Análise Espaço-Temporal de Indicadores da Saúde na Região Nordeste Usando Técnicas de Mineração de Dados Geográficas e a Ferramenta VIS-STAMP

Adeline Marinho Maciel [1]
Cláubio Landney Lima Bandeira [1]
Marcelino Pereira dos Santos Silva [2]
Angélica Félix de Castro [3]

[1] Programa de Pós-Graduação em Ciência da Computação - MCC
Universidade do Estado do Rio Grande do Norte – UERN/
Universidade Federal Rural do Semi-Árido – UFERSA
{adelsud6, claubiobandeira}@gmail.com

[2] Universidade do Estado do Rio Grande do Norte - UERN
BR 110 – Km 46 – Bairro Costa e Silva - Campus Central
59.625-620 Mossoró – RN, Brasil
prof.marcelino@gmail.com

[3] Universidade Federal Rural do Semi-Árido – UFERSA
BR 110 – Km 47 – Bairro Costa e Silva - Campus Central
59.625-900 - Mossoró – RN, Brasil
angelicafcastro@gmail.com

Abstract. The use of techniques that involve phenomena that vary in both space and time are indispensable. These phenomena, for example, can be related to deforestation, erosion, occupation of hillsides among others. This way, arise computational tools called Geographic Information System (GIS) that have evolved so that support the modeling of these phenomena, allowing since the storage and the visualization even the use of data mining techniques. By using geographical data, arise a new field called Geographical Data Mining. Among the methods used by it is the clustering. This method is very important in the mining process to be able to extract structures directly of the data, without any prior knowledge. Given this context, we aim to expose the use of GIS as a tool for time-space analysis and through the use of Geographical Data Mining, specifically using cluster methods, identifying patterns presents in the data. For this, we used the GIS A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP) and a case study is approached using indicators related to health of the Nordeste region of Brazil, from 2001 to 2006, specifically the dengue. With this analysis aims to verify the existence of correlations between the dengue and other indicators this region.

Palavras-chave: Geographic Information System, Geographical Data Mining, clustering, Sistemas de Informações Geográficas, Mineração de Dados Geográficos, agrupamento.

1. Introdução

O estudo de eventos relacionados a modelos espaço-temporais é de grande importância na análise e avaliação de diversos tipos de fenômenos, tais como: crescimento urbano, desmatamento, poluição, entre outros. Além disso, por envolver o estudo de grandes volumes de dados, a tarefa de análise espaço-temporal torna-se impossível de ser realizada através de técnicas manuais de forma eficiente, pois estas demandariam muito tempo e estariam propensas a vários erros. Assim, ferramentas que permitam desenvolver modelos que sejam capazes de representar apropriadamente fenômenos que variam tanto no espaço como no tempo são indispensáveis. Dessa forma, surge a necessidade de utilizar ferramentas computacionais.

As ferramentas que tratam dados relacionados à localização no espaço são denominadas Sistemas de Informações Geográficas (SIG). Estas estão aptas a armazenar dados, possibilitando o mapeamento, estruturação e análise destes. Porém, os SIG vêem evoluindo

para que sejam capazes de modelar o comportamento de determinados objetos em sua trajetória espaço-temporal. Assim, é possível a ordenação, visualização, análise quantitativa e até mesmo a utilização de técnicas de mineração de dados e descoberta de conhecimento, visando à identificação de padrões e tendências nestes dados. Isto faz surgir um novo conceito: Mineração de Dados Geográficos.

Diante deste contexto, este trabalho foi desenvolvido com o objetivo de aplicar técnicas de Mineração de Dados Geográficos, visando identificar padrões espaço-temporais existentes nos dados. Para tanto, será apresentado um estudo de caso utilizando o programa *A Visualization System for Space-Time and Multivariate Patterns* (VIS-STAMP) sobre índices referentes ao Dengue na região Nordeste. Além disso, foi utilizado outros indicadores desta região dos anos de 2001 a 2006 com o objetivo de verificar a correlação entre esses indicadores e a taxa de incidência de Dengue.

