RBRH – Revista Brasileira de Recursos Hídricos Volume 9 n.1 Jan/Mar 2004, 181-187

Preparação de Dados de Chuvas Intensas Utilizando Técnicas de Mineração de Dados

Fábio Teodoro de Souza, Nelson Francisco Favilla Ebecken
COPPE / UFRJ – fabio@coc.ufrj.br, nelson@ntt.ufrj.br

Recebido: 07/03/03 – revisado: 15/09/03 – aceito: 15/01/04

Resumo

As ferramentas baseadas em princípios de inteligência artificial, tais como em mineração de dados, aliadas aos Sistemas de Informações Geográficas (SIG), são componentes de grande utilidade nos estudos ambientais. A prévia preparação dos dados é necessária no processo de modelagem, pois o ajuste da base de dados permite que as informações contidas sejam expostas para as ferramentas de extração de conhecimento. Pretende-se nesse trabalho, apresentar a metodologia adotada para o preenchimento das falhas nos dados de chuvas intensas, coletados em intervalos de 15 minutos, da rede pluviométrica do sistema de alerta da cidade do Rio de Janeiro.

Palavras-Chave – mineração de dados, regionalização, predição.

INTRODUÇÃO

Os dados são uma coleção de observações de eventos, que acontecem em função de outros, e que possuem uma relação com o mundo real donde ele foi coletado. A manipulação dos dados promove o entendimento da natureza da qual ele foi medido. Contudo, antes de se aplicar qualquer método de análise nos dados, é preciso prepará-los previamente corrigindo qualquer inconsistência presente. A preparação dos dados é a parte mais importante de qualquer projeto, Pyle (1999). A acurácia dos modelos produzidos e o processo de tomada de decisão dependem da qualidade da preparação dos dados.

No estudo de escorregamentos de encostas da cidade do Rio de Janeiro, pretende-se desenvolver um modelo para a predição desses acidentes geotécnicos, relacionando aos dados de ocupação do solo e aos padrões acumulados de chuva.

Os índices de chuva acumulada devem ser calculados com dados do pluviômetro mais próximo ao bairro onde ocorreu o escorregamento.

Contudo, o banco de dados de chuva apresenta-se com registros ausentes. Para que se possa alcançar uma melhor estimativa dos padrões de chuva acumulada causadores dos escorregamentos, se faz necessário preencher essas falhas.

A predição das falhas pode ser obtida pela simulação do algoritmo de redes neurais artificiais (RNA´s), e os dados usados para o treinamento, teste e verificação das RNA´s, devem ser agrupados com técnicas de redução de dimensionalidade (regionalização da chuva).

Na literatura de mineração de dados é difícil encontrar estudos para o preenchimento de falhas de dados, na freqüência de coleta do período considerado neste trabalho (em intervalo de 15 minutos), e este fato motiva à implementação de uma metodologia para a realização de tal tarefa.

As técnicas em *mineração de dados* permitem a extração de conhecimento dos dados de forma automática, e são menos dependentes da subjetividade, se comparadas aos estudos de chuvas intensas encontrados na literatura de hidrologia.

DADOS DISPONÍVEIS.

Estudos têm sido mostrados, Menezes et al. (2000), que a cidade do Rio de Janeiro possui características físicas (posição geográfica e topografia) e padrões atmosféricos favoráveis ao desenvolvimento de fenômenos meteorológicos causadores de fortes chuvas.

O banco de dados de chuva é composto por registros de precipitação, a intervalos de 15 minutos, coletados em 30 pluviômetros automáticos instalados no município. Esta rede de pluviômetros compõe o sistema de alerta que também dispõe de análise de sondagens atmosféricas e de imagens de radar e satélite. Foram selecionados 20 eventos ou períodos chuvosos, durante os quais ocorreram escorregamentos nos anos de 1998 a 2000.

A figura a seguir ilustra os polígonos de Thiessen e o mapa de bairros do Rio de Janeiro.

