

SELEÇÃO DE ANOS REPRESENTATIVOS DE UMA BASE DE DADOS PLUVIOMÉTRICOS COM ATRIBUTOS ESPACIAL E TEMPORAL UTILIZANDO TÉCNICAS DE VISUALIZAÇÃO.

Rogério Galante Negri¹, Edilson Ferreira Flores, Milton Hirokazu Shimabukuro. – Inter-áreas - Licenciatura em Matemática – Departamento de Matemática, Estatística e Computação – Faculdade de Ciências e Tecnologia – Campus de Presidente Prudente.

Em estudos climatológicos são utilizados como objetos de estudo fenômenos climáticos que possuem intrinsecamente uma distribuição espacial e temporal. A Organização Mundial de Meteorologia orienta que sejam analisados, no mínimo, 30 anos da observação de um determinado fenômeno. Dentro desse período, busca-se a identificação de anos representativos, os quais são anos com comportamentos distintos e que são utilizados como representantes de determinadas classes de características semelhantes. Por exemplo, considerando um conjunto de dados que contenha registros de chuva, distribuída sobre determinada área, deseja-se verificar os anos padrão que caracterizem os tipos habitual, seco e chuvoso, e assim, identificar o melhor ano representativo para cada padrão.

Em virtude do longo período de coleta, os conjuntos de dados podem comportar volumes que trazem complexidade à sua manipulação e análise. Não só na Climatologia, como em diversas áreas de pesquisa, o acúmulo de dados é um fator que muitas vezes dificulta a extração de informação útil. Nesse contexto, ao observar os transtornos trazidos pela manipulação de grandes quantidades de dados, pesquisadores se concentraram no desenvolvimento de técnicas capazes de auxiliar o homem na manipulação desses conjuntos (Branco, 2003).

Técnicas para Mineração de Dados, ou DM (*Data Mining*), são capazes de fornecer suporte para perceber, ou descobrir, informações úteis que geralmente se apresentam ocultas em vastos conjuntos de dados. Objetivos comuns às diferentes aplicações de DM incluem a detecção, interpretação e previsão de padrões presente nos dados. Geralmente, técnicas de DM são utilizadas como sinônimo de Descoberta de Conhecimento em Base de Dados, ou KDD (*Knowledge Discovery in Database*), embora este último seja definido como um processo mais geral, no qual o DM está imerso.

Nesse contexto, uma ferramenta importante que pode ser utilizada no auxílio a Mineração de Dados é a Visualização, capaz de oferecer recursos para exploração visual de base de dados. A Visualização é um processo que realiza uma transformação do dado, da forma numérica ou textual para a forma visual, permitindo ao homem a sua observação gráfica, facilitando a percepção de característica ocultas nos dados, seja qual for a natureza dessa informação. Um sistema de Mineração Visual de Dados, ou VDM – *Visual Data Mining*, integra técnicas de Visualização e as para Mineração de Dados, buscando assim a combinação entre a habilidade de exploração da mente humana com o ágil processamento dos computadores, gerando uma ferramenta atraente e de eficiente utilidade.

Sendo assim, as técnicas de Visualização dentro da Climatologia podem ser utilizadas na determinação de anos representativos. Com esse tipo de técnica o processo de determinação de anos representativos pode ser melhorado pelo uso conjunto com técnicas analíticas, como algoritmos estatísticos.

Neste trabalho foi utilizado uma base de dados referente a precipitação pluviométrica distribuída pelo Estado de São Paulo no período de 1967 a 1997 e quantificado através de postos de coleta de dados (PCD). A origem desses dados é o Banco de Dados Pluviométricos do Estado de São Paulo, organizado pelo Departamento de Água e Energia Elétrica, DAEE. Sobre esta base de dados foram utilizadas três técnicas de Visualização, duas delas responsáveis pelo processo de identificação de anos representativos e uma terceira ligada ao tratamento dos dados.