2. Metodologia de Trabalho

2.1 Área de Estudo

A área de estudo escolhida foi a região Nordeste, umas das cinco regiões do Brasil. Essa região está localizada no norte do país e ocupa uma área de 1.561.177,8 km^2, o que corresponde a 18,26% da área total do país.

De acordo com o censo 2010 disponível em CENSO (2010) a região Nordeste possui mais de 51 milhões de habitantes, sendo a segunda região mais populosa do país. Possui o maior número de estado do país, sendo eles: Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe e Bahia. Estes estados possuem heterogeneidades entre suas características básicas, por exemplo, cultura, diversidade populacional, espaço territorial, renda, escolaridade, dentre outros fatores relevantes e específicos de cada estado.

2.2 Materiais e Métodos

Para essa pesquisa foi utilizada uma malha digital da região Nordeste disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em IBGE (2010), o qual forneceu os dados relacionados à extensão territorial e população.

Os dados para análise espaço-temporal foram retirados do *site* do Ministério da Saúde, o qual disponibilizou as bases de informações mediante endereço eletrônico do Departamento de Informática do Sistema Único de Saúde (DATASUS) disponível em DATASUS (2010). Nele foi realizada a seleção de índices que possuíam relação com os estados da região Nordeste. Esses índices foram retirados dos Indicadores de Dados Básicos (IDB), disponibilizados através da seção sobre as Informações de Saúde. Foi estabelecido um período de seis anos para os dados, caracterizando assim, o tempo correspondente aos anos de 2001 a 2006, e de acordo com a Pesquisa Nacional por Amostra de Domicílios - PNAD (2001 a 2007), para todos os índices correspondentes aos nove estados da região Nordeste.

Abaixo segue a descrição dos índices selecionados:

- Taxa de Analfabetismo - % da população de 15 ou mais anos não alfabetizada.
- Nível de escolaridade - % da população com 8 e mais anos de estudo.
- Produto Interno Bruto (PIB) *per capta* – Produto Interno Bruto (Referência 2000), nova metodologia de cálculo do PIB nacional e regional.
- Cobertura de esgotamento sanitário – % da população servida por esgoto.
- Cobertura de coleta de lixo – % da população servida por coleta de lixo.
- Gasto público com saúde *per capta* - valores brutos dos gastos em milhões de reais; valores *per capta* em reais.
- Taxa de incidência de doenças transmissíveis (Dengue) – Taxa de incidência: casos por 100.000 habitantes.

- Proporção de pobres – % de pobres, valor de referência, salário mínimo de 2007, é de R$ 380,00.
- Taxa de desemprego - % da população com mais de 10 anos desocupada.

Para a realização da análise espaço-temporal e a verificação de correlação entre os indicadores dos estados nordestinos, foi utilizado o SIG VIS-STAMP proposto por Guo *et al*. (2006). Esta ferramenta realiza a análise geo-visual entre dados espaço-temporais com o objetivo de compreender, explorar e visualizar padrões complexos através de dimensões multivariadas, espacial e temporal. Para o desenvolvimento dessa análise, ele utiliza métodos como agrupamento, classificação e visualização.

2.3 Pré-processamento/Transformação

Devido a heterogeneidade dos dados, pois estes foram provenientes de diferentes fontes, estes ficaram suscetíveis a ruídos, com dados faltando e inconsistentes. Assim, uma das tarefas mais importantes em todo o processo que envolve a mineração é justamente eliminar esses ruídos. Para isso, existe a etapa de pré-processamento que permite, dentre outras coisas, eliminar dados inconsistentes, incompletos e discrepâncias.

Uma outra etapa importe é a transformação. Após selecionados, limpos e pré-processados, os dados necessitam serem adequadamente armazenados e formatados para então, aplicar os algoritmos.

