20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Métodos Contextuais para Classificação Baseada em Casos e Classificação por Vizinhos mais Próximos: Estudo de caso para Mapeamento de Risco de Doenças Endêmicas no Brasil

Flávia de Toledo Martins Bedê[1], Luciano Vieira Dutra[2], Sandra Sandri[3], Corina da Costa Freitas[2]

[1]Programa de Doutorado em Computação Aplicada – CAP, INPE

[2]Divisão de Processamento de Imagens – DPI, INPE

[3]Laboratório Associado de Computação e Matemática Aplicada – LAC - INPE

{flavinha, dutra, corina}@dpi.inpe.br,sandri.at.lac.inpe.br@gmail.com

Abstract. *Of all the parasitic diseases that affect humans, schistosomiasis is one of the most widespread. Considered a serious public health problem, the disease affects thousands of people in Brazil. Since the implementation of schistosomiasis control program in Minas Gerais, surveillance and control actions are being undertaken. To contribute to the control and mapping of endemic areas, this paper proposes the application of data on prevalence in districts Gerais to versions of contextual methods KNN (K-Nearest Neighborhood) and CBR (Case-Based Reasoning), also proposed this work. The contextual versions KNN, identified as* $K^{l}NN$*, and CBR, identified as CCBR, incorporate spatial information in their solution.*

Resumo. *De todas as doenças parasitárias que afetam o homem, a esquistossomose é uma das mais difundidas. Considerada um grave problema de saúde pública, a doença afeta milhares de pessoas no Brasil. Desde a implementação do programa de controle da esquistossomose no estado de Minas Gerais, ações de controle e vigilância vêm sendo realizadas. Visando contribuir com o controle e mapeamento de áreas endêmicas, este trabalho propõe a aplicação dos dados de prevalência em localidades do estado de Minas Gerais às versões contextuais dos métodos KNN (K-Nearest Neighborhood) e CBR (Case-Based Reasoning), também propostas neste trabalho. As versões contextuais do KNN, identificada como* $K^{l}NN$*, e do CBR, identificada como CCBR, incorporam na sua solução informações espaciais.*

Palavras-chave: *Mapeamento de risco, KNN, CBR.*

1. Introdução

A esquistossomose é um dos graves problemas de saúde publica que afetam milhares de pessoas em todo mundo (WHO, 1985). No Brasil, a esquistossomose é causada pelo agente etiológico *Schistosoma mansoni*, que tem como hospedeiro intermediário caramujos do gênero *Biomphalaria* (Amaral et al., 2006). O parasita utiliza a água como meio para infectar o homem (hospedeiro definitivo), que através de suas fezes infectadas contamina a água, possibilitando a infecção do caramujo e dando origem a um novo ciclo. Assim, para estudar a transmissão dessa doença, além de combinar fatores ambientais e sociais, relacionados ao caramujo e ao homem, é importante relacionar esses fatores a aspectos espaciais, visto que locais próximos a áreas endêmicas são locais de potencial risco de contaminação.

Estudos anteriores apresentam alguns modelos elaborados para o mapeamento do risco da esquistossomose *mansoni* no estado de Minas Gerais. No entanto os resultados obtidos não são tão precisos e acurados (nenhum deles chegou a 70% de acertos). Visando melhorar esses resultados, propõe-se versões estendidas (generalizadas) da metodologia de *k*-vizinhos mais próximos (*K-Nearest Neighborhood* – KNN) e da metodologia de raciocínio baseado em casos (*Case-Based Reasoning* – CBR) para favorecer a elaboração de planos de ação por parte do programa de controle da esquistossomose do estado de Minas Gerais.

O KNN é uma técnica de classificação de padrões que consiste em atribuir uma classe a um elemento desconhecido usando a classe da maioria de seus vizinhos mais próximos, segundo uma determinada distância (no espaço de atributos). A sua versão estendida, identificada como $K^{l}NN$, irá basicamente incorporar na construção do modelo a utilização de l distâncias semanticamente distintas. A construção do modelo $K^{l}NN$ será baseada na procura dos vizinhos para cada distância utilizada. Por exemplo, $K^{3}NN$ (k_1, k_2, k_3) refere-se a k_1 vizinhos mais próximos, considerando uma distância espectral (exemplo distância euclidiana), a k_2 vizinhos mais próximos, considerando uma distância geográfica (distância real entre as sedes de duas cidades por ex.) e a k_3 vizinhos mais próximos considerando uma distância no tempo.