Em um primeiro momento, foi utilizado o software VisualGeo (Negri & Flores, 2006) que fornece um conjunto de técnicas estatísticas tais como Histograma, gráfico da Distribuição Normal, Estatística Descritiva e Distribuição Espacial dos postos, que permitem ao usuário fazer uma análise exploratória sobre os dados. As figuras 1 e 2 ilustram tais resultados.

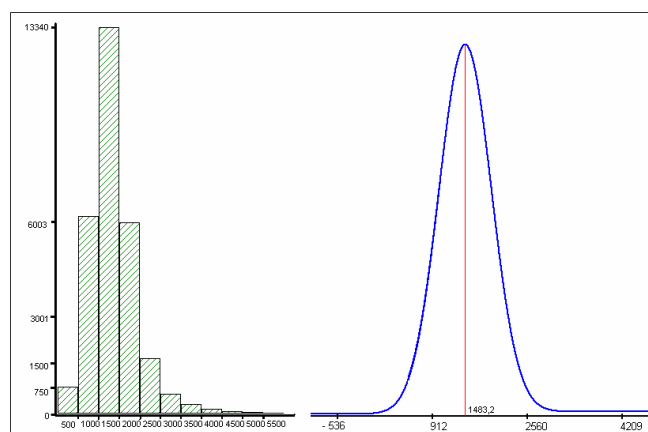


Figura 1: Análise Exploratória dos dados – Histograma e Distribuição Normal

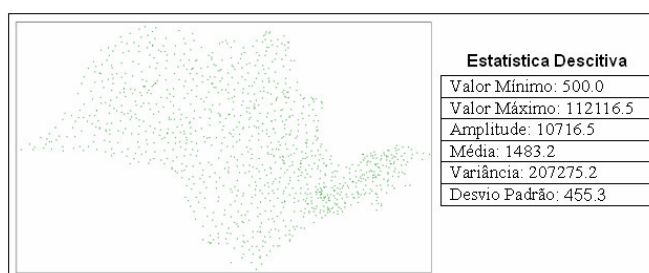


Figura 2: Análise Exploratória dos dados – Distribuição Espacial e Estatística Descritiva

A partir desses resultados foi possível detectar a forma geral da distribuição espacial, constatar a normalidade dos dados, assim como perceber a existência de dados discrepantes, pois foi encontrado como Valor Máximo a quantidade referente a 112.116,5 mm de Precipitação Pluviométrica na região de estudo no período de 1967 a 1997. Embora este dado esteja armazenado, sua ocorrência não se torna prejudicial no caso desse estudo. Com base nessas estatísticas foram definidas as classes de intervalos ilustradas na figura 3.

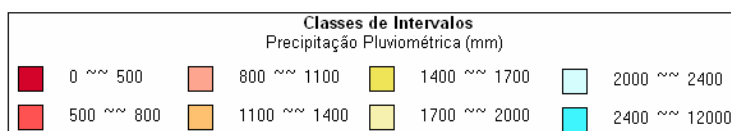


Figura 3: Classes de intervalos definidas com o uso do VisualGeo

Para a definição desses intervalos foi utilizado o gráfico da Distribuição Normal, reservando uma classe final que capturasse valores maiores que 2400 mm. Logo, de acordo esses intervalos o resultado gerado é o gráfico dado pela figura 4.