Nesta pesquisa, realizamos sobre os dados algumas destas tarefas. Dentre elas, podemos citar que, os dados selecionados correspondiam aos anos de 2001 a 2007, porém, como existiam um elevado número de dados ausentes na maioria dos índices, foi necessário remover os indicadores correspondentes ao ano de 2007, passando-se apenas a analisar o período de 2001 a 2006.

Ainda, observou-se que existiam apenas dois indicadores com uma de suas variáveis sem dados, ou seja, com dados ausentes. Nesse caso, a resolução desse problema se deu pela atribuição de valores nulos (zero) a esses atributos. Não foi escolhida outras técnicas, pois estas acarretariam na obtenção de resultados não tão corretos e precisos.

Além disso, para a geração dos arquivos de entrada no software escolhido, foi necessário a criação de três arquivos. O primeiro arquivo no formato *shapefile* (.shp) corresponde aos limites dos objetos espaciais. Já o segundo, no formato csv, refere-se a todos os atributos de cada objeto espacial presente no *shapefile*. Estes, juntos representam a malha da região Nordeste. Por fim, o terceiro arquivo, denominado "nord_data.csv", continha todos os dados espaço-temporais e multivariados referentes ao .shp.

2.4 VIS-STAMP

Agrupamento, ou *Clustering*, é um dos processos mais utilizados em Mineração de Dados Geográficos. Segundo Miller e Han (2009) ele permite o agrupamento de um conjunto de objetos que possuem alguma similaridade. Dessa forma, uma das características relevantes desse processo é a possibilidade de visualização dos dados, proporcionando assim uma melhor compreensão da estrutura dos objetos, possibilitando assim realizar comparações entre eles. O VIS-TAMP propicia a criação de *cluster* utilizando técnicas baseadas no Método Hierárquico Aglomerativo (MHA). Neste método, os *clusters* mais próximos (similaridade) são fundidos em um *cluster* maior (Neves *et al.*, 2001).

O VIS-STAMP é composto por um *MapMatrix* para a visualização temporal dos mapas; um *Space-Time Matrix* (S-T Matrix) o qual organiza os padrões multivariados no campo espaço-temporal; um *Self Organizing Map* (SOM) que possibilita a representação multivariada de agrupamento e de abstração (incluindo agrupamento de séries temporais); e um *Parallel Coordinate Plot* (PCP) que permite a visualização de padrões multivariados. Eles

são independentes, mas suportam interações para que o usuário possa visualizar detalhes da análise.

3. Resultados e Discussões

Nesta seção, são apresentados os resultados da análise espaço-temporal dos índices apresentados com o VIS-STAMP. Desta forma, na Figura 1 podemos observar o PCP onde cada linha corresponde a um *cluster* multivariado e colorido pelo SOM (ver Figura 2a).

