ISSN:1984-2295

Revista Brasileira de Geografia Física

Homepage: www.ufpe.br/rbgfe

Avaliação de modelos regressivo logístico e baseado em rede neural para previsão da probabilidade de ocorrência de alagamentos em Curitiba-PR[1]

Marciel Lohmann[2], Leonardo José Cordeiro Santos[3] Camila Cumico4

[2] Pesquisador e Professor Doutor na Universidade Estadual de Londrina, Departamento de Geociências. marciel_lohmann@yahoo.com.br (autor correspondente). [3] Pesquisador e Professor Doutor da Universidade Federal do Paraná, Departamento de Geografia. santos.ufpr@gmail.com. 4Pesq. UFPB.

Artigo recebido em 15/08/2016 e aceito em 27/09/2016

R E S U M O

O presente artigo tem como objetivo avaliar os modelos regressivo logístico e baseado em rede neural para previsão da probabilidade de ocorrência de alagamentos em Curitiba-PR, utilizando como base a integração de informações hidrometeorológicas. Para a construção dos modelos foram utilizados os dados de precipitação estimada a partir da integração das informações provenientes de radar meteorológico, satélite e pluviômetros, utilizando o método de Análise Objetiva Estatística (ANOBES). A partir dos dados de estimativas de precipitação foi calculada a chuva média acumulada de 6 em 6 horas, utilizando-se do método de Thiessen e do Inverso da Distância ao Quadrado, sendo os dois métodos comparados para verificar qual possui o melhor resultado para a geração dos dados de entrada dos modelos. Em relação ao desempenho dos dois métodos utilizados na construção dos modelos, verificou-se no caso estudado que o SOM (Self Organizing Map) apresentou desempenho superior quando comparado com a regressão logística tanto no período de calibração quanto de verificação.

Palavras-chave: chuva crítica, inteligência artificial, curva ROC.

Evaluation of logistic regression and neural network models for probabilistic forecasts of flooding in Curitiba-PR

A B S T R A C T

This paper aims to evaluate the logistic regression and neural network models for probabilistic forecasts of flooding in Curitiba-PR using as a basis the integration of hydrometeorological information. For the construction of the models were used rainfall data estimated from the integration of meteorological radar, satellite and rain gauges data, using the analysis for statistical purposes (ANOBES) method. Rainfall estimates were used to calculate cumulative average rain of 6 hours, using the method of Thiessen and Squared Inverse Distance. These the two methods were compared to see which has better results for data generation to be used as models' data input. Regarding the performance of the two methods used to construct the models, it was found that the SOM (Self Organizing Map) has superior performance when compared with the logistic regression, either for calibration and verification.

Keywords: critical rain, artificial intelligence, ROC curve.

Introdução

Os fenômenos atmosféricos despertaram o interesse e a curiosidade do homem desde as civilizações antigas, que consideravam estes fenômenos obra da força divina. A história do ajustamento do homem às condições do meio e da transformação destes por suas atividades tem sido uma relação de conflito e harmonia, mas durante muitos séculos tais condições se mantiveram dentro dos limites sem causar danos significativos, pelo menos até o início do período da Revolução industrial (Brandão, 2004).

De acordo com o mesmo autor, embora pesem, favoravelmente, o grande avanço tecnológico atual e os esforços para o conhecimento das forças da natureza, a sociedade moderna permanece ainda bastante vulnerável diante dos eventos naturais extremos, particularmente os de natureza meteorológica.

[1] Trabalho com base na tese de doutorado defendida na UFPR em 2012.

Sob o rótulo genérico de eventos naturais extremos encontra-se uma gama de fenômenos, variada em quantidade e complexa em intensidade. A grande maioria dos mais frequentes e intensos desses eventos está ligada, direta ou indiretamente, à atmosfera: enchentes, secas, nevoeiros, geadas, granizos, descargas elétricas, nevascas, tornados, ondas de calor, ciclones tropicais e vendavais, complementados por desmoronamentos de vertentes e ressacas, acrescidos por impactos pluviais concentrados (White, 1974 apud Monteiro, 1991).

A relação entre desenvolvimento urbano e a qualidade e quantidade dos recursos hídricos em uma determinada bacia hidrográfica, por exemplo, constitui atualmente um dos principais problemas de gestão ambiental em nosso país. Nesse contexto, fica cada vez mais evidente a questão das inundações urbanas e a crescente vulnerabilidade da população à sua ocorrência.

Segundo Tucci (2005), as inundações urbanas são devido a dois processos, que podem ocorrer isoladamente ou combinados: (1) inundações ribeirinhas, associadas ao extravasamento das águas dos rios, quando o escoamento pluvial excede a capacidade de seu leito principal e (2) inundações devido à urbanização, que ocorrem na drenagem urbana devido ao efeito de impermeabilização do solo, canalização ou obstruções ao escoamento. O primeiro tipo ocorre geralmente em bacias médias ou grandes e tem seu impacto associado à ocupação das áreas de risco pela população. O segundo envolve frequentemente bacias pequenas, sendo fortemente influenciado pela variabilidade das precipitações. Neste trabalho, utilizar-se-á o termo alagamento como sinônimo de inundações devido a urbanização.

Lohmann (2011 e 2013), estudando os dados de ocorrências de Defesa Civil de Curitiba, comenta que os alagamentos de certa forma acompanham o processo de expansão urbana de Curitiba, ou seja, têm aumentado concomitantemente com a incorporação de novos espaços ocupados, sobretudo em áreas consideradas de risco. Em outros estudos Zanella (2005) e Deschamps (2004) têm demonstrado a falta de sincronia entre as ações antrópicas e as leis da natureza.

O mesmo autor, traçando um paralelo entre o número total de ocorrências de Defesa Civil registradas (queda de árvores, incêndios, deslizamentos, erosão entre outras) e apenas as que dizem respeito aos alagamentos, mensurou que do total de ocorrências entre os anos de 2005 e 2010, 44,5% são de alagamentos, sendo, portanto o principal e maior problema enfrentado pela Defesa Civil municipal, já que perante o total de ocorrências registradas para todo o período analisado, praticamente 45% estão relacionadas aos alagamentos.

A ocorrência desse fenômeno impõe a cidade e à população atingida um conjunto de impactos físicos, humanos e econômicos, que tem sido discutido por Wind et al. (1999), Queensland (2002), Merz et al. (2004), Tucci (2005), Thicken et al. (2005) e Machado et al. (2007).

Na tentativa de mitigar tais impactos, é de suma importância a obtenção de dados hidrometeorológicos históricos, atuais e futuros para que se possa investigar os locais mais atingidos e mais prováveis de serem atingidos por algum evento natural extremo. No entanto, as informações hidrometeorológicas em geral, não são capazes de expressar, com toda a extensão e fidelidade, o estado atual e o comportamento futuro da atmosfera e dos recursos hídricos (Yevjevich, 1974; Wilks, 1995; WMO, 2007), já que as informações climatológicas e de monitoramento incluem as incertezas provenientes dos processos imperfeitos de mensuração e estimação, enquanto a previsões, baseadas em modelos preditivos de natureza objetiva ou subjetiva, incorporam as incertezas de propagação de erros, de compreensão e de representação dos fenômenos naturais (Leite, 2008).

Especificamente para modelos preditivos e que geram, portanto, previsões probabilísticas é a consideração de que a probabilidade preditiva evolui no tempo, conforme a interação da capacidade preditiva dos modelos, calibração dos parâmetros e comportamento combinado dos dados de entrada. Essa atualização dinâmica das incertezas pode agregar vantagens em relação ao caso anterior e contribuir com ganhos adicionais no resultado, conforme a sensibilidade do processo decisório.

Sendo assim, as vantagens da previsão probabilística sobre a determinística têm sido intensamente discutidas nos últimos tempos. O entendimento consensual é o de que as previsões probabilísticas são mais completas, pois expressam as incertezas e permitem a derivação da melhor estimativa sob o ponto de vista do usuário, ao contrário das previsões determinísticas, que são incompletas e só expressam a melhor estimativa sob o ponto de vista do previsor (Krzysztofowicz,1998 e 2001).

Ainda de acordo com os mesmos autores, as previsões probabilísticas são apontadas como cientificamente mais honestas, pois admitem e expressam as incertezas; elas permitem a emissão de avisos e alertas baseados em riscos; elas favorecem a tomada de decisão racional do

usuário; elas agregam potencial para benefícios econômicos adicionais para toda a sociedade; e evitam mal entendidos nas atribuições de responsabilidade entre previsor e tomador de decisão, desacoplando os processos de previsão e uso das informações (Krzysztofowicz,1998 e 2001).

Levando em consideração tais ideias, o objetivo deste trabalho foi avaliar dois modelos preditivos (regressão logística e rede neural artificial SOM) para a previsão da probabilidade de ocorrência de alagamentos em Curitiba-PR. Tais modelos podem servir, em uma fase posterior, como base para um sistema de alertas para o município em questão.

Material e métodos

Área de estudo

Como área de estudo foi escolhido o município de Curitiba (Figura 1). A escolha desta área justifica-se na medida em que grandes transformações socioeconômicas e ambientais foram observadas ao longo das últimas décadas deflagrando fortes pressões e modificações sobre o ambiente, muitas das quais se traduziram em perdas irreparáveis para esta área.

