the profile of localities with peculiar behavior, representing a potential tool for comprehend the behavior of Brazilian localities, as well as the relationship among them.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•**Human-centered computing** → **Visual analytics;**

Keywords

DataViva; Information Visualization; Multidimensional Projection.

1. INTRODUÇÃO

Atualmente, diversas entidades governamentais coletam um grande volume de informações socioeconômicas que representam importantes indicativos a respeito do crescimento social e econômico de uma comunidade. Desse fato, surge a preocupação de como processar e analisar esses dados, uma tarefa crucial para os especialistas, que necessitam extrair o máximo de informações para direcionar a criação de políticas públicas ou estratégias de negócio mais eficazes.

Técnicas de visualização de informação podem representar uma importante ferramenta de análise. A representação visual de coleções de dados comunica claramente ao usuário o conteúdo informacional desses dados, reduzindo o trabalho cognitivo necessário para realizar diversas tarefas [9]. Seguindo essa ideia, o principal papel dessas técnicas é criar representações visuais das relações contidas nos dados e associá-las a ferramentas interativas que permitam ao

usuário participar ativamente do processo de análise, utilizando *layouts* que podem ser úteis para fins educacionais, científicos e governamentais. A análise visual dos dados pode permitir ao usuário encontrar tendências e padrões que seriam difíceis de serem detectados pela análise instância a instância, ou utilizando algum tipo de formato tabular. Dentre as técnicas de visualização existentes na literatura, aquelas baseadas em projeção multidimensional apresentam bons resultados em diversos cenários [14, 6, 11]. A ideia principal é preservar no espaço de visualização (usualmente 2 ou 3 dimensões) os relacionamentos relevantes observados no espaço original (representado pelos indicadores medidos para cada instância), utilizando medidas de similaridade que posicionem próximas instâncias similares, e distantes instâncias não similares.

O **DataViva**[1] é uma plataforma aberta para pesquisa e análise de dados socioeconômicos e educacionais de diversas localidades brasileiras, utilizando um conjunto de técnicas de visualização. Ela representa uma importante ferramenta de análise e gestão estratégica, com capacidade de revelar tendências a respeito do andamento da sociedade, economia e educação de forma prática e eficiente.

Este artigo apresenta a aplicação de duas técnicas de projeção multidimensional aos dados fornecidos pelo DataViva, e analisa os *layouts* produzidos, de forma a destacar que padrões importantes foram identificados pela organização obtida. Para a análise, foi desenvolvido um sistema computacional que oferece, além dos *layouts* gerados pelas técnicas, um conjunto de ferramentas interativas para a exploração desses *layouts*. Acredita-se que a aplicação dessas técnicas destaque aspectos estruturais dos repositórios, em termos de relacionamentos entre as instâncias. Pela análise do posicionamento dos pontos no *layout*, pode ser possível, dentre outras tarefas, identificar grupos de localidades com perfil comportamental similar, ou localidades com comportamento peculiar, guiando a definição de políticas públicas semelhantes, ou a reutilização daquelas previamente empregadas com sucesso em uma ou mais localidades. Nenhuma das ferramentas de análise fornecidas pela plataforma atualmente oferece essa perspectiva dos dados de forma explícita, de forma que os *layouts* gerados pelas técnicas utilizadas aqui podem potencializar a análise dos dados e melhorar a tomada de decisão.

A principal contribuição deste trabalho é um sistema de análise visual de dados socioeconômicos que expande as capacidades exploratórias do sistema DataViva, provendo uma ferramenta capaz de analisar as localidades brasileiras sob a perspectiva dos relacionamentos entre elas, que por sua vez refletem seu perfil comportamental. É possível assim realizar uma análise do mercado de trabalho formal, do comércio exterior e educação em diversas localidades brasileiras, que possibilita uma melhor compreensão da situação no país, sob essa perspectiva.

As seções seguintes do texto detalham a plataforma DataViva, o sistema de análise proposto, além dos resultados obtidos com essa análise.

2. A PLATAFORMA DATAVIVA

A plataforma DataViva tem o objetivo de permitir a realização de análises visuais dos dados socioeconômicos, educacionais e de comércio exterior de localidades brasileiras, fornecidas por quatro repositórios. O primeiro deles, a **Relação Anual de Informações Sociais (RAIS)** contém dados fornecidos pelos ministérios do Trabalho e Emprego a respeito do mercado de trabalho e atividades econômicas dos municípios brasileiros, nos anos de 2002 até 2014, medidos anualmente. Já o **Banco de Dados da Secretaria de Comércio Exterior do Ministério do Desenvolvimento, Indústria e Comércio Exterior (SECEX/MDIC)** contém dados a respeito do movimento comercial do Brasil com as demais nações do mundo, em termos de importações e exportações, nos anos de 2000 até 2016, medidos mensalmente. Finalmente, as bases de **Censo Escolar (SC)** e **Ensino Superior (Hedu)** contém dados estatístico-educacionais do ensino básico e superior, envolvendo matrículas, vagas e cursos oferecidos, dentre outras informações, nos anos de 2007 até 2015. Os indicadores utilizados neste estudo são apresentados na Tabela 1.