Pluviômetros

1 - Vidigal	16 – Jardim Botânico
2 – Urca	17 – Itanhangá
3 – São Conrado	18 – Cidade de Deus
4 – Tijuca	19 – RioCentro
5 – Santa Tereza	20 – Guaratiba
6 – Copacabana	21 – Gericinó
7 – Grajaú	22 – Santa Cruz
8 – Ilha do Gov.	23 – Cachambi
9 – Penha	24 – Anchieta
10 – Madureira	25 – Grota Funda
11 – Irajá	26 – Campo Grande
12 – Bangu	27 – Sepetiba
13 – Piedade	28 – Sumaré
14 – Tanque	29 – Mendanha
15 - Saúde	30 – Itaúna

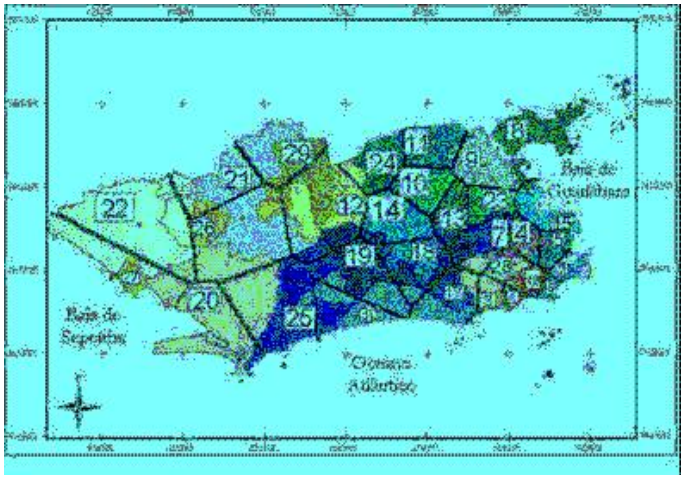

Figura 1 – Mapa da rede de pluviômetros automáticos e mapa de bairros da cidade do Rio de Janeiro.

A Secretaria do Meio Ambiente monitora os parâmetros do solo - tais como qualidade, mineralogia, vulnerabilidade, uso – e as taxas de cada taxonomia são calculadas para cada um dos 159 bairros existentes no município. A configuração de cores da figura 1 ilustra os bairros pertencentes aos polígonos de Thiessen.

METODOLOGIA

A figura a seguir ilustra a metodologia utilizada na substituição dos valores ausentes.

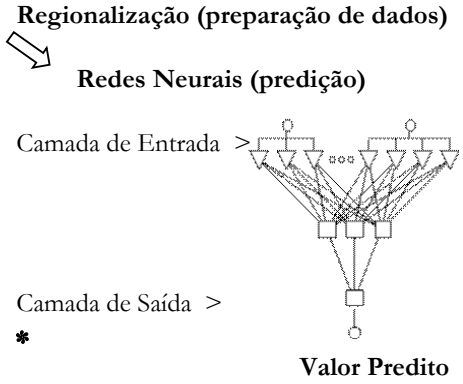

Figura 2 – Metodologia de substituição de falhas.

Uma vez identificado o pluviômetro com dado ausente (camada de *saída* ou variável de *predição* das *RNA´s*), então é necessário utilizar as técnicas de regionalização, que servem para selecionar os pluviômetros que farão parte da camada de entrada (variáveis de entrada das *RNA´s*).

A regionalização dos dados é necessária para identificar os padrões espaciais da chuva e permitir que o treinamento da *RNA* seja realizado com dados de pluviômetros próximos daquele pluviômetro com falhas. Ou seja, se o treinamento for realizado com todos os dados possíveis, tende-se a adicionar ruído proveniente de dados de pluviômetros distantes, uma vez que existe uma grande variabilidade espaço-temporal da chuva no município do Rio de Janeiro. Além disso, a regionalização configura um número reduzido de variáveis para o treinamento da *RNA*, reduzindo o esforço computacional.

Neste trabalho foram utilizadas quatro abordagens diferentes: *Análise de Componente Principal* (*ACP*), *Correlação*, *Árvore de Agrupamento* e o *Método de Partição k-means*.

A *Análise de Componente Principal* pode ser vista como um método de redução de dados, a partir da associação de duas ou mais variáveis correlacionadas dentro de um fator. Por exemplo, pode-se considerar um gráfico em que cada variável é representada por um ponto. Neste gráfico podem-se girar os eixos em qualquer direção sem mudar as posições relativas dos pontos uns aos outros, porém, com mudança das coordenadas atuais dos pontos; ou seja, a simples rotação dos eixos mudaria naturalmente a carga dos fatores. O objetivo da estratégia de rotação, elaborada e popularizada em discussões detalhadas por Wherry (1984), é o de obter um padrão de interpretação mais fácil e claro, através da associação de fatores com

altas cargas para algumas variáveis e com baixas cargas para outras

As medidas de correlação extraídas da *Matriz* de *Auto-Correlação* expressam uma medida da relação entre duas ou mais variável. Os coeficientes de correlação podem variar de -1 a +1. Os valores -1, +1 e 0 representam uma correlação negativa perfeita, positiva perfeita, e ausência de correlação, respectivamente.