Curitiba é a capital do Estado do Paraná e localiza-se na Região Sul do Brasil. Fundada em 1693, ocupa uma área de 432,17 km². É também a cidade pólo da Região Metropolitana composta por 26 municípios que, juntos, ocupam uma área de 15.622,33 km².

Localiza-se no Primeiro Planalto Paranaense, o qual foi descrito por Reinhard Maack (1981) como "uma zona de eversão entre a Serra do Mar e a Escarpa Devoniana", mostrando um plano de erosão recente sobre um antigo tronco de dobras.

Em relação a hidrografia, o município de Curitiba está localizado à margem direita e a Leste da maior sub-bacia do Rio Paraná, a bacia hidrográfica do rio Iguaçu. Os principais rios de Curitiba que constituem as seis bacias hidrográficas do município são: rio Atuba, rio Belém, rio Barigüi, rio Passaúna, Ribeirão dos Padilhas e rio Iguaçu (Figura 1.2). A maior bacia hidrográfica de Curitiba é a do Rio Barigüi, que atravessa o município de Norte a Sul e perfaz um total de 139,9 Km². Ao sul do município tem-se a menor bacia hidrográfica de Curitiba, a do Ribeirão dos Padilhas, com 33,6 km² de área.

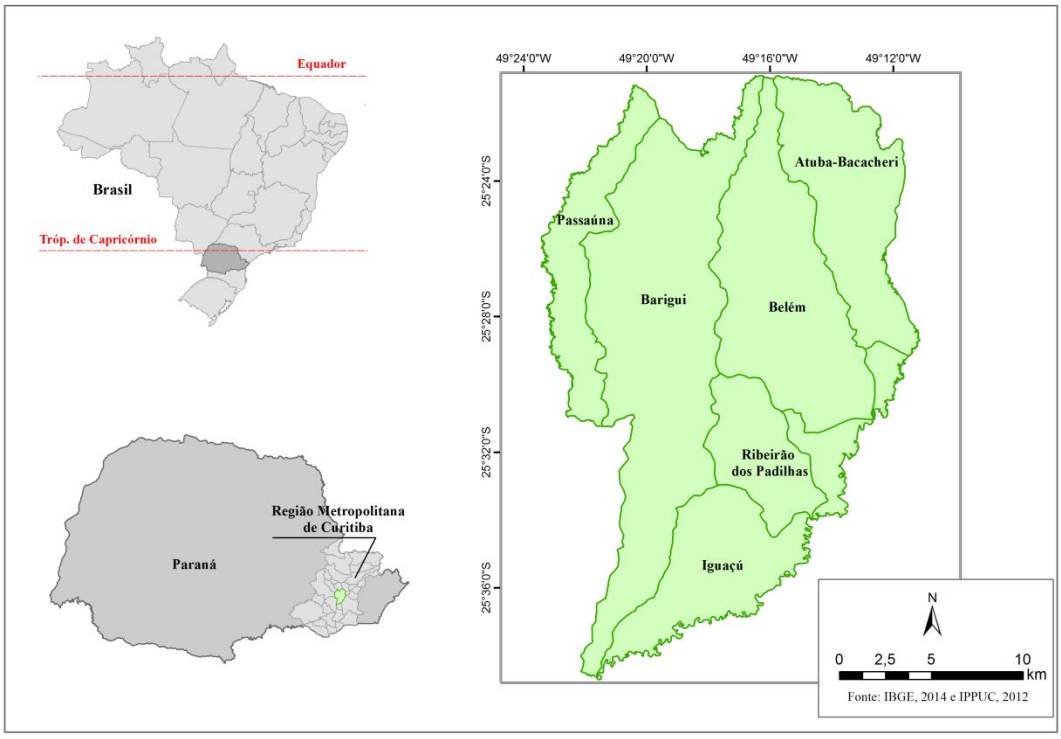

Figura 1. Mapa de localização de Curitiba.

Dados de chuva

Para o trabalho optou-se por utilizar como base os dados de chuva estimada a partir da integração das informações provenientes de radar meteorológico, satélite e pluviômetros, utilizando o método de Análise Objetiva Estatística (ANOBES)

De acordo com Pereira Filho (2004) é um dos mais eficientes esquemas de interpolação de dados. Este método foi inicialmente desenvolvido em 1963 e posteriormente recomendado pela Organização Meteorológica Mundial (WMO) em 1970, para interpolação de dados hidrometeorológicos. Este método de interpolação (Daley, 1991 apud Calvetti et al 2006), embora simples e dedutível matematicamente, foi pouco aplicado operacionalmente até recentemente, por causa das limitações de processamento computacional. Com o advento de computadores com grande capacidade de processamento e armazenamento, e a um custo baixo, estas restrições ao uso do método ANOBES foram superadas.

A equação de Análise Objetiva Estatística (ANOBES) é dada por:

$$P_a(x_r) = P_b(x_r) + \sum_{n-1}^{N} W_n[P_o(x_n) - P_b(x_n)]$$

onde,

Pa (xr) é a precipitação analisada;
Pb(xr) é a precipitação estimada pelo radar (“background*”*);
Po(xn) é a precipitação medida pelo pluviômetro (“observação”);
Pb(xn) é a precipitação estimada pelo radar no pluviômetro;
*Wn*é o peso *a posteriori* a ser determinado pela configuração de dados da rede;
N é o número total de pluviômetros;
Xr e *xn* são as respectivas posições dos pontos de grade do radar e dos pluviômetros.

Para derivar os pesos, assume-se que os erros de observação e a precipitação estimada pelo radar (“background”) não têm correlação e não tenham viés. A variância esperada do erro da análise, derivada a partir da equação descrita abaixo, é minimizada em relação aos pesos *Wn*. A precipitação integrada é derivada da soma das estimativas obtidas pela análise objetiva estatística dividida pela somatória do quadrado das diferenças de cada campo analisado.

$$P_{Est}\,(x_i, y_i) = \frac{P_{Rad}(x_i,y_i)\ x\ E_{Rad}^{-2}(x_i,y_i) + P_{Sat}(x_i,y_i)\ x\ E_{Sat}^{-2}(x_i,y_i)}{E_{Rad}^{-2}(x_i,y_i)+E_{Sat}^{-2}(x_i,y_i)}$$

Onde

$E_{rad}(x_i, y_i) = (P_{Rad}(x_i, y_i)$ - $P_{Obs}(x_i, y_i)$
$E_{Sat}(x_i, y_i) = (P_{Sat}(x_i, y_i)$ - $P_{Obs}(x_i, y_i))$

Assim a integração é ponderada pelas diferenças entre as estimativas de precipitação por radar e satélite e a precipitação medida na rede de pluviômetros. Desta forma, pretende-se obter o padrão espacial das medições por sensoriamento remoto e ponderá-la pela melhor medição volumétrica da intensidade de chuva obtida pela rede de pluviômetros.

Testes utilizando esta técnica foram feitos por Calvetti et al (2007), mostrando que a integração das informações torna-se importante não apenas para a análise de sistemas isolados de precipitação, mas também para sistemas frontais fornecendo estimativas mais apropriadas para estudos hidrológicos em bacias hidrográficas. Calvetti et al (opcit), comenta que janeiro é um mês de muitos eventos rápidos de chuva e de intensidade forte a extrema (acima de 50 mm). Não raras vezes não é possível identificar tais fenômenos devido a variação espacial e temporal das chuvas. Mesmo assim, a integração das informações proporcionou o melhor campo de precipitação utilizando as melhores características dos sistemas de medição.

Para este trabalho, foram compilados os dados de estimativas de precipitação horária de janeiro de 2005 a dezembro de 2010. Tais dados estavam disponíveis em pontos de grade de 4x4kmpara o estado do Paraná. Foram selecionados apenas os pontos que estavam inseridos nas bacias que drenam para a bacia do rio Iguaçu na RMC (Região Metropolitana de Curitiba). Foi calculada a chuva média utilizando-se o método de *Thiessen* e do *Inverso da Distância ao Quadrado (*r^{-2})

A escolha de tais métodos se deu em função dos mesmos serem extensivamente discutidos em trabalhos científicos e ainda utilizados para o cálculo de chuva média para diversas bacias de maneira operacional proporcionando bons resultados.

A diferença entre os dois métodos é que em Thiessen utilizou-se apenas os pontos de grade mais próximos a bacia, enquanto que em r^{-2} são considerados todos os pontos para todas as bacias, embora também seja atribuído peso maior aos pontos mais próximos.

Com os dados de chuva calculados para cada bacia e com o intuito de identificar qual a influência de dias anteriores ao que ocorreu o alagamento, convencionou-se utilizar o dia da ocorrência e três dias anteriores a este, por entender que a ocorrência de alagamentos é dinâmica, ou

seja, podem ocorrer com chuvas torrenciais ou ainda ter influência de chuvas ocorridas em dias anteriores. Para tanto, foi calculada a chuva acumulada de 6 em 6 horas obtendo-se 16 valores.