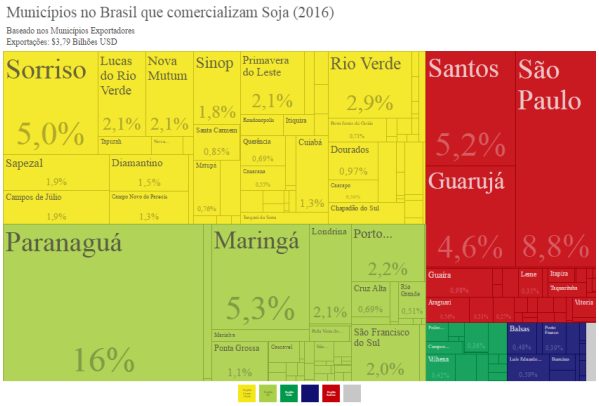
Table 1: Indicadores utilizados na análise.

Repositório	Atributos
RAIS	Renda Mensal Total Renda Mensal Média Total de Empregos Total de Estabelecimentos
SECEX/MDIC	Exportações Importações Peso das Exportações Peso das Importações Complexidade Econômica Crescimento Nominal das Exportações (1 e 5 anos) Crescimento Nominal das Importações (1 e 5 anos) Diversidade de Produtos Diversidade Efetiva de Produtos Diversidade de Destino das Exportações Diversidade Efetiva de Destino das Exportações
Censo Escolar (SC)	Alunos Matriculados Classes Idade Média Crescimento Nominal
Censo do Ensino Superior (Hedu)	Alunos Matriculados Alunos Concluintes Ingressantes (Manhã, Tarde, Noite, Integral) Idade Média

A plataforma DataViva contém 11 aplicativos que empregam diversas estratégias de visualização, tais como a ***Treemap*** [7] (Figura 1a). Nessa técnica, o espaço de visualização é particionado em um conjunto de áreas retangulares, com área proporcional ao conteúdo das instâncias ou grupo de instâncias, de forma que informações mais importantes tenham maior destaque no *layout*. Outra técnica empregada é a ***Heatmap***, integrada a mapas que mostram a distribuição geográfica dos dados (Figura 1b). Nessa técnica, o valor de determinado atributo é mapeado para um intervalo contínuo de cores, de forma que a primeira cor do intervalo é mapeada para o valor mais baixo do atributo, e a última cor mapeada para o valor mais alto. Valores intermediários são mapeados em cores intermediárias do intervalo.

Outras técnicas empregadas utilizam redes de relacionamento, gráficos de dispersão e de setores (gráficos de rosca).

[1]http://www.dataviva.info/



(a) Treemap.

(b) Heatmap aplicado ao mapa geográfico.

Figure 1: Exemplos de visualização na plataforma DataViva.

No geral, essas técnicas mapeiam cada valor de atributo de cada instância da coleção em uma metáfora visual, baseada em cor, forma ou ícone no plano. Em algumas dessas técnicas, os valores de dois dos atributos são mapeados em posições de um ícone no plano, e o restante é mapeado em propriedades visuais dos ícones, tais como cor, formato, orientação, tamanho, entre outros. A densidade dos dados e as dimensões de exibição do *layout* formam assim padrões que variam de acordo com as características da coleção, sendo facilmente percebidos pelo usuário.

A plataforma representa uma importante ferramenta de análise e gestão estratégica governamental e privada, com capacidade de revelar tendências a respeito do andamento socioeconômico de forma prática e eficiente. Pode-se analisar as mudanças de uma variável ao longo do tempo, observar a participação percentual de várias variáveis, analisar o crescimento da participação de uma variável ao longo do tempo, dentre outras tarefas. O DataViva tem potencial para facilitar a análise e auxiliar na tomada de decisão, organizando visualmente os dados dos repositórios que utiliza.

As técnicas de visualização empregadas no DataViva são úteis para a identificação de correlações entre os atributos, representados pelos indicadores medidos em cada localidade. No entanto, apresentam limitações relacionadas à capacidade de visualização e interpretação, quando a coleção de dados possui muitas instâncias. Além disso, os *layouts* produzidos, apesar de permitirem a inferência de conhecimento, podem não ser intuitivos para usuários que não estão acostumados com as técnicas, exigindo aprendizado e prática.