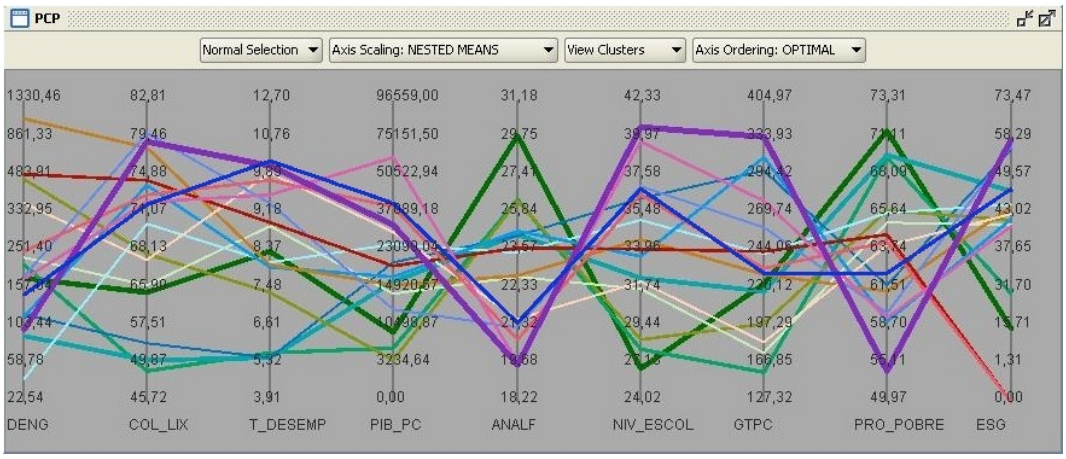
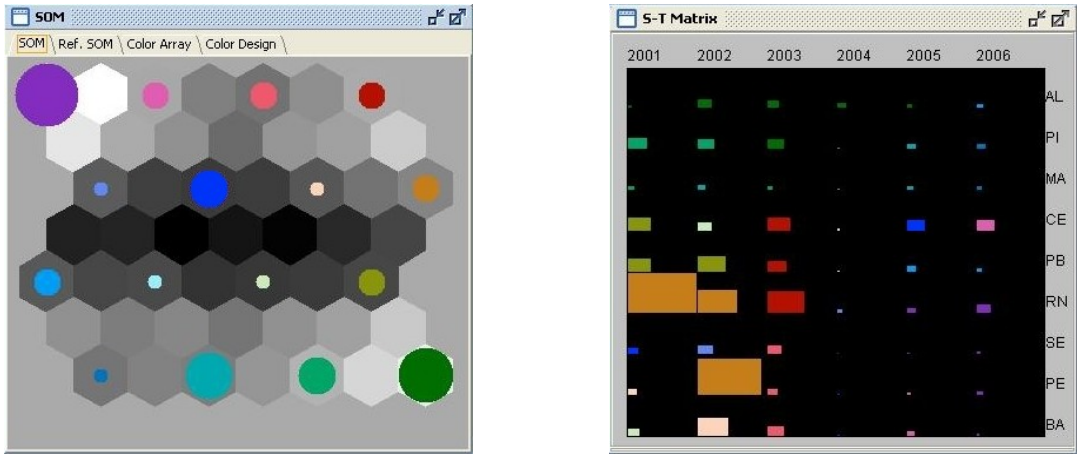


Figura 1: PCP com os *clusters* formados pelos índices.

A Figura 2a corresponde ao SOM que é utilizado em mineração de dados para reduzir a dimensionalidade dos dados, apresentando-os no formato bidimensional. É importante ressaltar que, há uma relação direta entre o número de dados e o tamanho de sua representação. Assim, quanto maior a quantidade de dados presentes no *cluster*, maior será a área do círculo no SOM. Já na Figura 2b, temos a S-T Matrix que permite visualizar as variações espaço-temporais dos padrões multivariados.



a) b)

Figura 2: a) SOM e b) S-T Matrix gerados pelos índices.

A representação do Map Matrix possibilita ao usuário visualizar os *clusters*, mostrando como eles se apresentam e como as mudanças ocorrem em uma determinada linha de tempo. Ele pode ser observado na Figura 3 a seguir.