A Regressão Logística

Com o objetivo de identificar a previsão da probabilidade de ocorrência de alagamentos nas bacias hidrográficas, inicialmente elaborou-se diversos testes com as três funções de ligação (Logit, Normit e Gompit). Os resultados de tais testes foram compilados e pode-se então escolher qual a melhor função de ligação bem como qual o melhor método (Inverso da Distância ao Quadrado ou Thiessen) para o cálculo da chuva média.

Os dados de entrada foram constituídos pelas séries de chuva acumulada de 6 em 6 horas para 4 dias, gerando portanto, 16 valores consecutivos de chuva (16 dimensões de chuva acumulada), sendo que o primeiro valor refere-se a chuva acumulada de 6h no início do 3º dia anterior ao da ocorrência e assim sucessivamente até o último valor que refere-se a chuva acumulada das 18 até as 24h do dia da ocorrência. A última coluna da tabela possui valor 0 (zero) ou 1 (um). Esta coluna tornou-se necessária em função de o modelo preditivo logístico trabalhar com variáveis binárias, ou seja, o cálculo do coeficiente logístico compara a probabilidade de um alagamento ocorrer (1) com a probabilidade de ele não ocorrer (0). A Tabela 1 mostra o exemplo de tabela construída que serviu de base para entrada no Minitab.

Tabela 1. Exemplo de tabela construída que serviu de base para entrada no Minitab

Ano	Mês	Dia	6|3	12|3	18|3	....	6|0	12|0	18|0	24|0	Sit.
2005	1	5	0	0	0	....	0,91	0,01	7,08	1,03	1
2005	1	6	0	0	1,62	....	0	0	6,52	1,5	0
2005	1	7	10,17	1,64	3,96	....	1,25	0	0,57	0,18	1

Data Chuva Ac. de 6 em 6h Ocorrência

Para um determinado modelo preditivo, existem basicamente duas fases que devem ser seguidas. A primeira de "Calibração" e a segunda de "Verificação". A calibração tem por finalidade promover ajustes de alguns parâmetros do modelo para que os resultados simulados tenham comportamento semelhante aos reais ou experimentais. Já a verificação, considerando que modelos são construídos a partir de uma série de pressupostos e simplificações sobre o comportamento do sistema real, consiste em avaliar se esses pressupostos e simplificações foram corretamente implementados no modelo computacional. A partir disso pode-se então, calcular a acurácia do modelo gerado e avaliar seus resultados via determinado método ou teste.

Para a calibração, foram utilizados os dados de 01/01/2005 a 31/12/2009 e para a verificação os dados de 01/01/2010 a 31/12/2010. Tal definição se deu em função da necessidade de se utilizar um período maior de dados para que o modelo possa ser treinado e, a partir de seu resultado, testado utilizando-se um período menor de dados. Ainda, levou-se em consideração para tal definição, a relação entre o número de ocorrências totais de alagamentos em cada ano e o número de dias com ocorrência para cada ano. Como o número de dias com ocorrência em cada ano permaneceram muito próximos com exceção do ano de 2007, optou-se então por se utilizar o ano de 2010 para testar o modelo.

Com tais considerações a cerca de quais períodos deveriam ser utilizados para a calibração e verificação do modelo, partiu-se para a geração das previsões probabilísticas para cada dia do período considerado na calibração.

Para tanto, no software MINITAB, existe a ferramenta de regressão logística binária. Utilizando-se desta ferramenta gerou-se as previsões probabilísticas de ocorrência de alagamentos para cada dia, utilizando-se as três modalidades de funções de ligação, ou seja, "Logit, Normit/Probit e Gompit" e ainda os dados de chuva calculados via método de Thiessen e r^{-2}.

Os resultados foram gerados individualmente para cada bacia hidrográfica (Barigui, Belém, Iguaçu, Atuba e Ribeirão dos Padilhas) e ainda para o município de Curitiba. Não foi considerada a bacia hidrográfica do Passaúna em função de possuir um número praticamente insignificante de ocorrência de alagamentos.

A ideia de gerar os resultados por bacia e também para o município de Curitiba está centrada no fato de poder avaliar e selecionar se deveria ser

gerado um modelo para cada bacia ou então um modelo geral para Curitiba e posteriormente aplicado nas diferentes bacias hidrográficas. Ainda, por experiências relatadas, nem sempre o melhor modelo calibrado gera os melhores resultados na verificação e vice-versa.

Rede Neural SOM (Self Organizing Map – Mapas Auto-Organizáveis)

Como todas as redes neurais, as de Kohonen são formadas por um conjunto de elementos simples, chamados neurônios, organizados em estruturas mais complexas, que funcionam em conjunto: a rede.

Cada neurônio é uma unidade de processamento que recebe estímulos (de fora do sistema ou de outros neurônios) (Figura 3.5), e produz uma resposta (para outros neurônios ou para fora do sistema). Tal como os neurônios do cérebro, os das redes neurais são interligados entre si por ramificações através das quais os estímulos são propagados. O processo de aprendizado consiste em reforçar as ligações que levam o sistema a produzir respostas mais eficientes.

Para a geração das previsões probabilísticas de alagamentos via a rede SOM, foi definido o mesmo conjunto de preditores utilizados para o modelo regressivo logístico, sendo constituídos pelas séries de chuva acumulada de 6 em 6 horas para 4 dias ou seja, chuva ao longo de 96 h, a partir de sua decomposição em 16 valores recursivos de 6 horas (diferenças entre a chuva acumulada no tempo *t e t-6*).

Com tais dados partiu-se para a inicialização do SOM, em que foram feitas diversas simulações e mensurados os erros. O conjunto de pesos iniciais geralmente é gerado aleatoriamente com valores pequenos e ao longo das simulações este valor deve ser alterado até encontrar o menor erro.

Este procedimento teve como objetivo escolher a topologia adequada, definir o número de interações para a rede e ainda gerar as curvas (Codebooks) que, baseadas nos dados de entrada, visam reproduzir os padrões, neste caso, os padrões de chuva para 4 dias. Existem metodologias, para definição de todos estes procedimentos, no entanto, normalmente estas escolhas são feitas de forma empírica através de inúmeros testes e avaliações (Ludwig Jr. e Montgomery, 2007). Por isso, a definição da configuração de redes neurais é ainda considerada uma arte que requer certa experiência por parte de seus usuários.

Os procedimentos foram executados utilizando o pacote de treinamento SOM desenvolvido por Kohonen, tendo como base o sistema operacional Linux bem como todas as funções para inicialização do SOM, treino e mensuração do erro, chamado de erro de quantização.

Como resultado, esta rede gera os chamados "codebooks" que objetiva a formação de padrões que representem as características de certo momento sendo seu uso baseado na comparação entre um elemento do codebook e o arquivo de entrada através de sua distância euclidiana.

Como os codebooks são traduzidos como os padrões de chuva, torna-se inviável utilizar um número muito elevado de codebooks e também um número muito pequeno. Observando-se a relação entre erro e número de codebooks, após os treinamentos, testes e mensurações dos erros, definiu-se que seriam considerados e gerados 4 arquivos de codebooks (saída) com topologia retangular, 2 com dimensão de 10 x 10, portanto, 100 (cem) codebooks (padrões de chuva) e mais 2 com dimensão de 16 x 16, portanto 256 codebooks.

Sabe-se que o objetivo do codebook é a formação de padrões que representem as características de certo momento e que seu uso é baseado na comparação entre um elemento do codebook e o arquivo de entrada através de sua distância euclidiana. O codebook escolhido para representar o evento será então o de menor distância.

Tendo-se como base estes 4 arquivos de codebooks, partiu-se para a fase de aprendizagem e geração das previsões probabilísticas, utilizando-se a aprendizagem do tipo fuzzy em que a função de pertinência foi a gaussiana, ajustada dinamicamente conforme a amostra apresentada.

Indicadores de Desempenho

De acordo com Krzysztofowicz (2001) citado por Alvim, Breda e Sabóia (2011), nas previsões probabilísticas, as incertezas devem tomar a forma de probabilidade preditiva, entendida como uma medida numérica do grau de acerto acerca da ocorrência de um evento, condicional a todas as informações (dados, modelos e julgamentos) utilizadas no processo de previsão.

Neste sentido, diversos métodos já testados e difundidos na literatura podem ser utilizados para avaliar a acurácia e desempenho de um determinado modelo. Neste trabalho foi utilizado como método a Curva ROC (Relative Operating Characteristic) e a área sob a curva bem como os diagramas de confiabilidade, discriminação e refinamento, métodos discutidos nos itens 2.10.1, 2.10.2 e 2.10.3.

De acordo com Leite (2009), o aspecto fundamental embutido na utilização da Curva ROC é a possibilidade de análise do desempenho do um

sistema previsor como um todo e não de uma seleção particular de situação operacional, que normalmente envolverá aspectos de qualidade conflitantes.

Assim, para a construção da Curva ROC, foi considerada a taxa POD e PoFD que são:

POD (*probability of detection*)

A taxa PoD pode ser definida como sendo a proporção de ocorrências que foram corretamente previstas. É a probabilidade condicional de que a previsão seja de ocorrência dado que o evento foi observado. Este indicador pode variar de 0 a 1 tendo seu melhor valor em 1.