3. SISTEMA DE ANÁLISE

Um sistema computacional foi implementado para a realização das análises dos dados fornecidos pelo DataViva, cuja estrutura é descrita na Figura 2. Cada módulo apresentado é descrito a seguir.

Módulo de Busca: coleta, nos repositórios do DataViva, os dados a serem analisados pelo sistema. O usuário, utilizando o módulo de Exibição/Interação, pode aplicar filtros que serão utilizados para construir dinamicamente um endereço web, representando uma consulta ao repositório.

Módulo de Pré-Processamento: realiza o tratamento dos dados brutos retornados pelo módulo de busca. Os dados do DataViva possuem diversos intervalos de valores e escalas, o que torna necessária a realização desse pré-processamento. Duas operações são feitas nos dados. A **normalização** evita que dados medidos em intervalos ou escalas diferentes atrapalhem o mapeamento para o espaço de visualização, prejudicando a análise. Já o **tratamento de valores ausentes** é aplicado porque, nos repositórios utilizados neste trabalho, nem todos os atributos foram colhidos para todos os anos. A técnica de tratamento de dados ausentes utilizada por esse módulo é a **imputação por média**, que atribui aos atributos ausentes o valor da média dos valores observados para as instâncias da coleção. Esse método representa um dos mais utilizados, tendo alcançado resultados satisfatórios em tarefas de classificação. Além disso, não apresenta diferenças significativas de resultado quando comparados a outros métodos de imputação simples, como imputação por mediana [1], sendo satisfatório para utilização nos experimentos com os dados do DataViva.

Módulo de Visualização: aplica uma técnica de visualização baseada em projeção multidimensional aos dados fornecidos pelo módulo de pré-processamento. Projeções, apesar de apresentarem considerável (e inevitável) perda de informação, especialmente em relação às configurações complexas do espaço original de características, representam uma das ferramentas mais comuns de exploração da estrutura de espaços multidimensionais. Considerando um conjunto m-dimensional $X = x_1,x_2,...,x_n$, com $\sigma(x_i,x_j)$ representando uma medida de similaridade qualquer entre x_i e x_j, e um conjunto bidimensional $Y = y_1,y_2,...,y_n$, com $d(y_i,y_j)$ representando a distância Euclidiana entre y_i e y_j no plano de visualização, uma projeção multidimensional representa uma função $f:X \rightarrow Y$ que procura minimizar a diferença entre σ e d, ou seja, $|\sigma(x_i,x_j)$ - $d(y_i,y_j)|$ se aproxime de zero [12]. As técnicas de projeção utilizadas nas análises são apresentadas na Seção 4.

Módulo de Exibição/Interação: exibe ao usuário os *layouts* criados pelo módulo de visualização. Um conjunto de ferramentas foi desenvolvido para auxiliar a interação do usuário com esses *layouts*, e consequentemente melhorar o processo de análise. É possível aplicar diversos filtros na consulta ao repositório de dados para direcionar a análise, tais como ano, localidade e nível de agregação (município,

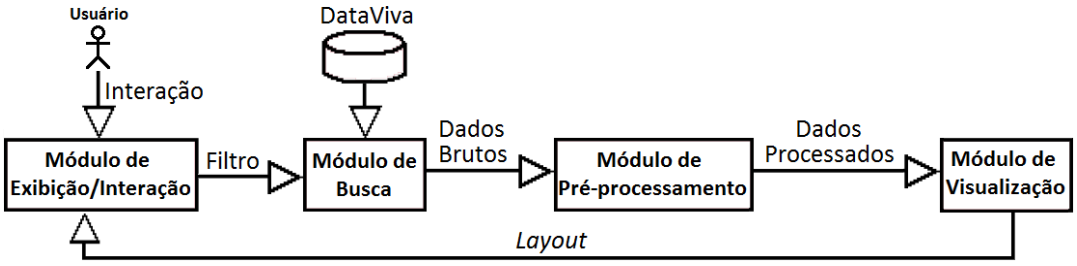


Figure 2: Diagrama de funcionamento do sistema.

microrregião, mesorregião, estado e região).

A Figura 4 apresenta um exemplo de *layout* PCA, que será detalhado na Seção 5. Cada círculo representa uma localidade (no exemplo, estados e o distrito federal do Brasil), e a cor de cada círculo representa uma informação específica de identificação retornada pela consulta (no exemplo, região geográfica à qual o estado pertence). A ideia deste *layout* é que o usuário possa identificar localidades com perfis similares ou complementares, compreender como é a estrutura de determinada região do país, ou como as localidades se distribuem em termos de perfil comportamental.