Os *agrupamentos de dados* ou *clustering* são técnicas em mineração de dados, que consistem em agrupar os dados dentro de classes ou '*clusters*' tal que os objetos dentro de uma classe tenham alta similaridade em comparação com um outro objeto dessa classe, Kamber (2001), mas têm baixa similaridade a objetos de outras classes.

A *árvore* de *agrupamento* (ou árvore hierárquica) usa as dissimilaridades ou distâncias entre os objetos para formar as classes. O cálculo das distâncias Euclidianas é o método mais direto de calcular as distâncias entre os objetos num espaço multidimensional, Kamber (2001).

Se os dados contêm uma clara "estrutura" em termos de classes de objetos (similares uns aos outros), então esta estrutura muitas vezes é refletida na árvore hierárquica como "galhos" distintos. A figura a seguir ilustra uma árvore hierárquica construída com os dados medidos dos 30 pluviômetros durante o período de 31 de dezembro de 1998 a 13 de janeiro de 1999 (1º evento).

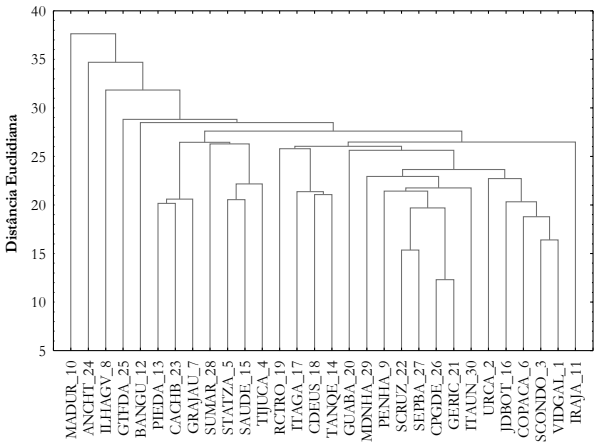

Figura 3 – Árvore Hierárquica construída com dados de chuva do município do Rio de Janeiro.

Outro método de agrupamento descrito por Kamber (2001), é o algoritmo *k-means*, que divide um conjunto de *n* objetos dentro de *k* classes, e, baseado na atualização do valor médio dos objetos de cada classe, o algoritmo re-classifica cada objeto para a classe da qual o objeto é mais similar, num processo iterativo até que haja convergência de uma função de critério.

Uma vez agrupados os pluviômetros segundo as técnicas de regionalização citadas, os bancos de dados de chuva podem ser preparados para o treinamento, teste e predição dos valores ausentes através das *RNA´s*.

As *RNA´s* de múltiplas camadas alimentadas adiante, basicamente consistem de um conjunto de unidades sensoriais, ou nós de fonte que constituem a *camada de entrada*, uma ou mais *camadas oculta* (de nós computacionais) e uma *camada de saída* (de nós computacionais), (Haikin, 2001). O sinal de entrada se propaga para frente através da rede, camada por camada. Estas *RNA´s* são chamadas de *Perceptrons de Múltiplas Camadas* (*PMC´s*), as quais representam uma generalização do *perceptron de camada única.* (Rosenblatt, 1958).

O treinamento dos *PMC´s* é realizado de forma supervisionada com o algoritmo de retro-propagação do erro, que é baseado na regra de aprendizagem por correção do erro e desenvolve-se em dois passos através das diferentes camadas da rede: um passo para frente, a propagação, e um passo para trás, a retro-propagação. Durante a propagação, um padrão de atividade (vetor de entrada) é aplicado aos nós sensoriais da rede e seu efeito se propaga através da rede, camada por camada. Finalmente, um conjunto de saídas é produzido como a resposta real da rede.

No passo de propagação, os pesos sinápticos são todos fixos. Durante a retro-propagação, os pesos sinápticos são todos ajustados de acordo com uma regra de correção de erro. Especificamente, a resposta real da rede é subtraída de uma resposta desejada (alvo) para produzir um sinal de erro, que é então propagado para trás através da rede, contra a direção das conexões sinápticas. Os pesos sinápticos são ajustados para fazer com que a resposta real da rede se mova para mais perto da resposta desejada, em um sentido estatístico.