PoFD (probability of false detection)

A taxa PoFD de falso alerta é a proporção de não ocorrências que foram incorretamente previstas. Assim como o PoD, este indicador varia de 0 a 1, porém tem-se como melhor valor o 0.

Além da avaliação via a curva ROC, foi calculada a área sob a curva (AUC), bem como utilizados os diagramas de confiabilidade, discriminação e refinamento.

Resultados

Resultados dos Modelos de Regressão Logística e Redes Neurais (Som)

A aplicação e o estudo do método de regressão logística e de redes neurais, para os dados de precipitação acumulados de 6 em 6h utilizando-se do método de Thiessen), resultou um modelo numérico preditivo. Esse modelo pode tornar-se uma ferramenta com a qual os tomadores de decisão, por exemplo, podem quando em função de um determinado padrão de chuva e grande probabilidade de ocorrência de um alagamento, emitir alertas para a população em geral, optar em interditar determinadas ruas ou ainda retirar moradores das áreas mais críticas.

Neste sentido, Leite (2008), comenta que a decisão de emissão do alerta deve ser realizada por autoridade pública municipal, ou delegada para a coordenação de defesa civil, com o objetivo primário de mitigar os impactos globais dos alagamentos, entre os quais salvar vidas e minimizar perdas econômicas. O mesmo autor ainda afirma que o alerta funciona como um gatilho para motivar a população a decidir e empreender medidas de proteção contra as inundações ou alagamentos.

A decisão de emissão do alerta pode ser orientada por diversos critérios, condicionantes de seu resultado e efetividade, e abranger as abordagens determinística e probabilística de análise da previsão. A seleção do critério e limites que determinam a emissão do alerta deve levar em consideração as consequências sociais, e seus efeitos econômicos e comportamentais, que são derivados do desempenho do sistema em termos de falso alerta, número de alagamentos detectados e perdidos e tempo disponibilizado para a população empreender medidas de proteção (LEITE, 2008).

Levando em consideração tais ideias e pensando na geração e avaliação de um possível sistema de emissão de alertas, neste item, serão apresentados os resultados gerados a partir da base de dados trabalhada para dois modelos, ou seja, o de regressão logística e de redes neurais (SOM) bem como a avaliação do desempenho dos modelos gerados para cada bacia hidrográfica e para o município de Curitiba.

Desempenho do Modelo de Regressão Logística

A regressão logística binária apresenta-se como um método para determinar a probabilidade de ocorrência dos valores preditos de uma variável dicotômica. Tendo como premissa que a construção de um modelo regressivo do tipo logístico tem como objetivo fazer a previsão da probabilidade de ocorrência de alagamentos para Curitiba, utilizou-se para sua avaliação a Curva ROC e a área sob a curva, que avaliam a acurácia e o desempenho de um modelo, tanto na calibração quanto na verificação. O modelo assim avaliado pode ser utilizado de forma operacional para a emissão de alertas de alagamentos em Curitiba em função de um determinado padrão de chuva como já explorado com Lohmann (2011).

Neste primeiro momento a atenção ficou voltada em investigar e avaliar qual o melhor método para cálculo da chuva média (Thiessen ou r^{-2}) e também qual a função de ligação (Logit, Normit, Gompit) possui desempenho que resulte em um bom ajuste dos dados para o modelo de regressão logística. Para tais avaliações foram construídas curvas ROC e calculada a área sob a curva, que por consequência também já serviram de base para avaliação geral do modelo de regressão logística.

A Figura 2 ilustra as curvas ROC para a bacia hidrográfica do Barigui, Belém, e também para o município de Curitiba e referem-se a Calibração. O cálculo da área sob a curva está apresentado na Tabela 2.

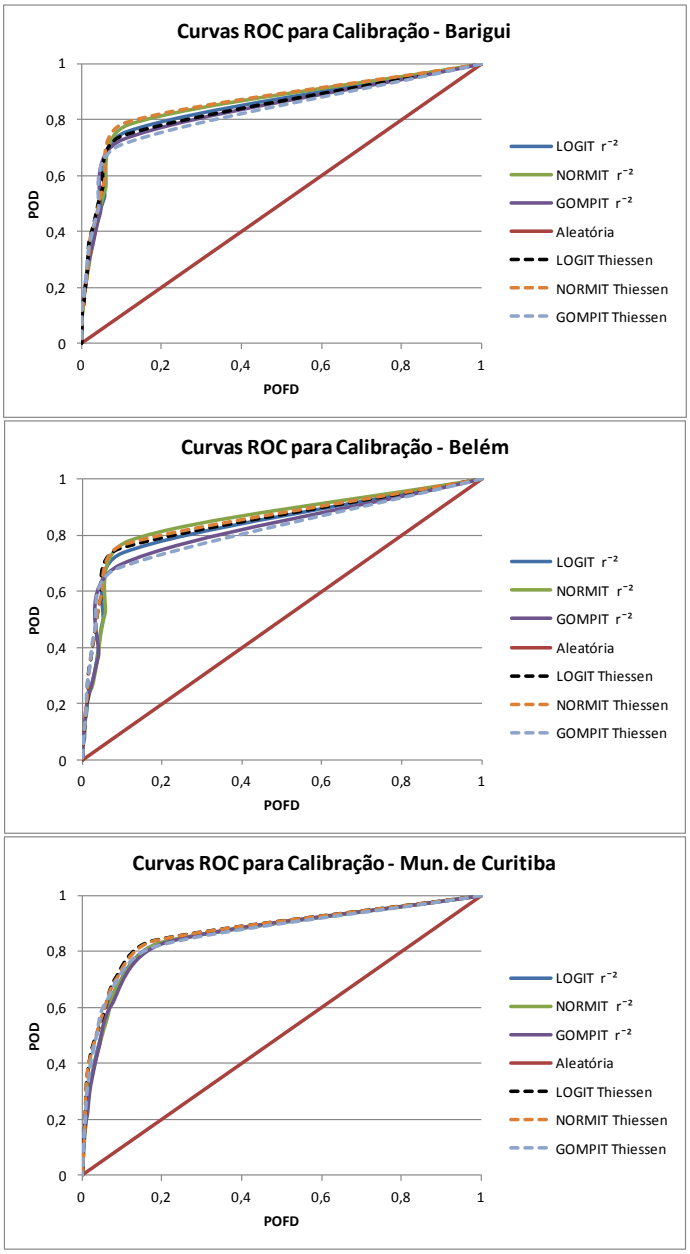

Figura 2. Gráficos apresentando as Curvas ROC geradas para as três funções de ligação e para os dois métodos de cálculo de chuva média no período de calibração para as bacias hidrográficas e para o município de Curitiba.

Tabela 2. Resultados da área sob a curva ROC para as bacias hidrográficas e o município de Curitiba para a Calibração, Regressão Logística

Função	Método	Área Barigui	Área Belém	Área Iguaçu	Área Atuba	Área Ribeirão	Área Curitiba
Logit	r^{-2}	0,8353	0,8262	0,6714	0,7946	0,7875	0,8569
Normit	r^{-2}	0,8464	0,8421	0,6879	0,8282	0,7908	0,857
Gompit	r^{-2}	0,8238	0,8057	0,5607	0,7765	0,7591	0,8513
Logit	Thiessen	0,8340	0,8419	0,6430	0,8545	0,7968	0,8704
Normit	Thiessen	0,8549	0,8468	0,6702	0,8535	0,8104	0,8693
Gompit	Thiessen	0,8185	0,8081	0,5081	0,7754	0,7385	0,8568

Fonte: elaborado pelos autores, 2011. Observação: em vermelho os valores máximos de área.

Observando-se a Figura 2 e a Tabela 2 nota-se que a chuva média calculada pelo método de Thiessen, utilizadas a para a calibração e avaliadas pela Curva ROC e área sob a curva ROC, possui melhor desempenho quando comparada com o r^{-2}, visto que, com exceção da bacia do Iguaçu em que o método r^{-2} se mostrou superior, em todas as outras bacias, inclusive para o município de Curitiba o método de Thiessen se mostrou superior.

Em média, a área sob a curva que varia de 0,5 (nenhuma acurácia aparente) a 1,0 (acurácia perfeita), ficou acima de 0,8, ilustrando que para a calibração, o método para cálculo da chuva média deve ser o de Thiessen.

Sob outra ótica, percebe-se que quando calculada a chuva média por bacia e para Curitiba, nota-se que a área sob a curva para Curitiba foi superior a 0,87, mostrando um ganho quando comparada com as bacias.

Tal diferença pode ainda possuir outra explicação que está centrada principalmente na resolução espacial em que é gerada a integração dos dados de satélite, radar e pluviômetros. Tal integração possui como resolução espacial uma célula de 4x4 km. Portanto em nível de bacia, há um número menor de células a serem consideradas no cálculo da chuva média e por consequência maior possibilidade de perda de informações. Para Curitiba como um todo, tem-se situação contrária, ou seja, maior número de células e menor perda de informações.