O CBR é uma técnica que consiste em construir soluções baseando-se em soluções para problemas similares (Aamodt e Plaza, 1994). A sua versão contextual, denominada CCBR, incorporará as informações espaciais em forma de peso, de tal forma que além de se basear em soluções similares, será dado um peso maior a soluções que sejam espacialmente mais próximas. Uma outra versão do CCBR consiste em definir, através de uma distância real entre as localidades (via estrada ou rio), uma região de influência onde serão consideradas apenas as similaridades dos casos de treinamento que estiverem contidos nesta região.

O objetivo deste trabalho é desenvolver e testar novos modelos para classificação de risco da esquistossomose *mansoni* em Minas Gerais que considera diferentes tipos de distâncias, particularmente a distância geográfica.

Neste trabalho, será dada particular atenção, as versões denominadas contextuais $K^{l}NN$ e CCBR que envolverem pelo menos uma distância geográfica, como por exemplo, distância euclidiana entre as sedes de município, entre dois municípios via estrada ou entre dois municípios via rio. Assim, os modelos contextuais propostos visam incorporar a informação do espaço usando as distâncias que serão obtidas a partir da matriz de proximidade generalizada, proposta por Aguiar et al. (2003) e,

posteriormente, geradas por Fonseca (2009) baseadas na distância entre municípios, através de rede de estradas pavimentadas e a rede de rio.

Este trabalho está divido em seis seções principais. Na Seção 2 serão apresentadas algumas informações sobre a doença. Na Seção 3 serão discutidos a técnica KNN juntamente com modelo K^{l}NN proposto. Na Seção 4 será apresentada a técnica CBR *fuzzy* e o modelo contextual CCBR proposto. Na Seção 5 será apresentada a metodologia com uma breve descrição da área de estudo e dos dados disponíveis juntamente com os dados espaciais. E finalmente na ultima seção, serão apresentados os resultados esperados e o cronograma de atividades.

2. Esquistossomose mansoni

Conhecida popularmente no Brasil como barriga d'água, xistose ou doença do caramujo, a esquistossomose é uma doença transmissível e parasitária, típica de locais sem saneamento ou com saneamento básico precário. A dispersão da doença é lenta e progressiva no país. As pessoas se contaminam com *Schistosoma mansôni* (agente etiológico) através de diversos tipos de contato (trabalho, lazer, banho) com água natural infestada por cercárias que são eliminadas na água através de hospedeiros intermediários, moluscos límnicos do gênero *Biomphalaria* (*B. glabrata*, *B. tenagophila*, *B. straminea*) (Doumenge et al., 1987). Estima-se que a doença afete seis milhões de pessoas no Brasil, principalmente na região nordeste do país (Coura e Amaral, 2004). Segundo os dados apresentados no Sistema de Informação de Agravos de Notificação (SINAN) do Ministério da Saúde, de 1995 a 2005 mais de um milhão de casos positivos foram relatados, 27% deles no Estado de Minas Gerais.

Os primeiros casos de esquistossomose registrados em Minas Gerais foram observados por Teixeira (1920) na cidade Belo Horizonte. Naquela oportunidade foram examinadas 9.995 pessoas “de todas as idades e condições”, sendo os ovos de *S. mansoni* encontrados nas fezes de 49 pacientes (0,5%).

O ciclo da doença é relativamente simples, mas de enorme complexidade social, e depende da eliminação de ovos do parasita pelo homem, de coleções hídricas habitadas por moluscos suscetíveis e das necessidades cotidianas das pessoas. Porém sabe-se que a transmissão da doença é mais dependente do comportamento do homem do que do vetor, pois a infecção se dá fora de casa no contato com águas naturais contaminadas por fezes de portadores do verme.

Os sintomas mais comuns da doença são: diarréia, dores abdominais, dores pelo corpo, febre, calafrios, dores de cabeça, falta de apetite, mal estar, emagrecimento, endurecimento e aumento de volume do fígado e baço, hemorragias, vômitos negros e fezes negras. Em alguns casos mais graves, pode provocar convulsões e paralisia dos braços e/ou pernas.

O diagnóstico e o tratamento são simples, mas a erradicação da doença só é possível com medidas que interrompam o ciclo evolutivo do parasita, como mudanças no comportamento humano mediante a educação da população e melhoramento das condições básicas sociais e econômicas da comunidade envolvida (Katz e Almeida, 2003).

Dado que a transmissão da doença se deve à combinação de características ambientais relacionadas ao homem e ao caramujo, é importante relacionar esses fatores a aspectos espaciais, já que locais próximos a áreas endêmicas são locais de potencial risco de serem ou de se tornarem endêmicos. Por esse motivo e por não existirem muitos estudos relacionados com o tema, este trabalho propõe o uso de modelos contextuais que incorporam estes aspectos espaciais para auxiliar a relacionar a incidência da doença com variáveis sociais ou ambientais para determinar locais que potencialmente podem desenvolver a doença.