É através da habilidade de aprender através do treinamento, que os *PMC´s* têm sido aplicado com sucesso para resolver diversos problemas difíceis, e também podem ser usados para a predição de falhas em banco de dados de chuva.

Os dados de chuva podem ser vistos como séries temporais, onde a variável precipitação é medida sobre um período de tempo. Também se espera que existam relações entre os valores de precipitação nos diversos tempos sucessivos. Portanto, a predição do valor da falha num dado tempo deve ser feita com informações espaço-temporal, a partir de um número de valores precedentes da própria variável (pluviômetro com falha) e de outras variáveis (pluviômetros agrupados por regionalização).

RESULTADOS E DISCUSSÕES

A figura a seguir ilustra o mapa do município e os polígonos de Thiessen com dados ausentes no 1º even-

to, além dos bairros atingidos por escorregamentos (contorno em amarelo).

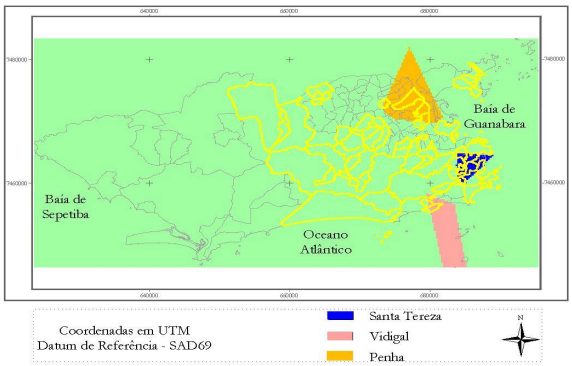

Figura 4 – Pluviômetros com falhas durante o 1º evento.

Descreve-se o procedimento de regionalização adotado para o pluviômetro instalado em Santa Tereza (polígono azul na figura 4), que foi escolhido como exemplo, por apresentar valores de precipitação elevados, pico de 26,8 mm/15min, conforme figura 9.

Na abordagem que considera a *ACP*, foram agrupados os pluviômetros com o mesmo fator do pluviômetro de Santa Tereza (inclusive). A figura a seguir ilustra os polígonos de Thiessen em vermelho, referentes aos pluviômetros selecionados no estudo de *ACP*.

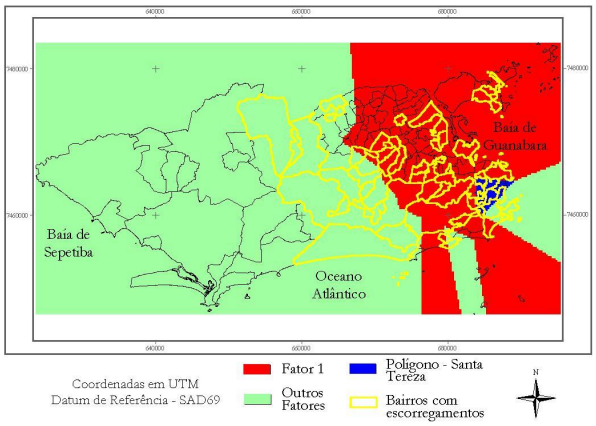

Figura 5 - Pluviômetros selecionados com o critério de ACP.

No critério de correlação foram agrupados os pluviômetros cujas medidas de correlação fossem maiores ou iguais a 0,70 ($c \geq 0{,}70$), se comparados com o pluviômetro de Santa Tereza (inclusive).

A figura a seguir ilustra os polígonos de Thiessen em vermelho selecionados no estudo de correlação.

Para o critério de árvore de agrupamento foram agrupados os pluviômetros pertencentes ao mesmo "galho" da árvore que possui o pluviômetro de Santa Tereza (inclusive).

Esta abordagem foi possível neste exemplo, devido à clara estrutura que a árvore apresenta, conforme se pode observar na figura 3.

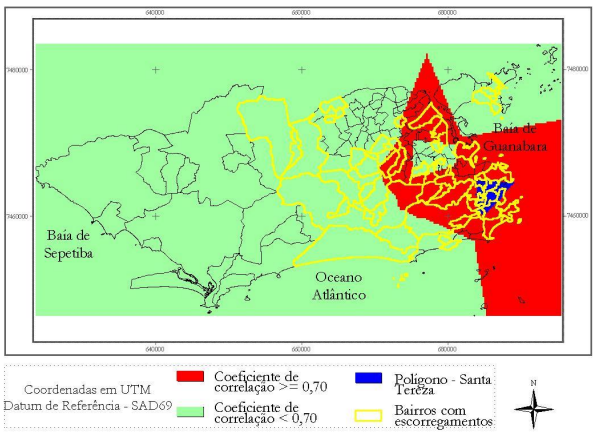

Figura 6 - Pluviômetros agrupados com correlação.