Neste sentido, chuvas convectivas localizadas e que se dissipam rapidamente, são identificadas e registradas com maior acurácia pelo radar meteorológico, sendo que tanto o satélite quanto o pluviômetro podem não registrar os valores de precipitação, dependendo da sua área de abrangência. Dessa forma, no processo de integração da chuva, esses dados são subestimados. Calvetti e Benetti (2007) já alertavam para este fato. Assim, possivelmente tais chuvas, por exemplo, são identificadas de melhor maneira considerando o limite municipal ao invés do limite das bacias hidrográficas.

Procurando responder qual a função de ligação (Logit, Normit e Gompit) possui melhor desempenho e ajuste dos dados para o modelo de regressão logística, há três possibilidades para função de ligação, que permitem ajustar uma classe ampla de modelos de resposta binária. São elas: o inverso da função distribuição logística cumulativa (logit), o inverso da função distribuição normal padrão cumulativa (normit = probit), e o inverso da função distribuição de Gompertz (gompit = complementar log log).

É necessário escolher uma função de ligação que resulte um bom ajuste dos dados coletados. Pode-se usar a estatística da qualidade do ajuste para comparar os resultados com diferentes funções de ligação.

Resumidamente, pode-se dizer que o modelo Normit (ou Probit) é uma alternativa do modelo Logit que admite a função de distribuição Normal (standard) para expressar a relação não linear entre as probabilidades estimadas da variável dependente e as variáveis explicativas.

O modelo Normit assim como o Logit é estimado pelo método da Máxima Verosimilhança, método de estimação não linear.

Uma vantagem da função de ligação Logit é que ela provê a estimativa da razão das chances para cada variável-preditora no modelo. Se a razão das chances for um, não há associação.

O método utilizado pela função Logit é o da máxima verossimilhança que tem como objetivo maximizar a função da verosimilhança (ou o logaritmo desta), isto é, obter (através de um processo iterativo) os valores dos parâmetros do modelo de modo que a probabilidade de observar os valores de Y_i seja a mais alta (máxima) possível. Este método, de acordo Lo (1986), é mais robusto no que se refere a confiabilidade dos resultados, do que a regressão linear.

Os resultados das estimações dos modelos Logit e Normit são similares em termos de significância estatística e precisão de ajustamento, contudo, os valores dos coeficientes estimados não são diretamente comparáveis.

De forma geral, a principal diferença entre os dois modelos está no fato de a distribuição logística apresentar caudas ligeiramente mais grossas do que a distribuição normal do modelo Normit, isto é, a probabilidade condicional P_i se aproxima mais lentamente para o 0 ou 1 do que no caso do modelo Logit.

Basicamente, não há razão forte que justifica optar por um dos dois modelos, uma vez que o método de estimação é o mesmo (método da Máxima Verosimilhança). Diferem apenas na função de distribuição acumulada.

Na prática, o modelo Logit é mais utilizado devido a sua especificação matemática mais simples.

Com tais indicações, partiu-se então para a fase de verificação dos resultados gerados na calibração. Para tanto, foram geradas as equações de regressão com os parâmetros e valores gerados na fase de calibração. No entanto, tais equações foram geradas apenas para os melhores resultados já indicados anteriormente e visualizados em vermelho na Tabela 4.8.

Com base nisso, optou-se ainda em já realizar e apresentar no gráfico das Curvas ROC uma comparação entre os resultados do melhor modelo gerado para Curitiba, no caso, Logit – Thiessen, e o melhor modelo gerado para cada uma das bacias. Tais resultados podem ser visualizados na Figura 3 e na Tabela 3.

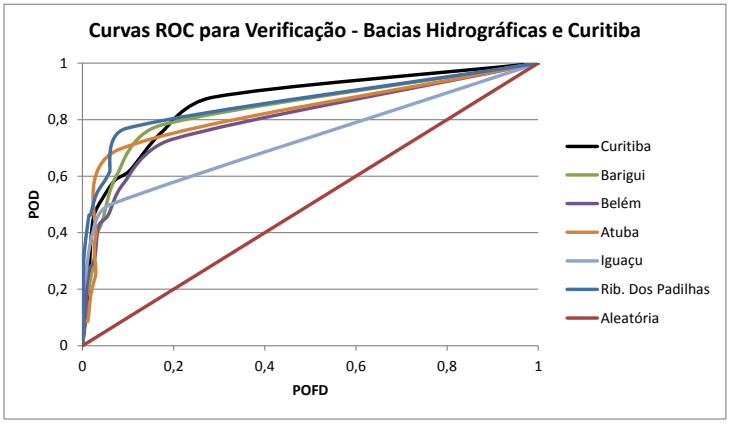

Figura 3. Comparação das Curvas ROC geradas para a verificação considerando apenas os melhores resultados para Curitiba e para as bacias hidrográficas.

Tabela 3. Comparação da área sob a curva gerada para os melhores resultados por bacia e para Curitiba - Verificação

Resultado	Modelo	Calibração	Verificação
Melhor Modelo Curitiba	Logit – Thiessen	0,8704	0,8604
Melhor Modelo Barigui	Normit – Thiessen	0,8549	0,8324
Melhor Modelo Belém	Normit – Thiessen	0,8468	0,7998
Melhor Modelo Iguaçu	Normit – R^{-2}	0,6879	0,7277
Melhor Modelo Atuba	Logit – Thiessen	0,8545	0,8113
Melhor Modelo Rib dos Padilhas	Logit – Thiessen	0,8104	0,8541

Fonte: elaborado pelos autores, 2011.

As curvas apresentadas na Figura 3 ilustram os resultados para a verificação do modelo em cada uma das bacias hidrográficas bem como para o município de Curitiba. Observa-se que para Curitiba a verificação apresentou resultado de 0,86 para a área sob a curva, mostrando bom desempenho quando considerado que para o período de verificação foi utilizado apenas os dados do ano de 2010.

Para as bacias de modo geral, verificou-se situação semelhante à de Curitiba, ou seja, os modelos gerados apresentaram bom desempenho, tendo em média 0,8 de área sob a curva. A bacia do Ribeirão dos Padilhas teve destaque, apresentado valor de área acima de 0,85.

Portanto, fazendo uso do método de regressão logística, atingiu-se o objetivo de gerar um modelo de previsão de alagamentos para Curitiba e para as bacias, utilizando-se como base os dados de chuva acumulados de 6 em 6 horas para o dia do evento e 3 dias anteriores, concluindo-se que os dados mencionados permitem gerar um modelo de previsão probabilística satisfatório. Salienta-se que não se considera que os modelos gerados por bacia hidrográfica não poderiam ser empregados, no entanto, apenas possuem resultados inferiores ao modelo gerado para Curitiba, mas que estão dentro de um patamar de confiança.

Ainda, os resultados permitem concluir que é possível identificar via regressão logística a relação entre o índice de precipitação e os alagamentos. Mesmo ficando associados os maiores valores dos parâmetros para as variáveis que compõem o dia do evento (como já era esperado), ou seja, os últimos 4 valores de precipitação acumulados de 6 em 6 horas, as variáveis referentes aos 3 dias anteriores (no caso mais 12 variáveis) também são importantes na construção do modelo regressivo logístico. Isso permite distinguir de melhor forma as situações com elevada probabilidade de um alagamento vir a ocorrer ou não.

Dessa forma, tais resultados revelam que o modelo regressivo logístico construído e testado (Calibrado e Verificado) pode ser utilizado para elaborar previsões probabilísticas de modo operacional, objetivando a emissão de alertas de alagamentos em Curitiba, já que o mesmo possibilita associar um determinado padrão de chuva à ocorrência de um alagamento.

Desempenho do Modelo Redes Neurais (SOM)

De posse das conclusões a cerca do modelo regressivo do tipo logístico partiu-se para a construção do modelo baseado em rede neural (SOM), também objetivando a previsão da probabilidade de ocorrência de alagamentos. Visto que para o modelo de regressão logística observou-se que o método para cálculo da chuva média mais adequado foi o de Thiessen e os melhores resultados foram gerados empregando-se como base Curitiba ao invés das bacias hidrográficas, definiu-se que seria utilizada a mesma base de dados gerada para Curitiba para a construção do modelo baseado em rede neural, até mesmo para que os métodos pudessem ser comparados.

Para tanto, levando-se em consideração os procedimentos adotados e já devidamente explicados na metodologia, foram geradas as probabilidades para a série de dados. Para tal, foram utilizados 4 arquivos de codebooks, 2 com dimensão de 10 x 10, portanto, 100 codebooks (padrões de chuva) e mais 2 com dimensão de 16 x 16, portanto 256 codebooks. Tais procedimentos foram empregados tanto para a Calibração quanto para a Verificação, sendo os mesmos avaliados via Curva ROC e área sob a curva ROC.

A Figura 4 mostra as curvas ROC produzidas para a calibração, já apresentando também a comparação com o melhor modelo gerado para Curitiba que teve 0,87 de área sob a curva usando regressão logística.

Para melhor entendimento e compreensão, convencionou-se chamar os dois arquivos de 100 codebooks de *"SOM 100 Codebooks A"* e *"SOM 100 Codebooks B"* e os dois arquivos de 256 codebooks de *"SOM 256 Codebooks A"* e *"SOM 256 Codebooks B"*. O cálculo da área sob a curva está apresentado na Tabela 4.