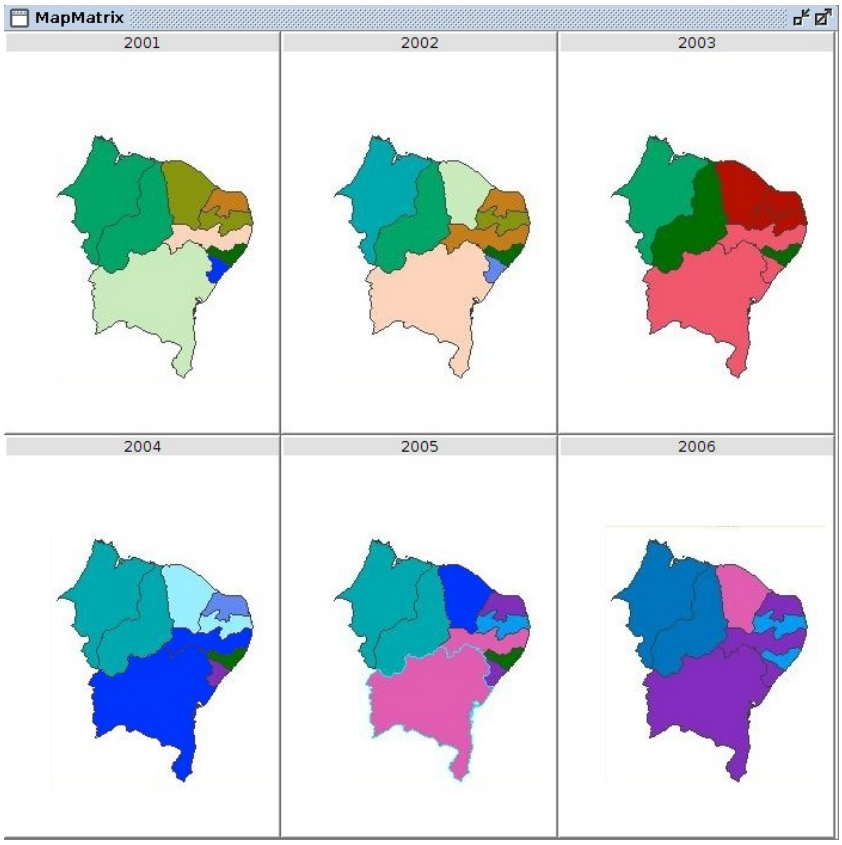


Figura 3: MapMatrix representando a formação de *cluster* espaço-temporal com os estados da região Nordeste.

Observando a Figura 4, a seguir, é possível visualizar a seleção de apenas um *cluster* e através dele, pode-se identificar uma correlação entre os índices. Nela, pode-se verificar relações já esperadas, revelando informações que não apresentam grau de novidade. Por exemplo, quando a Cobertura de Esgotamento Sanitário (ESG) e a Cobertura de Coleta de Lixo (COL_LIX) apresentam valores elevados, a taxa de incidência de dengue (DENG) tende a ser baixa nesse agrupamento. Outro padrão confirmado é que quando a Proporção de Pobres (PRO_POBRE) é baixa, o Gasto Total *Per Capta* e o Nível de Escolaridade são elevados, existe uma diminuição da taxa de analfabetismo (ANALF).

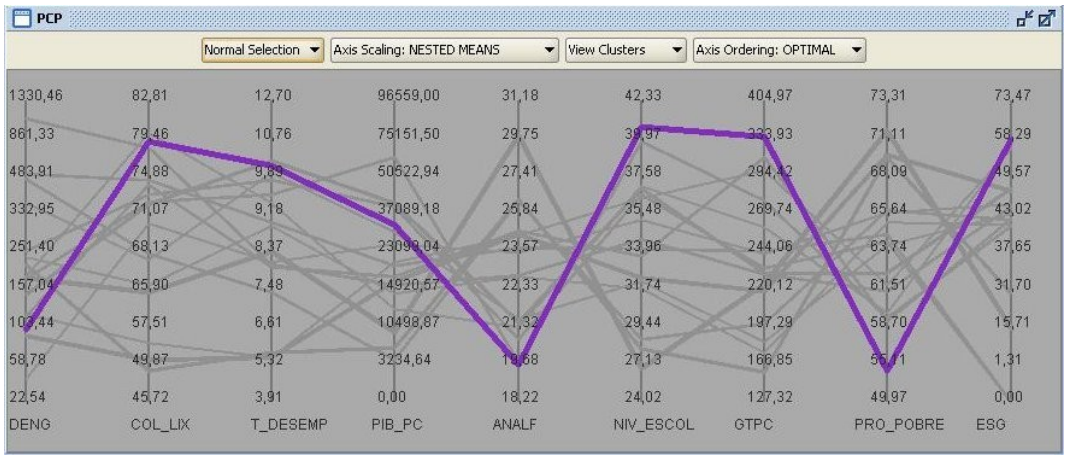


Figura 4: Visão em *cluster* da seleção de um *cluster*.