A figura a seguir ilustra os polígonos de Thiessen (em vermelho), referentes aos pluviômetros selecionados no estudo de árvore de agrupamento.

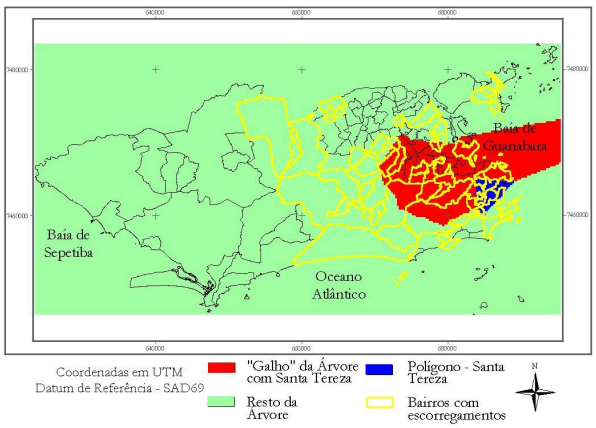

Figura 7 - Pluviômetros da árvore de agrupamento.

Para o método de agrupamento *k-means* foram agrupados os pluviômetros pertencentes à mesma classe que possui o pluviômetro de Santa Tereza (inclusive). A figura a seguir ilustra os polígonos de Thiessen, (em vermelho) referentes aos pluviômetros selecionados no segundo o método de *k-means*.

A escolha do método de regionalização deve ser feita em função do melhor resultado de predição das *RNA's*. Se em nenhuma das abordagens citadas forem

conseguidos bons resultados de predição, então se deve optar por outros métodos de regionalização, além do mero critério de proximidade geográfica (pluviômetros vizinhos daquele com falhas). Porém, todo método de regionalização deve considerar a importância da variabilidade espacial da chuva.

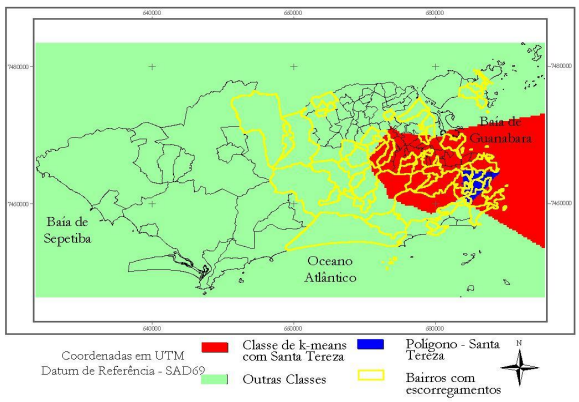

Figura 8 - Pluviômetros selecionados com critério *k-means*.

Depois de terminada a etapa de regionalização então podem ser realizadas as simulações com as *RNA´s*, visando a predição dos valores ausentes.

Para o preenchimento das falhas com *RNA´s* são selecionados somente aqueles intervalos com falhas em que se observa valor de precipitação diferente de zero em pelo menos um pluviômetro da rede, e em pelo menos duas horas antes ou depois da falha. Pois diferente disso, estas falhas também podem ser preenchidas com zero.

Para o treinamento das redes neurais, devem ser descartados os registros cuja soma de precipitação de todos os pluviômetros for igual a zero (não chove no município). Este tipo de registro dificulta o aprendizado da rede, além de consumir esforço computacional, pois muitas vezes abrangem mais de 90% do banco de dados de chuva.

As simulações com as *RNA´s* foram realizadas com os bancos de dados dos 20 eventos chuvosos, pois todo evento apresentava falha em pelo menos um pluviômetro. Foi utilizada a rede do tipo *PMC´s* séries temporais (3 camadas).

Devido à escassez de dados, foi preciso separar grande parte do banco de dados para o treinamento (mais de 90% dos registros algumas vezes), para que a rede obtivesse boa performance preditiva.