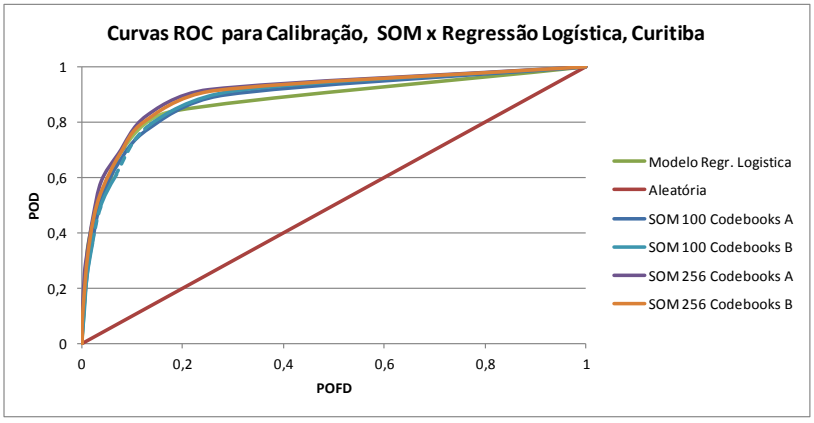

Figura 4. Curvas ROC geradas para a calibração considerando apenas os melhores resultados para Curitiba.

Tabela 4. Resultados da área sob a curva para o município de Curitiba para o período de Calibração

Resultados	Período de Calibração
Melhor Mod. Ctba – Regr. Log.	0,8704
SOM 100 Codebooks A	0,8818
SOM 100 Codebooks B	0,8856
SOM 256 Codebooks A	0,9046
SOM 256 Codebooks B	0,8984

Fonte: elaborado pelos autores, 2011.

Conforme pode-se observar na Figura 3 e na Tabela 4 as curvas apresentadas traduzem as expectativas de desempenho de um possível sistema de alerta usando um modelo baseado em

rede neural. Analisando os resultados para a calibração, observa-se que os valores de área sob a curva foram idênticos quando utilizados 100 e 256 codebooks, com 0,88 e 0,90 de área respectivamente. De forma geral, pode dizer que o desempenho é praticamente igual, com variação de apenas 0,02 no valor da área.

Em uma análise mais detalhada, verifica-se que o modelo que apresentou maior área foi o que utilizou o arquivo chamado *"SOM 256 Codebooks A"* com área acima de 0,90. Comparando-se tais resultados com o melhor resultado advindo da regressão logística, constata-se que houve um ganho de 0,03 no valor da área, ou seja, passou de 0,87 para 0,90, mostrando que o modelo construído via SOM pode ser considerado superior.

Com vistas ao que foi descrito e tendo como premissa que o SOM possui melhor desempenho do que a regressão logística utilizando Curitiba como base, decidiu-se investigar se os resultados também eram satisfatórios se fossem gerados por bacia hidrográfica. Para tanto, utilizou-se da base de dados de chuva para Curitiba e do melhor arquivo de codebooks (*"SOM 256 Codebooks A")* e aplicou-se o modelo construído para Curitiba nas bacias hidrográficas. São apresentados a seguir os resultados para as bacias analisadas.

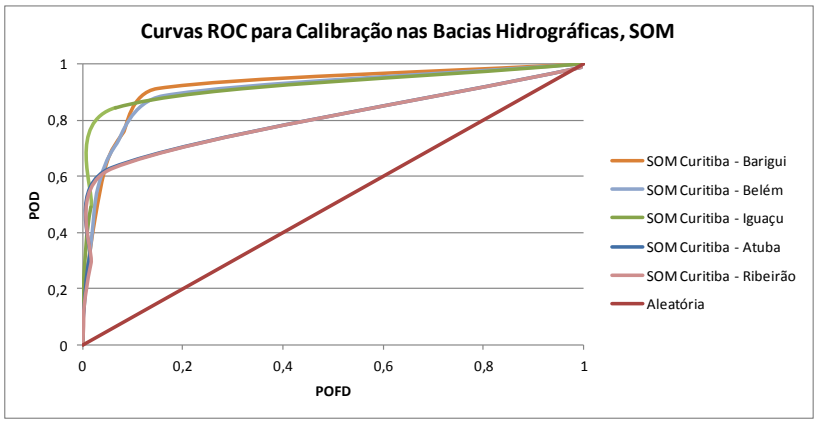

Figura 5. Curvas ROC geradas para a calibração considerando as bacias hidrográficas.

Tabela 5. Resultados da área sob a curva ROC para as bacias hidrográficas no período de Calibração referentes ao modelo SOM

Resultados	Período de Calibração
Curitiba – Barigui	0,9126
Curitiba - Belém	0,8978
Curitiba - Iguaçu	0,893
Curitiba - Atuba	0,7823
Curitiba – Ribeirão dos Padilhas	0,7789

Fonte: elaborado pelos autores, 2011.

Conforme ilustra a Figura 5 e a Tabela 5 observa-se que para a bacia do Barigui, Belém e Iguaçu os valores de área sob a curva ROC são em torno de 0,90, estando no mesmo nível de desempenho do modelo gerado para Curitiba. No entanto, para a bacia do Atuba e Ribeirão dos Padilhas, a área ficou menor mostrando desempenho inferior. Assim, mostra-se que é possível gerar os modelos utilizando o SOM por bacia hidrográfica, no entanto para algumas, com menor grau de desempenho, ou seja, não se obteve ganho quando os modelos foram gerados individualmente por bacia hidrográfica.

Levando em consideração que o modelo gerado para Curitiba possui melhor desempenho quando comparado com os modelos gerados por bacia, decidiu-se por investigar na verificação (que testa o modelo construído na calibração) apenas o melhor resultado para 100 e 256 codebooks, ou seja, *"SOM 100 Codebooks B"* e *"SOM 256 Codebooks A"*. Os resultados podem ser visualizados na Figura 6 e Tabela 6.

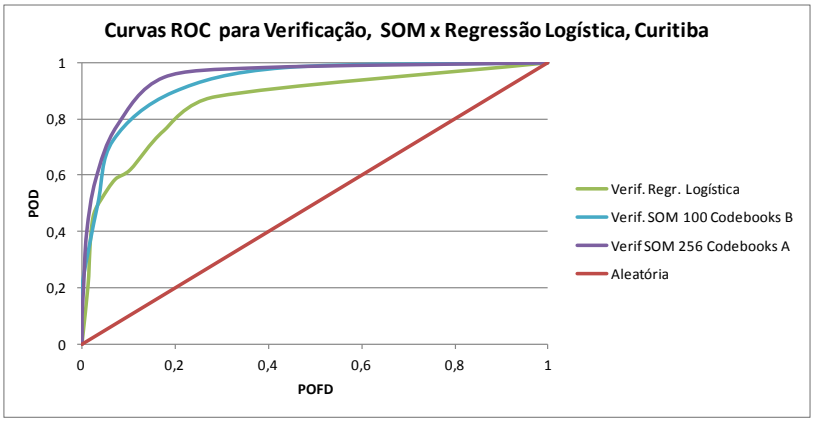

Figura 6. Curvas ROC geradas para a verificação considerando apenas os melhores resultados para Curitiba.

Tabela 6. Resultados da área sob a curva para o município de Curitiba para a Verificação, SOM

Resultados	Período de Verificação
Melhor Mod. Ctba – Regr. Logístico	0,8604
SOM 100 Codebooks B	0,9199
SOM 256 Codebooks A	0,9417

Fonte: elaborado pelos autores, 2011.

Analisando a Figura 6 e Tabela 6 fica bem caracterizado que o melhor desempenho se deu para as previsões que utilizaram 256 codebooks. A área sob a curva foi de 0,94 e 0,91 para 256 e 100 codebooks respectivamente. Comparando-se agora a área do melhor modelo gerado para Curitiba utilizando regressão logística, nota-se que o modelo baseado em rede neural foi superior tanto para 100 quanto para 256 codebooks, obtendo um ganho de 0,08 pontos quando considerada a área sob a curva ROC, passando de 0,86 para 0,94.

Elaborando-se uma análise geral e comparando os resultados gerados tanto para os modelos construídos utilizando regressão do tipo logística quanto para os modelos baseados em redes neurais e, considerando o período de dados analisados, conclui-se que os dois métodos mostraram que são passíveis de serem utilizados já que geraram resultados que ficaram acima de 0,87 para a calibração e acima de 0,86 para a verificação. Pôde-se mostrar por meio das comparações que os modelos gerados via regressão logística quanto via rede de Kohonen geram melhores resultados quando aplicados para Curitiba como um todo, mas não descarta-se a hipótese de que podem ser também utilizados tendo a bacia hidrográfica como base, no entanto com desempenho um pouco inferior.

De qualquer forma, pode-se afirmar que o modelo construído com a rede SOM apresentou desempenho muito superior ao construído utilizando-se da regressão logística, visto que as curvas ROC e a área sob a curva ROC apresentaram ganhos consideráveis, chegando em 0,90 e 0,94 de área sob a curva na calibração e verificação respectivamente.