Na Figura 5 temos o mesmo *cluster* apresentado na Figura 4. Porém, agora, os dados são mostrados individualmente. Por meio do MHA, foi possível a criação dos *clusters* através dos eventos de similaridades entre os dados. Estes possuem características intrínsecas dos seus atributos, o qual possibilita a ocorrência desse processo.

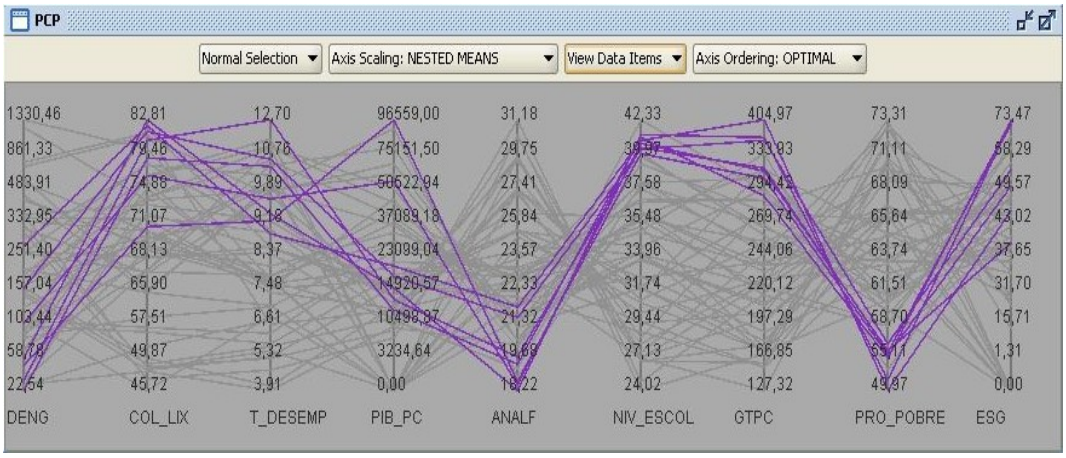
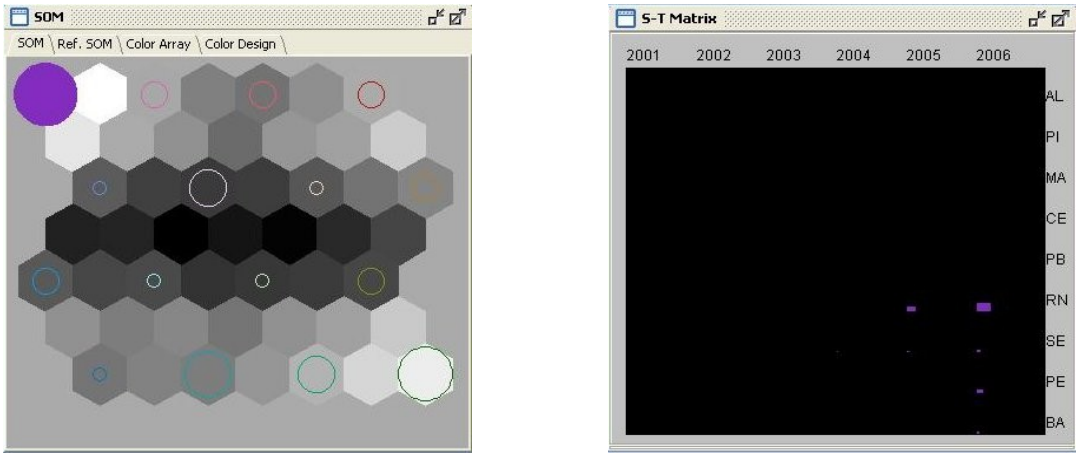


Figura 5: Visão individual dos dados do *cluster* selecionado.

A Figura 6a representa o SOM após a seleção do *cluster*, onde pode-se observar que a dimensão do círculo é elevada, caracterizando assim, a existência de uma grande quantidade de objetos presentes nele. Já a Figura 6b, exibe a S-T Matrix referente à taxa de incidência de dengue do *cluster* selecionado.



a) b)

Figura 6: a) SOM e b) S-T Matrix após a seleção do *cluster*.