A tabela a seguir apresenta um resumo estatístico da parcela de validação, com resultados de dois parâmetros determinantes na escolha da melhor predição.

Esta simulação teve na camada de saída o pluviômetro com falhas (Santa Tereza no 1º evento).

A razão de desvio padrão é o quociente entre o desvio padrão dos erros e o desvio padrão dos dados medidos, e quanto menor for seu valor, indica uma melhor predição.

Tabela 1 – Resumo estatístico das predições (verificação).

Métodos	Razão de Desvio Padrão	Correlação de Pearson-R
ACP	**0,34**	**0,94**
Correlação	**0,34**	**0,95**
Árvore	0,55	0,83
k-means	0,48	0,89

O coeficiente de correlação de Pearson-R é uma métrica usada para avaliar a relação entre os valores preditos e os dados, e quanto mais próximo do valor 1, indica uma melhor predição.

Pode-se observar na tabela 1 que as melhores predições foram obtidas das regionalizações com critérios de *Correlação* e *ACP*, pois estes métodos foram o que apresentaram menores valores da razão de desvio padrão e os maiores valores do coeficiente de correlação de Pearson-R.

A figura 9 a seguir ilustra o resultado da predição (simulação com as *RNA´s*) em perspectiva com os dados medidos, segundo o método de regionalização com medida de correlação.

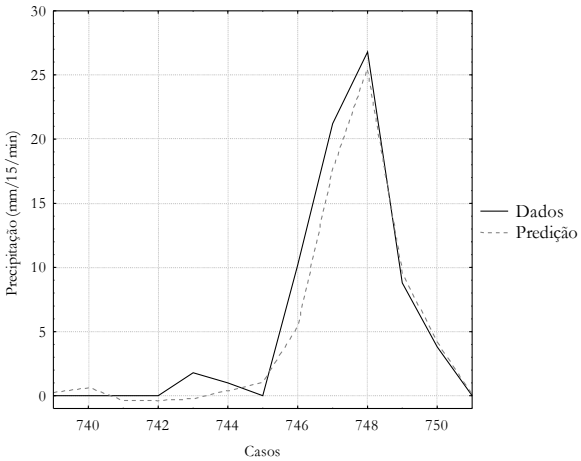

Figura 9 - Resultado da predição (correlação).

Pode-se observar na figura 9 que a linha de predição (pontilhada) apresenta boa aderência com a linha dos dados medidos (linha cheia).

A linha de predição consegue aproximar os valores máximos registrados e as variações bruscas no tempo com resultados satisfatórios.

A figura 10 a seguir ilustra os erros absolutos e médios quadrático, resultados da predição conforme figura 9.

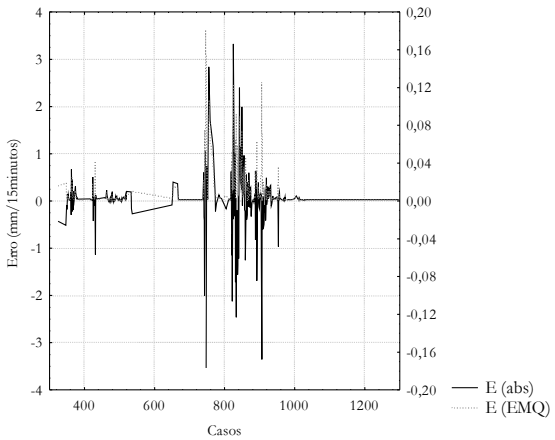

Figura 10 - Erros da predição (metodologia de correlação).

Pode-se observar na figura 10 que os erros absolutos (linha cheia) ficaram muito próximos do valor zero (eixo da esquerda) durante quase toda a dimensão da amostra, e quando a linha se deslocou do eixo, os valores dos erros nunca ultrapassaram o intervalo de -4,0 a +4,0 mm/15min.

Os erros médios quadráticos (linha pontilhada) também ficaram muito próximos do valor zero (eixo da direita) durante quase toda a dimensão da amostra, e quando a linha se deslocou do eixo, os valores dos erros nunca ultrapassaram o intervalo de -0,2 a +0,2.

A escolha do critério de regionalização depende, portanto, da análise dos vários resultados obtidos:

- Resultados estatísticos e erros;
- Aderência da linha de predição aos dados (identificação dos valores máximos e variações bruscas no tempo).

Estes resultados indicam que o critério de correlação ($c \geq 0{,}70$) pode ser adotado para o preenchimento da falha do pluviômetro de Santa Tereza.