Especificamente sobre o método utilizado para avaliação do desempenho dos modelos construídos, ou seja, a Curva ROC, vale salientar que a mesma é utilizada em diversas áreas do conhecimento e é indicada pela Organização Mundial de Meteorologia (OMM) como método para avaliar o desempenho de previsões probabilísticas de clima. A curva ROC pode ser considerada um sistema altamente flexível podendo ser utilizada para avaliar o desempenho de variáveis do tipo dicotômicas, categóricas, contínuas e probabilísticas.

Considerando-se que a curva de características operacionais pode ser utilizada em um sistema de emissão de alertas para indicar o ponto de corte entre emitir o alerta ou não, seu ajustamento operacional é realizado por meio da alteração do limite de determinado preditor a partir do qual se emite o alerta, o que significa um deslocamento sobre a mesma curva ROC. Por outro lado, de acordo com Leite (2008), o ajustamento tecnológico é realizado por meio do estabelecimento de uma nova curva ROC, na qual a relação entre *POD* e *POFD* se torne mais favorável.

Especificamente em relação à taxa POD e POFD, salienta-se que a definição do ponto de corte para emissão do alerta deve ser entendida como de fundamental importância, já que possui consequências diretas para a sociedade.

A taxa POD pode ser traduzida pela proporção de *eventos* que foram alertados, mostrando a eficiência do alerta propriamente dito. A taxa POFD traduz a proporção de situações de *não evento*, mas que foram alertados, ou seja, os falsos alertas. Quando utilizadas em um sistema de forma operacional impactam a percepção de risco. Por exemplo: uma taxa POD baixa cria na população a percepção que o sistema é ineficiente pois o evento ocorre porém sem emissão do alerta. Da mesma forma, se for selecionado uma taxa POFD alta, cria-se situações de perturbação para a população pois são emitidos muitos alertas sem a ocorrência subsequente do evento, ou seja, o sistema pode ser percebido como exagerado, causando sensação de descrédito junto a população usuária.

Refletindo sobre a situação apresentada, pode-se dizer que em ambos os casos a principal consequência é o sistema perder credibilidade junto a população e aos que deveriam ser beneficiados. Primeiro porque o sistema não possui capacidade de antecipar o evento e segundo por emitir quantidade exagerada de falsos alertas. Nesse cenário, tem-se desde a geração de perdas diversas como, por exemplo, a perda de bens, falta de energia elétrica, água entre outros, até situações de estresse ou ainda perda de vidas, já que as medidas de proteção (no caso dos alertas) que deveriam ser implementadas, não o são de forma satisfatória.

Neste sentido, deve haver um balanço adequado entre POD e POFD, objetivando otimizar da melhor maneira possível a eficiência do sistema para emissão dos alertas. Portanto, o melhor sistema é aquele que possui a maior taxa POD e a menor taxa POFD, pois é o que gera motivação para o empreendimento de medidas de proteção pela população. Ressalta-se ainda que para que a população possa ser beneficiada verdadeiramente, é necessário que a mesma participe do processo e seja educada no sentido de entender que para que se tenha uma taxa POD *desejável* é necessária certa taxa POFD.

Ainda discorrendo sobre métodos de avaliação de previsões probabilísticas, decidiu-se investigar os aspectos de qualidade traduzidos via os diagramas de confiabilidade, discriminação e refinamento para as previsões geradas com o SOM e regressão logística (Figura 7). A confiabilidade traduz o grau de concordância entre probabilidades previstas e subsequentes frequências relativas observadas. A discriminação retrata a maneira como o sistema previsor identifica situações de eventos e situações de não eventos. O refinamento demonstra o grau com que o sistema provê previsões probabilísticas próximas aos extremos, ou seja, zero ou um.

A Figura 7 apresenta os diagramas de confiabilidade, discriminação e refinamento elaborados tendo como base as previsões geradas via a rede de Kohonen e regressão logística, tanto para o período de calibração quanto para de verificação. Salienta-se que em virtude dos dados utilizados para a calibração serem os mesmos utilizados para o treinamento da rede, obviamente os resultados são melhores. Os resultados a serem considerados mais relevantes são os gerados para a verificação, já que se trata de novas amostras sendo testadas nos modelos.

A confiabilidade é indicada pela proximidade das curvas analisadas, da curva traçada na diagonal. Se a curva está abaixo da linha (vermelha), isto indica superestimativa (probabilidades previstas mais altas), e pontos acima da linha indicam subestimativas (probabilidades previstas mais baixas).

Conforme se pode observar na Figura 7, para a calibração, tanto a curva que representa a regressão quanto a que representa o SOM, estão próximas da curva traçada na diagonal, indicando que as previsões foram satisfatórias. Para a verificação, nota-se que há diferenças significativas quando comparadas as curvas da regressão logística e do SOM. Para probabilidade até 0,25 a curva da regressão é muito próxima a diagonal, indicando bom desempenho. De 0,3 a 0,65 há superestimativa, de 0,65 a 0,85 subestimativas e de 0,85 a 1 superestimativas. Nota-se que há uma oscilação da curva, ora super ora subestimando, não representando portanto, noção de continuidade . No geral, a regressão mostra tendência de superestimar as probabilidades previstas o que é preferível quando se considera situações em que há risco humano, por exemplo.

Para a curva do SOM, observa-se supertimativas até 0,25 e tendência a subestimar as previsões para valores maiores de 0,3. No entanto, nota-se que a partir do limiar de 0,3 tem-se uma condição de sempre subestimar, mostrando certo grau de continuidade. Como o interesse para emissão de alertas são as probabilidades mais altas, podem ser aplicadas técnicas de recalibração para tratar o problema. Neste sentido, em função da oscilação da curva que representa a regressão, a recalibração poderia ser prejudicada.

Em relação à discriminação, nota-se que para a calibração e para a verificação, as curvas que representam a regressão e o SOM para situações de não evento (x=0) são praticamente iguais,

mostrando que os dois sistemas de previsão possuem boa capacidade de discriminar tais eventos, como já era de se esperar, já que ao longo do ano tem-se muito mais dias sem alagamentos do que com alagamentos. Para as situações de evento (x=1), as curvas da regressão e do SOM mostram comportamentos parecidos tanto na calibração quanto na verificação, no entanto mostrado dificuldade do sistema em produzir previsões probabilísticas altas para tais situações.

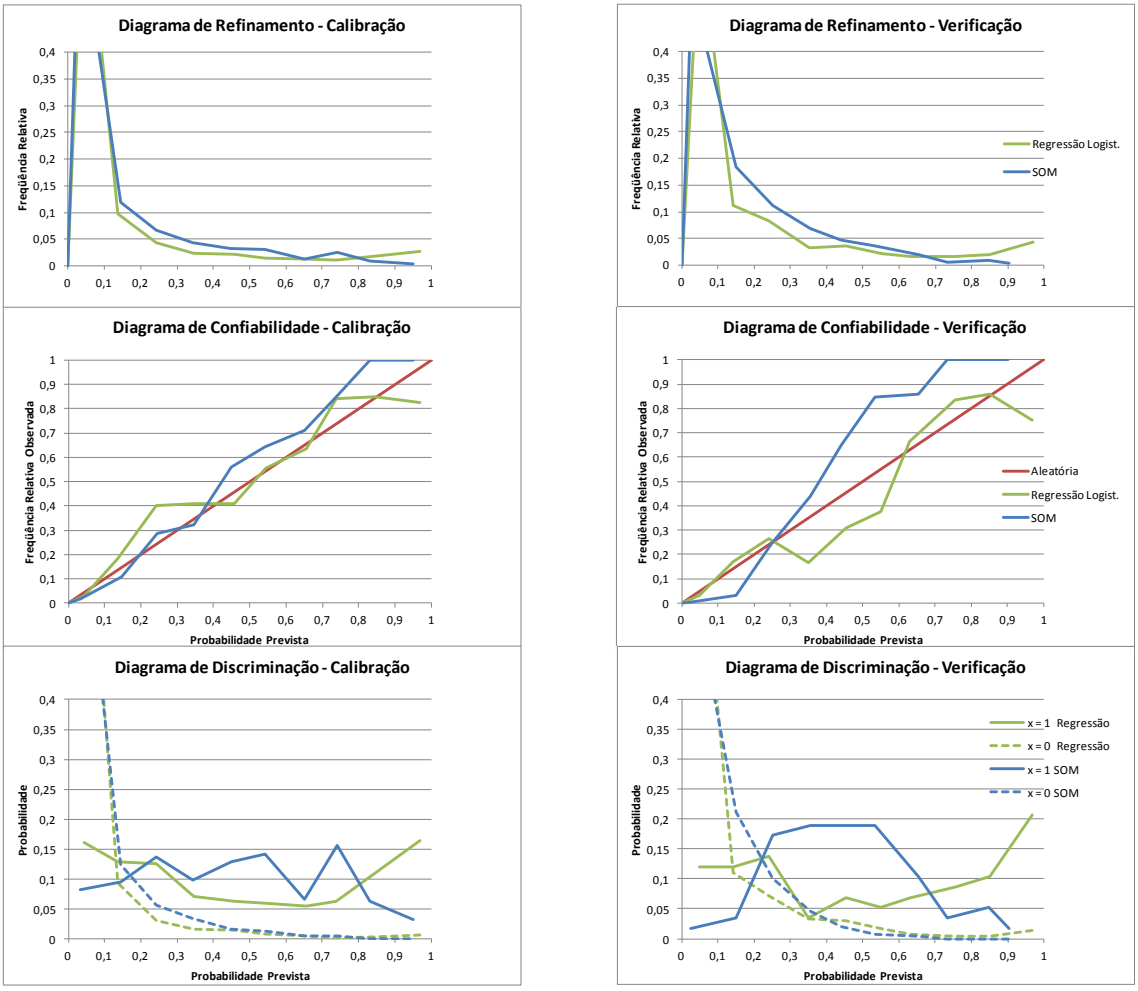

Figura 7. Diagramas de confiabilidade, discriminação e refinamento das previsões geradas via regressão logística e SOM.