O gráfico ilustrado pela figura 4, denominado *GeoBar*, expõe a distribuição temporal e espacial do fenômeno natural Precipitação Pluviométrica. A informação contida neste gráfico é apoiada segundo duas variáveis, o Período, descrito pelo eixo horizontal e o Número de Ocorrências, descrito pelo eixo vertical. Além disso, dentro de cada barra, é expressa a distribuição espacial dos dados que fazem parte de um determinado intervalo em um determinado ano. Fundamentalmente, é necessário perceber que o número de ocorrências de cada barra está ligado diretamente a sua altura, pois a medida que aumenta o número de ocorrências em determinada classe, de determinado ano, maior deve ser sua variação, ou seja, maior deve ser sua altura, e ainda, maior deve ser a quantidade de localizações espaciais dentro dessa barra, pois a quantidade de ocorrências é igual a quantidade de pontos plotados dentro dessa barra. A visão mais detalhada do produto gerado (a direita do gráfico na figura 4), apresentando a variação dos anos no eixo horizontal, a distribuição espacial da informação e o número aproximado de ocorrências ao se referir a altura da barra. O número de ocorrências é encontrado tomando como base a escala escrita no eixo vertical e comprando-a com a altura da barra, e não se referindo unicamente a altura geral em que se encontra a barra. Os pontos que denotam a

posição de uma amostra podem assumir três tipos de coloração, verde claro, médio ou escuro, indicando subclasses dentro da classe definida anteriormente.

Considerando ainda o mesmo intervalo utilizado para aplicação da técnica *GeoBar*, e utilizando uma escala de cores próxima para as classes, foi gerado um outro gráfico (figura 5) com o uso do software GGobi. Neste gráfico, cada eixo horizontal representa um ano e cada ponto nele se refere a um posto de coleta. Ambos os gráficos permitem concluir que o ano de 1983 possui um comportamento distinto dos demais e pode ser considerado um ano excepcional chuvoso.

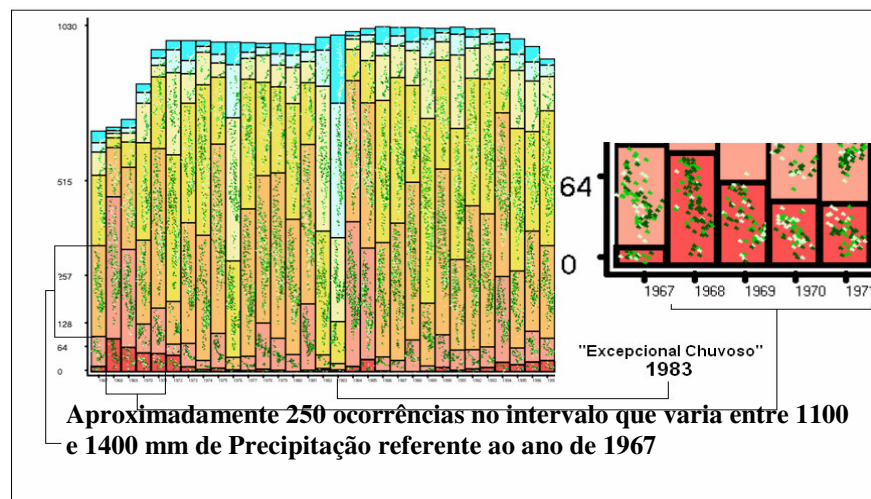


Figura 4: Produto gerado pela técnica *GeoBar*, aplicada ao período de 1967 a 1997, junto com uma visão detalhada do gráfico