Na Figura 7 tem-se a ilustração dos estados que formam o *cluster* selecionado. Vale salientar que, devido à ferramenta favorecer a análise temporal, logo a formação do *cluster* também se dá na trajetória do tempo.

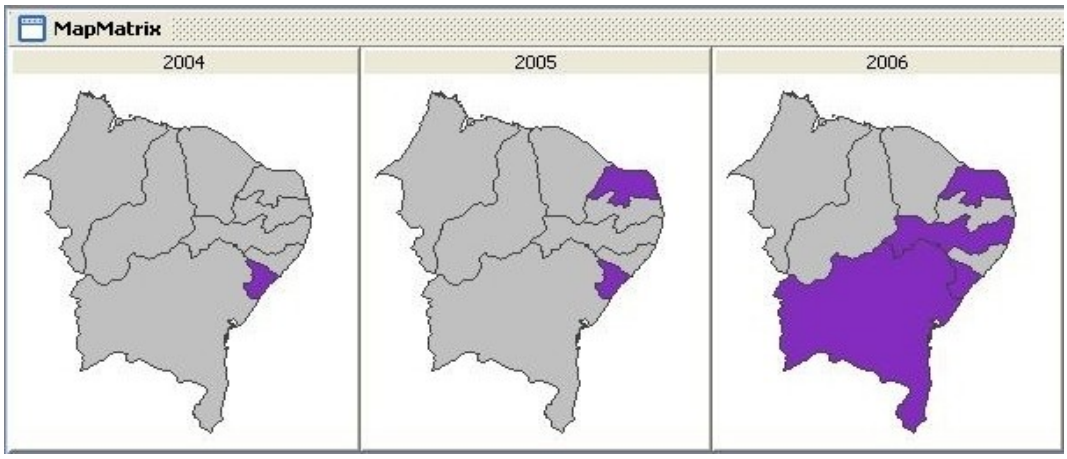


Figura 7: MapMatrix com os estados do *cluster* selecionado.

Por fim, na Figura 8 pode-se observar a escala min-max, do qual é útil quando se tem valores em diferentes eixos e estes são diretamente comparáveis, como por exemplo, com os valores percentuais utilizados nesta pesquisa. Desssa forma, podemos verificar que o (PIB_PC) possui uma forte relação para a ocorrência dos outros índices, dentre eles podemos destacar o (DENG).

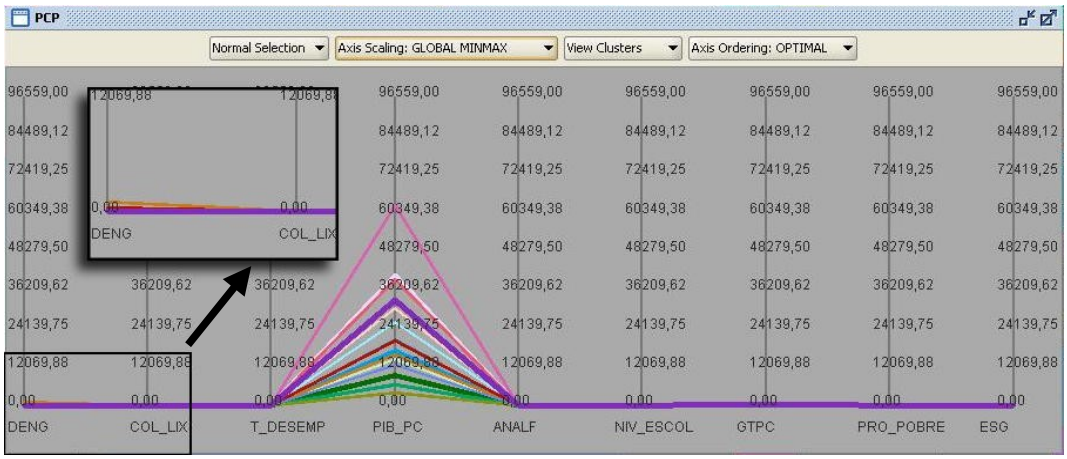


Figura 8: Escala Min-Max.