Observa-se que a regressão apresentou uma capacidade um pouco maior (em torno de 10%) quando comparada com o SOM, refletido já nos diagramas de confiabilidade, onde a regressão mostrou tendência de superestimar as previsões de maior probabilidade.

No diagrama de refinamento (sharpness diagram) pode-se perceber que as curvas, tanto para a calibração quanto para a verificação, são muito parecidas ficando evidente que para previsões entre 0 e 0,2, tem-se maiores frequências observadas para o SOM e menores frequência para a regressão. Para previsões acima de 0,8, observa-se o contrário, ou seja, maior frequência para a regressão e menor para o SOM.

De forma geral, tanto o diagrama de refinamento quanto o de discriminação refletem o que é observado no diagrama de confiabilidade, como já discutido.

Aqui vale uma ressalva e explicação a cerca do que foi discutido sobre a qualidade dos métodos de avaliação para previsões probabilísticas, ou seja, ora o modelo baseado em regressão do tipo logística mostra-se melhor e ora o modelo baseado em redes neurais. Esse fato é justificável e esperado se for considerado o limitado tamanho da amostra utilizada para tais análises, ou seja, os dados trabalhados foram de apenas 6 anos (2005 a 2010).

Além disso, essa amostra ainda foi dividida, pois se necessitava calibrar os modelos

construídos e posteriormente testá-los. Assim, foram utilizados os dados de 5 anos (2005 a 2009) para construir os modelos e de 1 ano (2010) para testá-los. Em função disso, infere-se que se aumentada a amostra, os preditores utilizados para construir os modelos ainda podem ser melhorados e, como consequência, gerar aumento de desempenho e maior qualidade das avaliações ora apresentadas.

Conclusões

Em relação aos dois métodos utilizados na geração das previsões de probabilidade e avaliados via o emprego da Curva ROC, área sob a curva ROC e os diagramas de confiabilidade, discriminação e refinamento, pode-se considerar que os dois métodos empregados forneceram resultados satisfatórios do ponto de vista estatístico.

A partir da leitura de diversos trabalhos que utilizam o modelo baseado em redes neurais (SOM) como método na modelagem de diversos problemas, pode-se confirmar as expectativas que eram de que o SOM possui melhor desempenho quando comparado com a regressão logística. Portanto, este trabalho possui como ponto forte a confirmação de tais expectativas já que o mesmo as corrobora, utilizando uma mesma base de dados e mesmo método de avaliação de desempenho. Entretanto cabe destacar que os resultados deste caso não são garantia de melhor desempenho das redes neurais em outros contextos e aplicações.

A modelagem por regressão teve, inicialmente, em sua fase de concepção, o objetivo de escolher qual a melhor função de ligação entre as três possíveis (Logit, Normit, Gompit), tanto para os dados de chuva produzidos por bacia hidrográfica como para Curitiba. Obteve-se como resultado que a melhor função de ligação foi a Logit tendo como entrada para o modelo os dados produzidos para Curitiba ao invés dos dados por bacia. Tais resultados foram verificados tanto na calibração quanto verificação.

Os resultados do modelo gerado com redes neurais obtiveram melhor qualidade que os por regressão, fato identificado tanto na calibração quanto na verificação, sendo que nesta última, a área sob a curva ROC foi de 0,94.

Espera-se que os resultados obtidos com o SOM melhorem ainda mais no momento que serão disponibilizadas as informações de satélite com resolução de 1 x 1km. Dessa forma, também a resolução dos campos de chuva estimados poderão ser produzidos com resolução de 1 x 1km, reduzindo erros associados a grade hoje possível de ser explorada com resolução de 4 x 4km.

A utilização de inteligência artificial, especificamente a rede SOM, é um método alternativo de modelagem de processos relacionados não só ao ambiente natural, mas também relacionado a outras ciências como a medicina, informática, administração entre outras, tendo como grande vantagem a possibilidade de modelar processos não-lineares de forma implícita. Mas como o trabalho indica, uma limitação encontra-se no tamanho da amostra utilizada para o treinamento e aprendizagem da rede. Acredita-se que para este estudo, os resultados poderiam ser melhorados com uma amostra maior, ou seja, um maior número de dados de estimativa de chuva e também de alagamentos. No entanto, foram utilizados os dados de 2005 a 2010, disponíveis no momento.

Referências

Brandão, A. 2001. Clima Urbano e Enchentes na Cidade do Rio de Janeiro, in: Guerra, A. J. T., Cunha, S. B. (Orgs.), Impactos ambientais urbanos no Brasil. Bertrand Brasil, Rio de Janeiro.

Calvetti, L., Beneti, C., Pscheidt, I., Stringari, D. E., Pereira Filho, A. J., 2007. Integração de Estimativas de Precipitação por Radar, Satélite e Pluviômetros: Análise Espacial para o Paraná. Anais do XVII Simpósio Brasileiro de Recursos Hídricos, São Paulo.

Daley, R., 1991. Atmospheric Data Analysis. Cambridge University Press, London.

Deschamps, M. V., 2004. Vulnerabilidade Socioambiental na Região Metropolitana de Curitiba. Tese (Doutorado). Curitiba, UFPR.

Krzysztofowicz, R., 1998, Probabilistic hydrometeorological forecasts: toward a new era in operational forecasting. Bulletin of the American Meteorological Society 79, 243-251.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. Journal of Hydrology 249, 2-9.

Leite, E. A., 2008. Gestão do Valor da Informação Hidrometeorológica: O Caso dos Alertas de Inundação para Proteção de Bens Móveis em Edificações Residenciais de União da Vitória. Tese (Doutorado). Rio de Janeiro, COPPE.

Lo, A. W., 1986. Logit versus Discriminant Analysis: A Specification Test and Application to Corporate Bankruptcies. Journal of Econometrics 31, 151-178.

Lohmann, M., 2011. Regressão logística e redes neurais aplicadas à previsão probabilística de alagamentos no município de Curitiba, PR. Tese (Doutorado). Curitiba, UFPR.

Lohmann, M., 2013. Análise dos alagamentos no município de Curitiba entre os anos de 2005 a 2010. Revista Ciência Geográfica 17, 135-149.

Ludwig Jr., O., Montgomery, E., 2007. Redes Neurais: Fundamentos e Aplicações com Programas em C. Ciência Moderna, Rio de Janeiro.

Machado, M. L., Nascimento, N., Baptista, M., et al., 2007. Curva de danos de inundação versus profundidade de submersão: desenvolvimento de metodologia. Disponível: GSearch

Merz, B., Krelbich, H., Thicken, A., B. R. Schmidtke. 2004. Estimation uncertainty of direct monetary flood damage to building. Natural Hazards and Earth System Sciences 4, 153-163.

Monteiro, C. A. F., 1991. Clima e excepcionalismo: conjecturas sobre o desempenho da atmosfera como fenômeno geográfico. Editora da UFSC, Florianópolis.

Queensland Government, 2002. Guidance on the assessment of tangible flood damages. Disponível: http://www.nrw.qld.gov.au/water/use/pdf/tangible_flood_damages.pdf. Acesso: 15 jun. 2013.

Thicken, A. H., Müller, M., Merz, H. K., 2005. Flood damage and influencing factors: new insights from the August 2002 flood in Germany. Water Resources Research 41. **DOI:** 10.1029/2005WR004177

Zanela, E., 2005. O impacto das precipitações, as inundações e a percepação das comuniddades atingidas, da imprensa e dos gestores públicos: um estudo de caso no bairro Cajuri, Curitiba – PR. Tese (Doutorado). Curitiba, UFPR.

Tucci, C. E., 2005. Gestão das inundações urbanas. Global Water Partnership South America, Unesco.

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences. Academic Press San Diego, California.

Wind, H. G., Nierop, T. M., Blois, C. J., Kok, J. L. 1999. Analysis of flood damages from the 1993 and 1995 Meuse floods. Water Resources Research 35, 3459-3465.

Wmo., 2007. WMO statement on the scientific basis for, and limitations of weather and climate forecasting. in; WMO, , Elements for life, Annexe 7, Geneva, Tudor Rose, Switzerland.

Yevjevich, V., 1974. Determinism and stochasticity in hydrology. Journal of Hydrology 22, 225-237.