3. K^lNN – um modelo de vizinhos mais próximos para *l* distâncias distintas

Nesta seção será apresentada a definição do modelo KNN além de uma breve explicação sobre o modelo K^lNN proposto.

Em reconhecimento de padrões, o algoritmo KNN é um dos mais simples de todos os algoritmos de aprendizado de máquina. Baseado na analogia, um objeto é classificado pelo voto da maioria de seus vizinhos. Este processo de classificação pode ser computacionalmente exaustivo se considerado um conjunto com muitos dados. Por isso, a grande desvantagem é o tempo de computação para a obtenção dos *k* vizinhos mais próximos. Então a maioria dos estudos envolvendo KNN tem o objetivo de aumentar a eficiência computacional e reduzir a taxa de erro de generalização deste método (Bishop, 2007; Michie e Spiegelhalter, 1994; Webb, 2002).

O KNN possui apenas um parâmetro livre (o número de *k* vizinhos) que é controlado pelo usuário com o objetivo de obter uma melhor classificação. O melhor valor de *k* pode ser determinado experimentalmente. Começa-se com $k = 1$, e utiliza-se um conjunto de testes, para estimar a taxa de erro do classificador. Para cada *k*, classificam-se as tuplas do conjunto de testes e verifica-se quantas tuplas foram classificadas corretamente. O valor de *k* que apresentar a menor taxa de erro será o escolhido. Normalmente, os valores de *k* escolhidos são 1, 2, 3 ou $\sqrt{n}$, onde *n* é o tamanho da base de treinamento (Bishop, 2007; Webb, 2002).

Basicamente, uma base de dados de treinamento composta por um conjunto de tuplas $\{a_1, ..., a_n, cl\}$, onde *cl* é a classe à qual pertencem as tuplas $\{a_1, ..., a_n\}$, é usada para classificar um novo caso c_0 (representado como $c_0 = (a_1(c_0),..., a_n(c_0))$). A classificação é realizada da seguinte maneira (Theodoridis e Koutroumbas, 2006):

- Inicialmente estabelece-se um valor para k (geralmente se determina um valor impar para *k*, de forma que este valor não seja múltiplo do número total de classes);
- Calculam-se as distâncias *x* do caso c_0 às tuplas de treinamento;
- Identifica-se os *k* vizinhos mais próximos, independentemente do rótulo de classe;
- Dentro das *k*-tuplas identificadas, identificar o número de tuplas que pertencem a cada classe;
- Classifica-se o caso c_0 associando-se a ele a classe mais frequente, ou seja, a classe que a maioria das *k*-tuplas pertence.

O problema se resume então à definição de um valor para *k* e de como é calculada a distância *x*. As métricas mais comuns no cálculo de distância entre as tuplas são a distância Euclidiana (mais usada), a distância de Manhattan e a distância de Minkowski. Esta terceira é uma generalização das outras duas, podendo ser denotada por:

$$x(a, a(c)) = \left(|a_1 - a_1(c)|^q + |a_2 - a_2(c)|^q + ... + |a_n - a_n(c)|^q \right)^{1/q} \quad (1)$$

onde, $q \in N$. Quando $q = 1$ é a distância Manhattan e quando $q = 2$ é a distância Euclidiana.

Pode ser útil atribuir pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuem mais para a média do que os mais distantes. O mais comum é atribuir a cada vizinho o peso de *1/x*. Também é possível atribuir um peso relativo à importância de cada atributo (Bishop, 2007; Webb, 2002).

Como dito anteriormente, o processo de classificação pode ser computacionalmente exaustivo. A maioria das pesquisas em KNN se concentra na tentativa de aumentar a velocidade para calcular o vizinho mais próximo (Fukunaga e Narendra, 1975; Gates, 1972).

3.1. Principais elementos do modelo proposto

Como o KNN possui apenas um parâmetro livre, *k*, a proposta do K^lNN é inserir mais um parâmetro, *l*, que basicamente considerará além dos *k* vizinhos espectrais, *l* distâncias semanticamente distintas. Assim, o usuário poderá controlar mais de um parâmetro, dependendo dos dados e da aplicação.

No modelo K^lNN, onde $K^l = (k_1, k_2, ..., k_l)$ a definição de quantos vizinhos devem ser usados para cada tipo de distância (espectral, espacial e temporal, por exemplo) pode ser feita da mesma maneira que se define o valor *k*, i.e., experimentalmente. Para este estudo, serão usados dados espectrais e espaciais, usando uma distância qualquer para os atributos espectrais e uma distância geográfica (distância real, via estrada e rio) para os atributos espaciais. A classificação pelo modelo K^lNN será realizada seguindo os seguintes passos:

1. Estabelece-se um valor para K^l de acordo com o conjunto de dados, onde $K^l = (k_1, k_2, ..., k_l)$. Cada k_l, estabelece-se o número de vizinhos para cada tipo de atributo (espectral, espacial e temporal, por exemplo).