4. Conclusões

Com este artigo foi possível observar o quanto o auxílio de ferramentas, para apoio na análise de dados espaço-temporais, constitui-se de grande relevância para a pesquisa realizada. Além disso, notou-se que a utilização de ferramentas que permitem esse tipo de análise também proporciona um diferencial significante com relação aos resultados obtidos, pois isso permitiu minimizar o uso de técnicas manuais.

Com relação a análise realizada sobre os dados da região Nordeste observou-se que, com exceção do PIB *per capta*, não existe uma forte correlação entre os indicadores examinados com o índice da dengue, o qual constituía o objetivo principal de nossa pesquisa. Entretanto, com relação aos resultados obtidos, verificou-se que a utilização de métodos de *clustering* realmente constitui uma técnica eficaz na busca de padrões. Mediante a utilização desse método foi possível visualizar e analisar, com maior precisão, padrões que individualmente não apresentavam informações relevantes, mas que após serem direcionados para um *cluster* específico passaram a expressar informações importantes.

Esse estudo proporcionou visualizar e confirmar resultados dos quais, em pesquisas anteriores, haviam sido amplamente estudados (Fraga e Dias, 2007), (Lima, 2010). Por exemplo, a relação entre o nível de escolaridade e a taxa de desemprego, o que mais uma vez corrobora que a utilização desse tipo de ferramenta é deveras importante.

Agradecimentos

Os autores agradecem à CAPES e ao CNPq pela concessão das bolsas de pesquisa e pelo apoio financeiro para realização da mesma. Além disso, agradecem também ao Dr. Diansheng Guo e sua equipe por disponibilizarem gratuitamente a ferramenta.

Referências Bibliográficas

Fraga, G. J.; Dias, J. Taxa de desemprego e a escolaridade dos desempregados nos estados brasileiros: estimativas dinâmicas de dados em painéis. **Econ. Apl.**, Ribeirão Preto, v. 11, n. 3, Sept. 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-80502007000300005&lng=en&nrm=iso>. Acesso em: 11.nov.2010. doi: 10.1590/S1413-80502007000300005.

Guo, D; Chen, J.; MacEachren, A. M.; Liao, K. A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP). **IEEE Transactions on Visualization and Computer Graphics**, v. 12:6, p. 1461-1474, 2006.

Instituto Brasileiro de Geografia e Estatística (IBGE). Disponível em: <http://www.ibge.gov.br> Acesso em: 18.out.2010.

Instituto Brasileiro de Geografia e Estatística (IBGE) - Dados do Censo 2010 publicados no Diário Oficial da União do dia 04/11/2010. Disponível em : <http://www.censo2010.ibge.gov.br/dados_divulgados/index.php> Acesso em: 10.nov.2010.

Lima, F. A relação entre o nível de escolaridade e o mercado de trabalho em 2009. Tema em análise, Estatísticas do Emprego - 1º trimestre de 2010, 36-43. Disponível em: <https://dspace.ist.utl.pt/bitstream/2295/654098/2/2010_lima_ine_publicacao_1t2010.pdf> . Acesso em: 11.nov.2010.

Miller, H.; Han, J. **Geographic Data Mining and Knowledge Discovery**. CRC Press, 2009, 443 p.

Ministério da Saúde. Departamento de Informática do SUS (DATASUS). Disponível em: <http://www2.datasus.gov.br>. Acesso em: 19.out.2010.

Neves, M. C.; Freitas, C. C.; Câmara, G. (2001) Mineração de Dados em Grandes Bancos de Dados Geográficos. INPE. Relatório Técnico-CTBRASIL. Disponível em: <www.dpi.inpe.br/geopro/modelagem/relatorio_data_mining.pdf> Acesso em: 19.out.2010.