Ao passar o mouse sobre uma instância é exibido um texto com sua identificação. Ao clicar sobre ela são exibidas informações detalhadas oriundas do repositório correspondente, como mostra a Figura 7. Também é possível utilizar uma ferramenta de *zoom* para analisar regiões específicas do *layout* em um nível maior de detalhes. Essa funcionalidade pode ser acessada de duas formas: utilizando um clique duplo na instância sobre a qual se deseja realizar a análise, ou utilizando a rolagem do mouse na região de interesse.

4. TÉCNICAS DE PROJEÇÃO MULTIDIMENSIONAL

Diversas técnicas de visualização de dados multidimensionais podem ser encontradas na literatura [10, 13]. Com o intuito de verificar se projeções multidimensionais são capazes de comunicar a estrutura geral dos repositórios fornecidos pelo DataViva, bem como realçar relacionamentos entre as instâncias (determinados por seus perfis comportamentais), duas estratégias clássicas foram escolhidas para a análise realizada neste projeto, ambas com reconhecida aplicação em diversas áreas do conhecimento. A primeira estratégia utiliza a técnica de análise de dados multivariada ***Principal Component Analysis*** **(PCA)** [8], largamente utilizada em diversos cenários envolvendo mineração de dados e visualização de coleções, com resultados satisfatórios [2, 3]. PCA representa uma técnica de projeção linear que busca encontrar direções ortogonais, chamadas de **componentes principais**, com idealmente máxima variância. A primeira componente representa, assim, a direção com a maior variância, a segunda componente representa a direção, com maior variância, ortogonal à primeira componente, e assim por diante. Neste trabalho, as duas primeiras componentes foram consideradas como dimensões a serem utilizadas na criação do *layout*.

A segunda técnica utiliza uma projeção baseada em ***Multidimensional Scaling*** **(MDS)** [5]. Essa classe de técnicas tem como objetivo preservar os relacionamentos entre as instâncias, observados no espaço original de atributos, no plano de visualização [4]. Assim, considerando um espaço com m dimensões, as técnicas MDS buscam um espaço com p dimensões, $p < m$, geralmente Euclidiano, no qual as distâncias entre as instâncias coincidem ou se aproximam das distâncias entre essas mesmas instâncias no espaço original [5]. O MDS métrico foi o escolhido para fazer parte deste sistema, cujo objetivo é encontrar um formato de pontos em que as distâncias entre eles são associadas às dissimilaridades entre os objetos por uma função de transformação.

5. RESULTADOS

Esta seção apresenta os resultados de diversas análises nos dados oriundos dos repositórios do DataViva. Essas análises buscam encontrar localidades de destaque, grupos de localidades com perfis semelhantes ou complementares em seus segmentos, além de padrões e tendências com relação ao comportamento dessas localidades, de forma a direcionar a criação de medidas que estimulem seu desenvolvimento.

5.1 Repositórios RAIS e SECEX/MDIC

A Figura 3 mostra os municípios das mesorregiões do Triângulo Mineiro/Alto Paranaíba e do Norte de Minas organizados em relação ao comércio internacional em 2014 (repositório SECEX/MDIC), e a cor dos círculos representa a mesorregião a qual pertencem (verde: Norte de Minas, roxo: Triângulo Mineiro).

Figure 3: *Layout* PCA dos dados de 2014 do repositório SECEX/MDIC para os municípios das mesorregiões do Triângulo Mineiro e Norte de Minas.

É possível perceber alguns grupos bem definidos no *layout*. Quatro deles foram destacados para esta análise. Nesses grupos, a maioria dos municípios pertence a apenas uma das mesorregiões, mas há a presença de municípios de outras mesorregiões, mostrando que mesmo com a distância

geográfica, existem semelhanças no perfil de negócios internacionais que podem justificar uma parceria de fornecimento de produtos entre esses municípios. O grupo 1 possui apenas um município da mesorregião do Norte de Minas, Pirapora. Os principais produtos exportados pelos municípios desse grupo são produtos de origem vegetal e gêneros alimentícios, sendo a China o principal destino de exportação. Os indicadores relativos à importações foram determinantes no posicionamento próximo dessas instâncias no *layout*. O mesmo pode ser observado no grupo 2, que exporta principalmente produtos de origem vegetal e gêneros alimentícios, e no grupo 3, que exporta principalmente produtos de origem vegetal e animal. O *layout* exibe assim, de maneira simples, como essas localidades tem comportamento similar, permitindo a análise de grupos de interesse por parte do especialista. Finalmente, o grupo 4 é formado apenas pelo município de Uberaba, no Triângulo Mineiro. Seu volume e peso de importações diferenciado, bem como sua diversidade de produtos, faz com o que o município se distancie dos demais no *layout*, e isso se deve ao fato da cidade possuir uma empresa de agrotóxicos com grande mercado nacional, que demanda a importação de grandes quantidades de produtos químicos.