2. Calculam-se as distâncias do caso c0 às k-tuplas de treinamento para todas as distâncias usadas;

3. Identifica-se os kl vizinhos mais próximos, independentemente do rótulo de classe;

4. Dentro das kl-tuplas identificadas, identificar o número de tuplas que pertencem a cada classe;

5. Classifica-se o caso c0 associando-se a ele a classe mais frequente, dentre k1, k2, ..., kl vizinhos identificados.

Para calcular as distâncias reais podem ser usados os mesmos tipos de distâncias usados comumente para o cálculo das distâncias espectrais, porem deve-se considerar a

posição geográfica dos casos. Neste trabalho propõe-se usar as distâncias reais relativas, descritas na Seção 5.2, em que serão consideradas as conexões com estradas e/ou rios.

4. CCBR – um modelo contextual do método de raciocínio baseado em casos

Neste capítulo será apresentada a proposta do modelo CCBR e uma breve explicação sobre o modelo CBR e sobre o modelo CBR *fuzzy*, que será usado como base na geração do modelo CCBR.

O CBR pode ser considerado como um modelo baseado em similaridade ou em raciocínio analógico. O princípio básico adotado implicitamente nesta metodologia de resolução de problemas é que problemas similares possuem soluções similares. Versões mais simples do modelo seguem o princípio de que somente é plausível (mas não necessário) que problemas similares tenham soluções similares. Em Armengol et al (1994), os autores propõem um novo algoritmo *Fuzzy* CBR para a classificação, em que este princípio é estudado, no contexto de classificação (Dubois et al., 1998), como "quanto mais similares são as descrições de dois casos, é mais possível que a classificação seja similar".

Basicamente, neste contexto, uma base de dados de CBR é composta de casos da forma c = (*d*, *cl*), onde *d* é o caso modelado descrito como um vetor de valores para um conjunto de atributos e *cl* sua classe associada. A atribuição de uma classe para um novo caso descrito d_0 dependerá da similaridade entre d_0 com os casos descritos na base de dados. Na abordagem proposta em Armengol et al. (1994), esta similaridade é calculada como uma média ponderada em função da similaridade entre cada atributo dos casos descritos no conjunto de dados de aprendizagem. Nesta abordagem, os vetores de peso são usados para minimizar os erros de classificação dos casos já contidos na base.

4.1. CBR fuzzy baseado em similaridade

Antes de entrar em mais detalhes, deve-se assumir um conjunto de casos resolvidos em uma base de casos *CB*, em que um caso é representado por uma tupla (completa) de valores de atributos que descrevem a situação ou problema a resolver, juntamente com uma classe de solução ou resultado. Em outras palavras, denota-se como $\mathbf{A} = \{a_1, ..., a_n\}$ um conjunto de atributos descritos e *cl* o atributo classe. Além disso, os domínios dos atributos a_i e classe são denotados por $D(a_i)$ e $D(classe)$, respectivamente (então $D(classe)$ é o conjunto de classes de solução). Em seguida, um caso $c \in CB$ será representado como um par $c = (d, cl)$, onde $d = (a_1(c),..., a_n(c))$ é uma *n*-tupla com os valores de descrição do problema e $cl = classe(c)$ é a classe de solução para o caso *c*. Se escrevermos $\mathbf{D} = D(a_1) \times ... \times D(a_n)$ (**D** para descrições) e **Cl** = D(*classe*), então uma base de casos *CB* é apenas um subconjunto de **D**×**Cl**. As definições deste texto usarão essas notações.

Neste método, tomando uma base de casos $CB = (c_i = (d_i, cl_i))_{i \in I}$ e um novo problema descrito *d**, a tarefa CBR é encontrar (prever) uma classe *cl** para *d**, aplicando o princípio geral, descrito acima, de alguma forma, ou seja, tendo em conta a similaridade de *d** com casos já resolvidos $c_i \in CB$.

Em geral, uma relação de similaridade fuzzy em um domínio Ω é um mapeamento $S : \Omega \times \Omega \to [0,1]$, que atribui a cada par de elementos (*w*, *w'*) um número que mede a similaridade entre *w* e *w'* (*S*(*w*, *w'*)), de acordo com alguns critérios

definidos, de forma que quanto maior $S(w, w')$, maior a semelhança entre eles. Em particular, $S(w, w') = 1$ significa que w e w' são indistinguíveis, enquanto $S(w, w') = 0$ significa que w e w' não têm nada em comum. Pode-se também entender $\delta(w, w') = 1 - S(w, w')$ como um tipo de distância entre w e w'. Essas funções exigem as propriedades de reflexividade e simetria (Dubois e Prade, 2000), ou seja, $S(w, w) = 1$ e $S(w, w') = S(w', w)$, para qualquer $w, w' \in \Omega$. Neste trabalho, serão consideradas como relações de similaridade as relações binárias fuzzy reflexiva e simétrica.