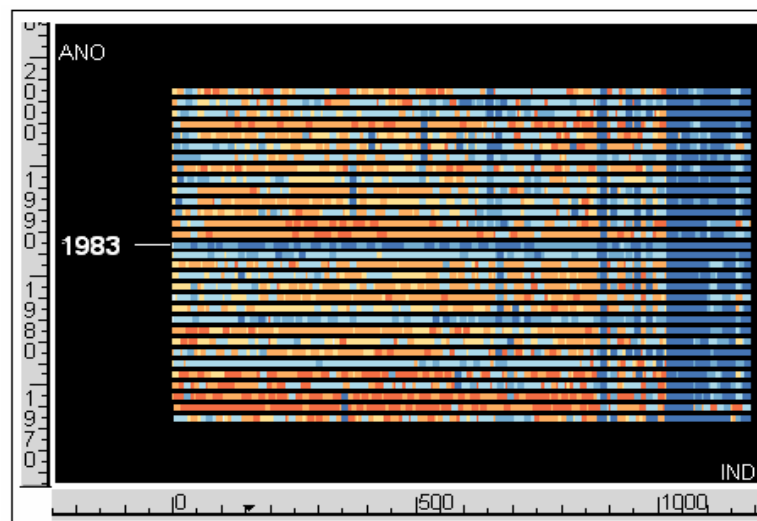


Figura 5: Representação gráfica gerada no GGobi

O próximo passo é tratar valores anômalos ou não existentes dos postos de coleta, substituindo tais valores com aproximações razoáveis. Para isso é necessária uma inspeção do conjunto de dados de um determinado posto, e de seus vizinhos caso necessário, com objetivo de inferir tais valores a partir de outros mais confiáveis. Esta tarefa é apoiada pela “Visualização Temporal Multi-Escala” (Shimabukuro et al, 2003) (figura 6), que permite apresentar os dados simultaneamente nas escalas diária, mensal e anual. Os dados de cada ano são apresentados em uma coluna com 25 quadros, sendo os 12 superiores usados para os dados de precipitação pluviométrica dia a dia, os 12 seguintes e o último quadro representam os totais mensais e anual, respectivamente.

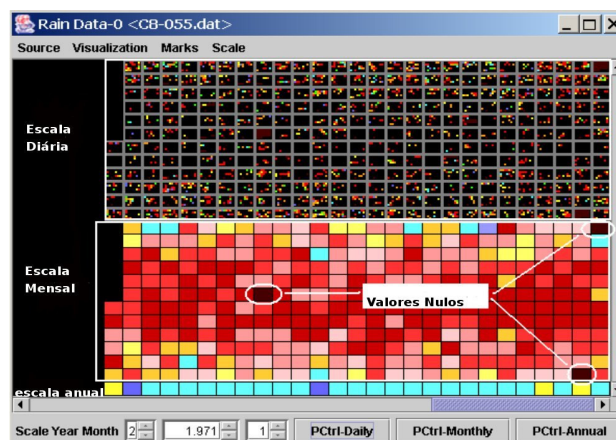


Figura 6: Esquema da Visualização Temporal Multi-Escala

Na base de dados estudada foi possível verificar os anos que apresentam uma característica “Chuvosa”, e ainda com característica “Excepcionalmente Chuvosa”, pois, a alta ocorrência de observações nas classes superiores e a baixa ou inexistente ocorrência nas classes inferiores determina essa caracterização. Ao contrario dos anos “Chuvosos” existem também os anos “Secos”, assim como os anos “Habituais” e com características intermediárias em relação aos “Secos” e “Chuvosos”.

Os resultados permitiram concluir que a utilização conjunta das técnicas aqui apresentadas, com o complemento das técnicas analíticas, oferece um ambiente poderoso para a tarefa de seleção de anos representativos e o tratamento de valores anômalos ou nulos. Desta forma, as técnicas possibilitaram a escolha de anos representativos bastante adequada e com confiabilidade no auxilio de estudos de caráter climatológico.

Referências

Branco, V.M.A. - Visualização como Suporte à Exploração de uma Base de dados Pluviométrica, Tese de Mestrado, USP – São Carlos (2003)

Negri, R.G.; Flores, E.F. - Visualgeio Visualização de Dados Geográficos: Técnica de Visualização de Dados Aplicada a Fenômenos Naturais com Atributos Espaciais e Temporais, I Simpósio de Matemática, FCT, Unesp Campus de Presidente Prudente, CD-ROM (2006)

Shimabukuro, M.H; Branco, V.M.A; Oliveira, M.C.F; Flores, E.F. - Visual Exploration of Spatio-Temporal Database. In GEOINFO 2003 V Simpósio Brasileiro de Geoinformática, 2003, Campos do Jordão – SP (2003)