A Figura 4 mostra os estados e o distrito federal do Brasil organizados em relação ao comércio internacional em 2014 (repositório SECEX/MDIC). A cor dos círculos representa a região geográfica do Brasil à qual o estado pertence. Três grupos foram destacados para esta análise.

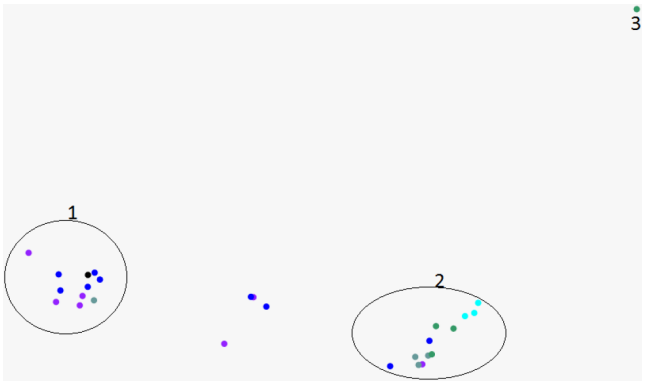


Figure 4: *Layout* PCA dos dados de 2014 do repositório SECEX/MDIC para os estados brasileiros.

O *layout* aqui também destaca grupos com perfis distintos, principalmente em relação os valores dos indicadores relativos à importação/exportação. A composição de grupos no *layout* reflete a separação entre as regiões geográficas do Brasil em termos de produção, e a diferença existente entre a maior parte dos estados das regiões norte e nordeste em relação aos estados das demais regiões. Tais estados possuem um nível de industrialização menor e seus territórios não são tão propícios para a agricultura e pecuária, apesar deles se destacarem em suas regiões respectivas. O grupo 1 é formado basicamente por estados das regiões geográficas Norte e Nordeste, sendo o distrito federal a única exceção. Já o grupo 2 concentra basicamente estados da região Sudeste (com exceção de São Paulo) e Sul, e esse posicionamento reflete a importância da agropecuária, agricultura e principalmente produtos minerais nesses estados, devido ao volume de reservas minerais de estados como Rio de Janeiro, Minas Gerais e Espírito Santo. O grupo 3 é formado apenas pelo estado de São Paulo, que exibe um comportamento peculiar, apresentando os valores mais altos de exportação e importação do país, provavelmente devido à alta concentração de indústrias no estado.

A Figura 5 mostra os municípios de Minas Gerais organizados em relação ao mercado de trabalho formal em 2013 (repositório RAIS).

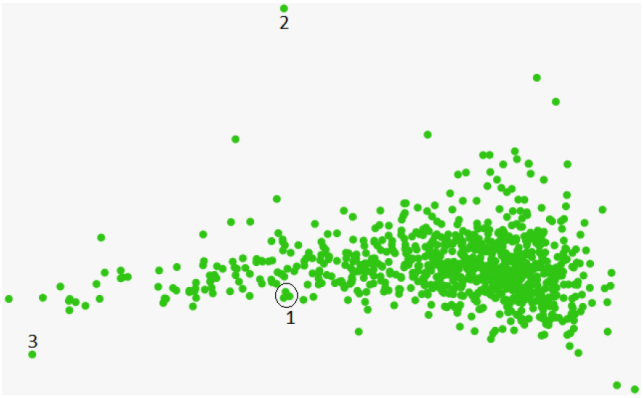


Figure 5: *Layout* PCA dos dados de 2013 do repositório RAIS para os municípios de Minas Gerais.

De uma forma geral, os municípios formam diversos grupos não bem definidos, e o *layout* reflete o comportamento de um estado heterogêneo e extenso geograficamente, limítrofe a três regiões geográficas brasileiras distintas. Isso dificulta a identificação e seleção de grupos para análise, mas ainda assim é possível perceber alguns pontos isolados, representando municípios com comportamento peculiar. Dois desses municípios isolados foram destacados para análise, juntamente com um pequeno grupo (grupo 1), composto pelos municípios de Pitangui, na região metropolitana de Belo Horizonte, e os municípios de Santos Dumont e Leopoldina, na região da Zona da Mata. O *layout* destaca a semelhança do perfil de empregabilidade entre esses municípios, de forma que um empreendedor bem sucedido em um deles estude a possibilidade de começar uma expansão de seu negócio nos outros municípios desse grupo. O grupo 2 mostra o município de Jeceaba, também na região metropolitana de Belo Horizonte, cujo posicionamento no *layout* destaca seu comportamento consideravelmente diferente dos demais municípios do estado. Esse posicionamento é determinado por diversos indicadores: sua renda mensal média é maior do que as rendas dos municípios de Uberlândia ou mesmo Belo Horizonte. O crescimento nominal de salários e de empregos (1 e 5 anos) também é elevado (55%), e os salários aumentaram 108%. Jeceaba é um exemplo de município pequeno (5.395 habitantes em 2010), com boa parte da sua população bem remunerada, representando uma localidade potencial para geração de novos negócios. Finalmente, o grupo 3 mostra o município de Belo Horizonte, cujo posicionamento no *layout* é determinado pelo total de empregos e de estabelecimentos. Esse resultado é esperado, uma vez que o município, que é a capital do estado, é naturalmente atrativo para investimentos diversos.