4.2. Principais elementos do modelo proposto

As etapas principais do método proposto são: primeiramente, a definição de uma relação de similaridade S_D adequada entre as descrições de casos na base de casos *CB* e uma descrição do problema arbitrário; e em seguida ou será atribuído um peso maior às descrições de casos que estiverem mais próximos geograficamente do caso particular ou serão considerados apenas as similaridades dos casos que estiverem a uma distância geográfica máxima definida.

Para a definição da relação de similaridade S_D adequada, são necessárias informações adicionais para avaliar a relevância (peso) de cada atributo para a recuperação de um caso particular. A relação de similaridade fuzzy induzida S_D^w: **D** x **D** $\rightarrow$ [0, 1] sobre as descrições de casos, é definida como segue:

$$S_D^w(d_1, d_2) = \sum_{a \in A} w(a).S_a\left(a(d_1), a(d_2)\right) \quad (2)$$

usando a notação $d_1 = (a_1(d_1), \ldots, a_n(d_1))$ e $d_2 = (a_1(d_2), \ldots a_n(d_2))$.

Note que, S_D^w é definida de acordo com a relação de similaridade descrita na Seção 4.1, ou seja, é reflexiva e simétrica.

Uma vez definida a relação de similaridade S_D^w, é possível definir como solução adequada uma classe *cl* para um problema descrito *d**, apenas comparando *d** com as descrições de todos os casos em *CB* (ou comparando com as descrições dos casos que estiverem dentro de uma região de influência, determinada por uma distância geográfica definida), associando a solução de classe *cl*, agregando todos os valores similares como feito em Martins-Bedê et al. (2009), mas agora dando um peso maior aos casos que estiverem mais próximos geograficamente de *d**. De acordo com Calvo et al. (2002), dependendo da aplicação, as funções de agregação adequadas podem ser, por exemplo, funções disjuntiva, como máxima ou outra, ou alguns tipos de funções de média, como quase-médias aritméticas ou funções de agregação mais sofisticadas ainda. Então, dada uma base de casos *CB*, um conjunto de atribuições de pesos $W = \{w_c\}_{c \in CB}$ e uma função de agregação adequada *F* em [0, 1] que considera a distância geográfica (no caso, de estrada ou rio), é possível designar a solução *cl* para uma descrição do caso $d^* \in$ **D.**

4.3. Determinando o peso para cada caso

Em seguida, é descrito como atribuir um peso adequado para cada caso na base *CB*. Então, supondo que um caso $c_0 \in CB$ é conhecido e fixado ao longo do processo. Na verdade, o mesmo processo que é descrito abaixo para c_0 será aplicado para cada caso do *CB*. Naturalmente, para cada caso c_0 será atribuído um peso correspondente w_0.

Para fazer isso, além do caso c_0, será necessário fixar um subconjunto de casos $LS_0 \subseteq CB$, ou seja, uma coleção de descrições cuja classe é conhecida. Então o problema de determinação do peso em relação ao c_0 é determinar uma atribuição de peso w_0: $\mathbf{A} \rightarrow [0, 1]$ tal que, para cada caso $c = (d, cl) \in LS$, a similaridade entre d_0 e d, $S_D^w(d_0, d)$, aproxima-se tanto quanto possível da similaridade entre as classes cl_0 e cl, denotada por $S_{Classe}(cl_0, cl)$.

É possível reformular o problema usando o quadrado da diferença para medir a divergência entre as duas similaridades (a similaridade entre as duas descrições de caso e a similaridade entre suas classes), para que o problema de determinação do peso em relação a c_0 seja encontrar os valores de $w(a_1)$, ..., $w(a_n)$ que minimizam a seguinte expressão:

$$\sum_{(d,cl)\in LS_0} \left[S_D^W(d, d_0) - S_{Classe}(cl, cl_0) \right]^2 \qquad (3)$$

sujeita às seguintes restrições sobre w_{c_0} :

(1) $\sum_{a\in A} w(a) = 1$, e

(2) $w(a) \geq 0$ para todo $a \in \mathbf{A}$.