Quando o intervalo de preenchimento da falha ultrapassar dois valores (maior que meia hora), então se deve realizar um estudo adicional considerando a "vizinhança". A figura a seguir ilustra um exemplo de comparação da predição da *RNA* e os dados de pluviômetros vizinhos:

Neste exemplo da figura 11, os dados medidos (em azul) durante o 20º evento chuvoso são do pluviômetro instalado na Tijuca (número 4 da Figura 1). A predição da rede (em vermelho) apresenta coerência quando comparada com os dados dos pluviômetros vizinhos, e, portanto, pode ser adotada para a substituição das falhas.

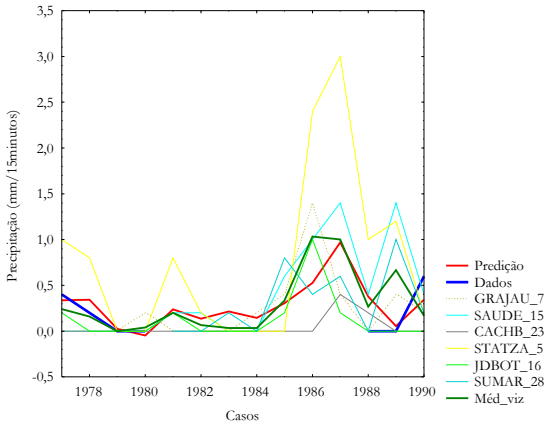

Figura 11 – Estudo regional - pluviômetros vizinhos.

A tabela a seguir resume os quantitativos dos critérios de regionalização adotados como eficientes para a substituição das falhas:

Tabela 2 – Resumo da eficiência dos métodos adotados.

Métodos	Boa Performance (%)	Pobre Performance (%)
ACP	**67,6**	25,0
Correlação	**13,3**	20,0
Árvore	10,5	35,0
k-means	7,6	15,0

Conforme ilustra a tabela 2, o método de *ACP* foi o que alcançou mais vezes as melhores predições (considerando os *resultados estatísticos*, *aderências* da linha de predição aos dados *e erros*), portanto, o método que foi mais utilizado durante o preenchimento das falhas.

CONCLUSÕES

O estudo de preenchimento das falhas nos dados de chuvas intensas desenvolvido com as técnicas em mineração de dados, aliadas aos SIG's, proporcionou resultados satisfatórios quanto à qualidade de predição, descritos pelos bons resultados estatísticos obtidos nas simulações com *RNA's*.

Durante o preenchimento das falhas, deve-se priorizar os métodos de regionalização por *ACP* e *Matriz de Auto-Correlação*, pois estes foram os métodos que forneceram melhores resultados, e depois tentar os métodos, tais como *Árvore de Agrupamento*, *k-means*, além de outros.

A regionalização da chuva também mostrou uma descrição abrangente e clara da relação espacial da pluviometria e as áreas atingidas por escorregamentos.

REFERÊNCIAS

HAIKIN, S., 2001. *Redes Neurais – Princípios e Prática*, 2.ed., Porto Alegre, Bookman, pp. 183-281.

KAMBER, H., 2001. *Data Mining - Concepts and Techniques* – Chapter 7 and 8, pp. 279-393.

MENEZES, W. F., Paiva, L. M. S, Silva, M. G. A. J. e Belassiano, M., 2000. *Estudo do Ambiente Favorável à Propagação de Sistemas Convectivos de Mesoescala sobre o Município do Rio de Janeiro*, XI Congresso Brasileiro de Meteorologia, Rio de Janeiro, pp. 1635-1645.

PYLE, D., 1999. *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, Inc., San Francisco, California, pp. 9-43.

ROSENBLATT, F., 1958. *The Perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review, vol. 65, pp. 386-408.

WHERRY, R. J. (1984). *Contributions to correlational analysis*. New York: Academic Press, pp. 300-301.

Preparing Data on Intense Rainfall using Data Mining Techniques

ABSTRACT

Data mining tools, together with Geographical Information Systems (GIS), are very useful components in environmental studies. In this modeling process it is necessary to prepare the data in advance because adjusting the databases allows the information contained to be exposed to knowledge extraction tools.

This study intends to show the methodology adopted to supply the missing values in the rainfall data collected at 15-minute intervals from the rain gauge network of the Warning System in the city of Rio de Janeiro.

Key Words – data mining, regionalization, prediction