5.2 Repositórios Hedu e SC

A Figura 6 mostra os municípios de Minas Gerais orga-

nizados em relação aos dados do ensino superior em 2013 (repositório Hedu). Três grupos estão destacados para esta análise.

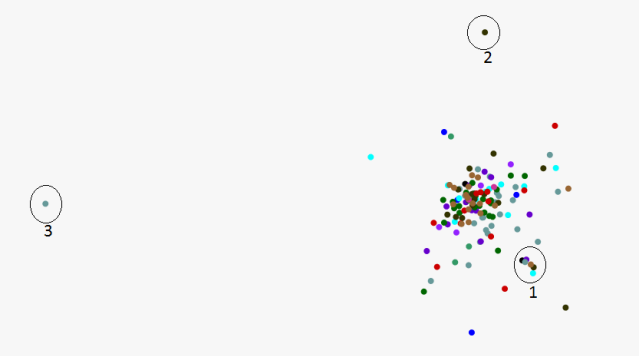


Figure 6: *Layout* MDS dos dados de 2013 do repositório Hedu dos municípios de Minas Gerais.

De uma forma geral, os municípios formam diversos grupos não muito bem definidos, e o *layout* reflete novamente o comportamento de um estado heterogêneo e extenso. O grupo 1 é formado por alguns dos municípios que só possuem uma instituição de ensino superior de pequeno porte, que oferecem apenas cursos no período noturno. Uberlândia é o único componente do grupo 2 por ser o município de Minas Gerais com o segundo maior número de alunos matriculados, concluintes, ingressantes e que estudam nos períodos da manhã, tarde e noite, além de ser o município mineiro com o maior número de alunos no período integral. Belo Horizonte é o único componente do grupo 3, e seus indicadores são mostrados na Figura 7 pela utilização da ferramenta de interação apresentada na Seção 3. Essa ferramenta possibilita verificar os valores dos indicadores **Alunos Matriculados**, **Alunos Concluintes** e **Ingressantes** (todos os períodos), que são altos para esse município, representando os maiores valores para o estado.

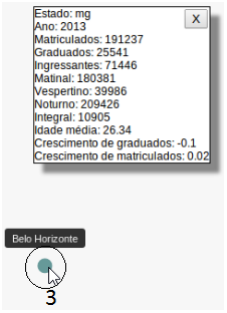


Figure 7: Seleção do município de Belo Horizonte (grupo 3), com exibição dos valores correspondentes dos indicadores.

Finalmente, é interessante notar que Uberlândia e Belo Horizonte representam grupos isolados que refletem seus papéis de polo educacional. Há também uma grande distância entre os municípios destacada no *layout*. É natural que Belo Horizonte, por ser capital do estado, e dentro de uma área metropolitana densa, apresente altos valores para os índices medidos. Uberlândia, por outro lado, está distante geograficamente da capital, e esse perfil pode refletir, dentre outros fatores, a necessidade de crescimento isolado da região na qual está inserida.

A Figura 8 mostra microrregiões dos estados de Santa Catarina e Rio Grande do Sul organizados em relação ao ensino superior em 2013 (repositório Hedu). A cor dos círculos representa o estado a qual pertencem (Preto: Rio Grande do Sul, roxo: Santa Catarina). Nesta análise também estão destacados três grupos.

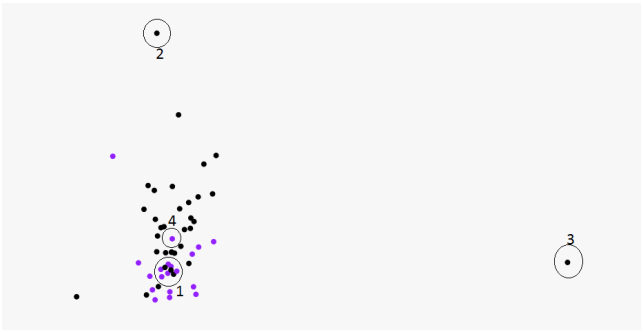


Figure 8: *Layout* MDS dos dados de 2013 do repositório Hedu das microrregiões dos estados de Santa Catarina e do Rio Grande do Sul.