Para resolver este problema de minimização será aplicado o algoritmo apresentado em Torra (2000) com a extensão descrita em Torra (2002). É importante ressaltar que esse algoritmo, como os similares existentes na literatura, deixa de dar uma (única) solução quando existe dependência linear entre as colunas da matriz de dados. Na realidade, como mostrado em Torra (2002), o problema só surge quando há uma coluna/ atributo a_i que pode ser escrito como uma combinação linear dos outros na forma $a_i = \Sigma_j p_j a_j$ de tal forma que $\Sigma_j p_j = 1$. Neste caso, ao retirar uma das colunas linearmente dependentes temos o mesmo mínimo que teríamos ao se considerar todos os atributos. Portanto, uma abordagem alternativa é considerar a mesma quantidade de subproblemas e atributos dependentes, onde cada subproblema corresponde a um original depois de remover um dos atributos dependentes. A solução com um erro mínimo corresponderia à solução do problema original.

5. Metodologia

Os modelos propostos neste trabalho serão usados para classificar a prevalência da esquistossomose em localidades de Minas Gerais, usando variáveis derivadas de Sensoriamento Remoto (SR), climáticas, e socioeconômicas. Sendo assim, nesta seção serão apresentadas a área de estudo com uma breve descrição dos dados que serão usados, juntamente com uma breve explicação dos dados espaciais, as distâncias via estradas e/ou via rios.

5.1. Área de estudo/ Dados disponíveis

Em Minas Gerais a distribuição da esquistossomose *mansoni* não é regular, intercalando-se em áreas de maior prevalência com outras onde a transmissão é baixa ou nula (Figura 1). A doença é endêmica nas regiões norte (compreendendo as zonas do Médio São Francisco e Itacambira), oriental e centro (zonas do Alto Jequitinhonha, Metalúrgica, Oeste e Alto São Francisco). Os maiores índices de infecção são

encontrados nas regiões nordeste e leste do Estado que compreendem as zonas do Mucuri, Rio Doce e da Mata (Carvalho et al., 2005; Pellon e Teixeira, 1950).

Os pontos indicados na Figura 1 representam as 1583 localidades de onde se tem informação positiva sobre a existência da esquistossomose em 1216 localidades. Destas, 565 possuem até 5% de prevalência, 404 possuem prevalência entre 5% e 15 % e finalmente, 247 possuem mais de 15% de prevalência (faixas de prevalências definidas como baixa, média e alta, respectivamente, de acordo com classificação da Secretaria de Saúde). Estes dados foram cedidos pela Secretaria de Estado de Saúde de Minas Gerais e correspondentes à prevalência da esquistossomose que será usada como sendo variável dependente.

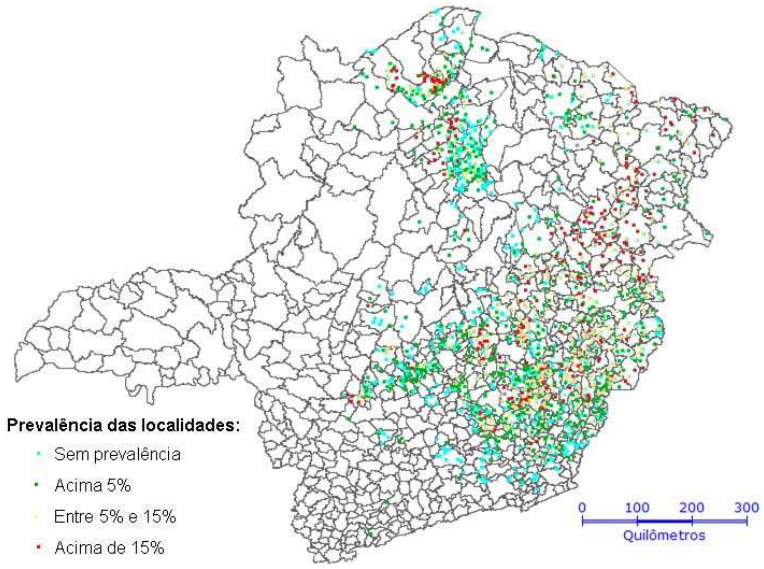

Figura 1. Localidades com prevalência da esquistossomose no estado de Minas Gerais. Fonte: Secretaria da Saúde de Minas Gerais

Para explicar a prevalência da doença, a análise conterá variáveis extraídas de SR (obtidas pelos sensores MODIS e SRTM), variáveis climáticas e variáveis socioeconômicas.

Das variáveis de SR derivadas do MODIS, serão usadas: banda azul (Blue), vermelho (Red), infravermelho próximo (NIR), e infravermelho médio (MIR), os índices de vegetação melhorado (EVI), índice de vegetação da diferença normalizada (NDVI), e os índices derivados do modelo linear de mistura espectral, vegetação (Veg), solo (Solo) e sombra (Somb).

Das variáveis obtidas através do SRTM, serão usadas o modelo digital de elevação (DEM) e a declividade (Dec), derivada do DEM. Das variáveis climáticas, serão usadas a precipitação acumulada (Prec), a temperatura mínima (Tmin) e a temperatura máxima (Tmax). Em Martins (2008) pode-se obter uma descrição mais detalhada das variáveis de SR e climáticas. No entanto, é importante ressaltar que para este trabalho serão usados ou o pixel referente à localidade ou uma média dos pixels contidos em uma região (*buffer*) ao redor da localidade.

Já para as variáveis socioeconômicas serão usadas as informações referente a setores censitários disponíveis no Sistema Nacional de Indicadores Humanos (SNIU, 2005). No total são 3.216 variáveis com informações sobre:

- os domicílios (tipo de domicilio, quantidade de moradores (total, por sexo, por faixa etária, por nível de escolaridade, com receita), tipo de abastecimento de água, quantidade de banheiros, tipo de saneamento, coleta de lixo);
- a quantidade de pessoas que vivem no setor (total, casal, filhos, por sexo, por tipo de domicilio, por faixa etária, por nível de escolaridade, com receita, por tipo de abastecimento de água, por quantidade de banheiro, por tipo de saneamento, por tipo coleta de lixo);
- o responsável pelo domicílio (por faixa etária, por escolaridade, por renda, por sexo, por quantidade de dependente).

5.2. Dados espaciais

As informações espaciais que se pretende usar para geração dos modelos contextuais serão obtidas através do algoritmo desenvolvido por Fonseca (2009). Baseada em Aguiar et al (2003), Fonseca (2009) gerou duas matrizes de proximidade generalizada (*Generalized Proximity Matrix – GPM*), uma para rede de estradas e outra para rede de rios.

De acordo com Aguiar et al (2003), a GPM descrita em seu trabalho, é uma extensão da matriz de pesos espaciais utilizada em muitos métodos de análise espacial (Bailey e Gattrel, 1995), onde as relações espaciais são calculadas levando-se em conta não apenas relações do espaço absoluto (tais como distância euclidiana), mas também relações do espaço relativo (rede de transporte). Usando a GPM, dois objetos geográficos que são vizinhos de fronteira podem não ser considerados como vizinhos se não houver alguma rede de comunicação (por ex. rede de estradas, rede de rios) entre eles; mas se os objetos geográficos forem conectados por uma rede, mesmo que estejam separados por milhares de quilômetros, serão considerados vizinhos.

Essa relação de vizinhança se torna muito importante para este estudo por vários motivos, entre eles pode-se citar dois motivos principais. Primeiro porque localidades onde passa um rio que possui prevalência baixa ou nula, mas que rio acima existe outra localidade com alta prevalência da doença e saneamento precário ou inexistente, o risco de contaminação da doença para a primeira localidade aumenta. Em vários locais menos favorecidos de Minas Gerais existe o hábito da população de se refrescar em rios e lagos. Então o segundo motivo principal se refere a localidades que possuem fácil acesso (rede de estradas) a outra localidade que possui um rio (ou lago) contaminado pela larva do *Shistossoma*, a probabilidade de que os moradores de localidades próximas (conectados por estradas) se refresquem e se contaminem também aumenta.

5.3. Aplicação

Através do conjunto de dados apresentado na Seção 5.1, primeiramente, serão gerados dois modelos de regressão linear, global e regional (conforme Martins (2008)) com dois objetivos, usar as variáveis (atributos) selecionadas pelos modelos de regressão global e regional para a aplicação nos modelos K^{l}NN e CCBR propostos nas Seções 3 e 4

respectivamente e para usar como meio de avaliação da precisão dos modelos contextuais propostos.

6. Resultados esperados

Pretende-se gerar dois algoritmos $K^{l}NN$ e CCBR para serem aplicados ao estudo de caso aqui proposto, podendo ser aplicados a qualquer outro problema que tenha dependência espacial.

Referencias

Aamodt, A.; Plaza, E. Case-based reasoning: foundational issues, methodological variations and system approaches. AI Communications, v. 7, p. 39–59, 1994.

Aguiar, A. P. D.; Câmara, G.; Monteiro, A. M. V.; Souza, R. C. M. Modeling spatial relations by generalized proximity matrices. In: Simpósio Brasileiro de Geoinformatica 5. , 2003, Campos do Jordão. p. 8.

Amaral, R. S.; Tauil, P. L.; Lima, D. D.; Engels, D. An analysis of the impact of the schistosomiasis control programme in Brazil. Mem Inst Oswaldo Cruz, v. 101, p. 79-85, 2006.