É possível perceber uma clara separação entre os estados no *layout*, mas com uma região de interseção. O grupo 1 abrange as microrregiões catarinenses de Joinville, Blumenau, Itajaí, Criciúma, que, dentro do grupo, possuem os maiores números de alunos matriculados, concluintes e que estudam no período noturno. Além disso, o grupo contém os municípios gaúchos de Ijuí, Santa Rosa e Erechim. A análise desse grupo revela microrregiões com perfis educacionais semelhantes, mas que estão distantes geograficamente, especialmente as microrregiões catarinenses. Essa informação, destacada pelo *layout* dificilmente seria notada em uma análise a dados tabulares, e pode auxiliar governantes a criar eventuais parcerias, caso não existam, além de aplicar experiências de sucesso, empregadas em uma ou algumas dessas localidades, em outras do mesmo grupo. Cabe ressaltar também que o grupo ressaltado pelo *layout* apresenta localidades com bons índices de matrículas, e com aumento no ano considerado. Dessa forma, pode ser interessante realizar um estudo das estratégias utilizadas aqui, de forma que sejam aplicadas também a localidades de outros grupos. O grupo 2 é formado pela microrregião de Sananduva, cuja posição isolada no *layout* é justificada pelos valores dos indicadores **alunos matriculados**, **concluintes** e **ingressantes**. Nesse caso, o *layout* consegue destacar, de forma simples, uma microrregião com um comportamento peculiar, aqui representado pela estrutura educacional modesta ou inexistente, e propiciar um estudo mais profundo que guie a criação de estratégias de desenvolvimento para essa localidade. Finalmente, o grupo 3 é formado somente pela microrregião de Porto Alegre. É a microrregião, dentre as representadas no *layout*, com a maior quantidade de alunos matriculados, concluintes e ingressantes, e possui a maior quantidade de alunos estudando em todos os períodos (manhã, tarde, noite e integral), o que determinou sua posição isolada no *layout*. O grupo 4, formado pela microrregião de Florianópolis, não apresenta um posicionamento isolado, naturalmente esperado para uma capital. Cabe uma análise aprofundada das políticas educacionais adotadas nessa mi-

crorregião, de modo a verificar se há alguma deficiência que possa ser corrigida.

A Figura 9 mostra as mesorregiões de todo país organizadas em relação aos dados do censo do ensino básico de 2014 (repositório SC). A cor dos círculos representa a região geográfica do Brasil à qual a mesorregião pertence. Cinco grupos foram destacados para esta análise.

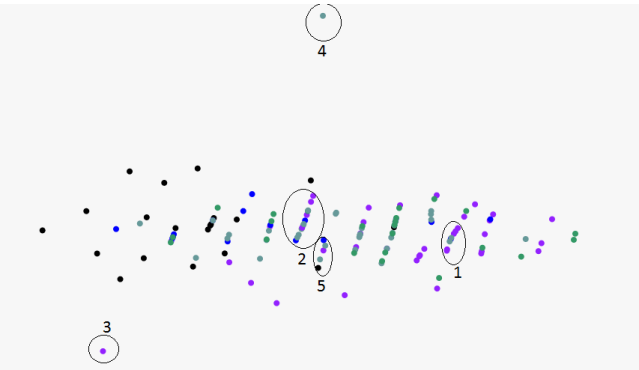


Figure 9: *Layout* MDS dos dados de 2014 do repositório SC das mesorregiões de todo país colorido por região.

O grupo 1 é composto pelas mesorregiões Centro Sul e Norte Baiano, Jequitinhonha, Presidente Prudente, Norte Goiano, Jaguaribe, Sertão Alagoano e Agreste Paraibano. Esse é um grupo que, com exceção da mesorregião de Presidente Prudente, é composto por localidades com menos recursos financeiros. O *layout* ressalta a semelhança no perfil educacional dessas mesorregiões, e o posicionamento curioso da mesorregião de Presidente Prudente - em especial considerando a região geográfica do Brasil na qual se localiza, destacando assim um fato que merece atenção por parte dos respectivos governantes. O grupo 2 é bem heterogêneo, com 6 mesorregiões da região Sudeste (5 do estado de São Paulo), 4 da região Nordeste (3 do estado do Maranhão), além de 2 do estado do Mato Grosso, na região Centro-Oeste. O posicionamento dessas mesorregiões no *layout* facilita a identificação, por parte dos especialistas, de potenciais parcerias entre essas localidades com perfil educacional semelhantes, além do compartilhamento de experiências que tenham resultado satisfatório. O grupo 3 é formado pela mesorregião metropolitana de Recife, que possui os estudantes com a maior idade média, além do maior crescimento no número de alunos matriculados no período de 1 ano, indicadores que influenciaram no seu posicionamento no *layout*. A mesorregião metropolitana de São Paulo, representada pelo grupo 4, foi posicionada isoladamente no *layout*, devido aos valores dos indicadores relativos ao número de escolas, classes e alunos matriculados, com valores sensivelmente maiores do que os do restante das mesorregiões. Finalmente, o grupo 5, composto pelas mesorregiões do Rio de Janeiro, região metropolitana de Belém, Mata Paraibana, Oeste de Minas e Pantanal Sul mato-grossense destaca um fato interessante: a concentração de localidades tão distantes geograficamente, mas com perfis educacionais tão semelhantes.

6. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou a aplicação de duas técnicas de projeção multidimensional aos dados de quatro repositórios da plataforma DataViva, com o intuito de aumentar a capacidade de extração de informação dessas fontes de dados. A ideia principal foi complementar as ferramentas de análise já existente na plataforma, potencializando o poder de compreensão desses dados pelo usuário.

Os resultados obtidos permitiram identificar informações relevantes dos repositórios. Os *layouts* destacaram grupos de localidades com perfis comportamentais semelhantes, assim como localidades com comportamento peculiar em relação aos indicadores considerados, que merecem uma análise detalhada por parte dos especialistas. A maioria das localidades com grande concentração de habitantes se destacou com relação à maioria dos indicadores, mas também foi possível identificar localidades de destaque com menos habitantes. Em alguns casos, isso pode representar o resultado de estratégias interessantes desenvolvidas pelos governos, do ponto de vista de desenvolvimento. Em outros casos, a causa pode ser a ocorrência de eventos específicos ligados à iniciativa privada, ou mesmo fatos ocorridos que sejam excepcionais ou inusitados. De qualquer forma, a análise com base nos *layouts* permitiu a identificação de ocorrências que seriam mais difíceis de se verificar pela análise localidade por localidade, no formato tabular. Além disso, como os dados do DataViva são disponibilizados publicamente, as estratégias visuais empregadas aqui permitem que os cidadãos explorem a estrutura comportamental das localidades brasileiras, representando assim um canal efetivo de comunicação entre governo e população.

Os próximos passos deste trabalho envolvem a investigação de novas estratégias de visualização, para destacar outras perspectivas nos dados, além de corrigir limitações conhecidas nas técnicas empregadas neste estudo. Pretende-se também explorar a combinação dessas técnicas com algoritmos de aprendizado de máquina e mineração de dados, tais como técnicas de agrupamento, para potencializar as informações exibidas para o usuário. Finalmente, é interessante desenvolver novas funcionalidades de interação, utilizando inclusive a coordenação de múltiplas técnicas de visualização, para que seja possível uma inserção ainda mais efetiva do usuário na análise. Nesse sentido, um avaliação com usuários é prevista, no intuito de avaliar a percepção dos especialistas ao analisar os *layouts*, bem como propiciar a descoberta de necessidades que guiem a criação de ferramentas de interação adicionais.

7. ACKNOWLEDGMENTS

Os autores agradecem à FAPEMIG, CAPES e CNPQ pelo auxílio no desenvolvimento deste projeto.

8. REFERENCES

[1] E. Acuna and C. Rodriguez. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pages 639–647. Springer, 2004.

[2] J. P. Anzola, L. A. Rodríguez, and G. M. Tarazona. Exploring data by pca and k-means for ieee xplore digital library. In *Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society*, page 15. ACM, 2016.

[3] F. Begam and S. Kumar. Visualization of chemical space using principal component analysis. *World Applied Sciences Journal*, 29:53–59, 2014.

[4] P. Borg, I.; Groenen. Modern multidimensional scaling: Theory and applications. *New York: Springer Series in Statistics*, 1997.

[5] T. F. Cox and M. A. Cox. *Multidimensional scaling*. CRC Press, 2010.

[6] D. Jamróz. Application of multidimensional data visualization in creation of pattern recognition systems. In *Man-Machine Interactions 3*, pages 443–450. Springer, 2014.

[7] B. Johnson and B. Shneiderman. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the IEEE Conference on Visualization, 1991, Visualization'91*, pages 284–291. IEEE, 1991.

[8] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[9] D. Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[10] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, 23(3):1249–1268, 2017.

[11] S. R. Panta, R. Wang, J. Fries, R. Kalyanam, N. Speer, M. Banich, K. Kiehl, M. King, M. Milham, T. D. Wager, et al. A tool for interactive data visualization: Application to over 10,000 brain imaging and phantom mri data sets. *Frontiers in neuroinformatics*, 10, 2016.

[12] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008.

[13] G.-D. Sun, Y.-C. Wu, R.-H. Liang, and S.-X. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867, 2013.

[14] E. Tejada, R. Minghim, and L. G. Nonato. On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization*, 2(4):218–231, 2003.