Armengol, E.; Esteva, F.; Godo, L.; Torra, V. On learning similarity relations in fuzzy case-based reasoning. Lecture Notes in Computer Science, v. 3135, p. 14–32, 1994.

Bailey, T.; Gattrel, A. Spatial data analysis by example. London: Longman, 1995.

Bishop, C. M. Pattern Recognition and Machine Learning Springer, 2007. 738 p.

Calvo, T.; Kolesarova, A.; Komornikova, M.; Mesiar, R. Aggregation operators: Properties, classes and construction methods. Aggregation operators: new trends and applications. v. 97. Publisher Physica-Verlag GmbH, 2002, p. 3-104.

Carvalho, O. S.; Dutra, L. V.; Moura, A. C. M.; Freitas, C. C.; Amaral, R. S.; Drummond, S. C.; Freitas, C. R.; Scholte, R. G. C.; Guimarães, R. J. P. S.; Melo, G. R.; Correia, V. R. M.; Guerra, M. Desenvolvimento de um sistema de informações para o estudo, planejamento e controle da esquistossomose no Estado de Minas Gerais. In: Simpósio Brasileiro de Sensoriamento Remoto, 2005, Goiânia. INPE, 16-21 abr. 2005. p. 2083-2086. ISBN 85-17-00018-8.

Coura, J. R.; Amaral, R. S. Epidemiological and control aspects of schistosomiasis in Brazilian endemic areas. Mem. Inst. Oswaldo Cruz, v. 99 (Suppl. !), p. 13-19, 2004.

Doumenge, J. P.; Mott, K. E.; Cheung, C.; Villenave, D.; Capui, O.; Perrin, M. F., 1987, Atlas of the global distribution of schistosomiasis, Bordeaux, France, Universitaires de Bordeaux Press

Dubois, D.; Esteva, F.; Garcia, P.; Godo, L.; L`opez de Mantaras, R.; Prade, H. Fuzzy Set Modelling in Case-based Reasoning. International Journal of Intelligent Systems, v. 13, n. 4, p. 345–373, 1998.

Dubois, D.; Prade, H. E. Fundamentals of Fuzzy Sets, The Handbooks of Fuzzy Sets Series Dordrecht, 2000. 672 p.

Fonseca, F. R. Modelagem espacial da esquistossomose mansoni no estado de Minas Gerais, utilizando a conectividade de redes via estradas e rios. 2009. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

Fukunaga, K.; Narendra, P. M. A branch and bound algorithm for computing nearest neighbours. IEEE Trans. Comput, p. 917–922, 1975.

Gates, G. W. The reducednearestneighbour rule. IEEETransactions on InformationTheory, p. 431, 1972.

Katz, N.; Almeida, K., 2003, Esquistossomose, xistosa, barriga d'água, Ciencia e Cultura, p. 38-43.

Martins-Bedê, F. T.; Godo, L.; Sandri, S.; Dutra, L. V.; Freitas, C. C.; Carvalho, O. S.; Guimarães, R. J. P. S.; Amaral, R. S. Classification of Schistosomiasis Prevalence Using Fuzzy Case-Based Reasoning. . Lecture Notes in Computer Science, v. 5517, p. 1053-1060, 2009.

Martins, F. T. Mapeamento do risco da esquistossomose no estado de Minas Gerais, usando dados ambientais e sociais. 2008. 145 p.MSc em Computação Aplicada) – INPE, São José dos Campos.

Michie, D.; Spiegelhalter, D. J. Machine Learning, Neural and Statistical Classification Prentice Hall, 1994. 289 p.

Pellon, A. B.; Teixeira, I. Distribuição da esquistossomose mansônica no Brasil. In: Congresso Brasileiro de Higiêne, 1950, Recife. p. 117.

SNIU, 2005, Sistema Nacional de Indicadores Urbanos (SNIU), [database on the Internet].

Teixeira, M. J. A schistosomose mansônica na infância em Belo Horizonte. 1920. 107 p. Tese (Concurso à Cadeira em Pediatria da Universidade Federal de Minas Gerais) – Imprensa Oficial, Belo Horizonte.

Theodoridis, S.; Koutroumbas, K. Pattern recognition Academic Press, 2006. 885 p.

Torra, V. On the learning of weights in some aggregation operators: the weighted mean and OWA operators. Math. and Soft Comp, v. 6, p. 249-265, 2000.

___. Learning weights for the quasi-weighted means. IEEE Trans. On Fuzzy Systems, v. 10, n. 5, p. 653–666, 2002.

Webb, A. R. Statistical Pattern Recognition. 2nd. London, U.K.: Wiley, 2002. 496 p.

WHO. The Control of Schistosomiasis. Technical Report Series, v. 728, p. 113, 1985.