UFRRJ
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

MODELAGEM ESPACIAL DOS DADOS DE FASCIOLA HEPÁTICA BOVINA NO SUL DO ESPIRITO SANTO

BRUNA ALVES SANTIAGO

2015

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

MODELAGEM ESPACIAL DOS DADOS DE FASCIOLA HEPÁTICA BOVINA NO SUL DO ESPIRITO SANTO

BRUNA ALVES SANTIAGO

Sob a Orientação do Professor
Dr. Wagner de Souza Tassinari
e Co-orientação do Professor
Dr. Marcelo Dib

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Modelagem Matemática e Computacional**, no curso de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Estatística.

Seropédica, RJ
Setembro de 2015

519.7
S235m
T

Santiago, Bruna Alves, 1988-
Modelagem espacial dos dados de fasciola hepática bovina no sul do Espírito Santo / – 2015.
59 f.: il.

Orientador: Wagner de Souza Tassinari.
Dissertação (mestrado) – Universidade Federal Rural do Rio de Janeiro, Curso de Pós-Graduação em Modelagem Matemática e Computacional.
Bibliografia: f. 46-49.

1. Programação (Matemática) – Teses. 2. Programação heurística – Teses. 3. Epidemiologia veterinária – Espírito Santo (Estado) – Teses. 4. Bovino – Parasito – Espírito Santo (Estado) – Teses. 5. Doenças parasitárias – Espírito Santo (Estado) – Teses. I. Tassinari, Wagner de Souza, 1976-. II. Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Modelagem Matemática e Computacional. III. Título.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM MATEÁTICA E COMPUTACIONAL

BRUNA ALVES SANTIAGO

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Modelagem Matemática e Computacional**, no curso de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Estatística.

DISSERTAÇÃO APROVADA EM _____ / _____ / _____

_____
Wagner de Souza Tassinari. Dr. UFRRJ
(Orientador)

_____
Izabel Cristina dos Reis. Dr. FIOCRUZ

_____
Alba Regina Moretti. Dr. UFRRJ

DEDICATÓRIA

À minha amada mãe Lucimei Alves de Araujo.
(in memorian)

AGRADECIMENTOS

A Deus, por me dar forças nos piores momentos.

Às minhas irmãs Lohane, Kátia, Keila, ao meu irmão Junior, ao meu esposo Diego, ao meu pai Adriano, aos meus sogros Janete e Sebastião, as minhas irmãs de alma Thalyta, Bia e Anne, aos meus amigos e familiares em geral pela paciência, apoio, cuidado e compreensão.

À minha prima professora Drª Adria Lyra pela inspiração.

Ao meu orientador prof. Dr. Wagner de Souza Tassinari pela seriedade com a qual conduziu esse trabalho, pelos ensinamentos e pela confiança depositada em mim.

Ao meu co-orientador prof Dr. Marcelo Dib Cruz pelos ensinamentos e pela paciência.

Ao PPGMMC pela oportunidade de concluir mais uma etapa em minha vida profissional.

A Fundação de Amparo à pesquisa do Estado do Rio de Janeiro (FAPERJ) pela concessão da bolsa.

A profª Dr Isabella Martins (UFES) pela confiança em me ceder os dados para a elaboração deste trabalho.

À todas as pessoas que direta ou indiretamente colaboraram para que o presente trabalho fosse concluído.

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO GERAL

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda que parasita o fígado e as vias biliares de bovinos e ovinos, mas diagnosticada também em caprinos, equinos, búfalos, humanos e animais silvestres, como a capivara e ratão do banhado . Atualmente, a fasciolose tem se tornado um grave problema econômico para os produtores da região Sul do estado do Espírito Santo e possivelmente para toda a extensão territorial do estado, tendo em vista a possibilidade de contaminação de rebanhos bovinos, ovinos, caprinos e bubalinos pelo transporte de animais parasitados e pela presença do hospedeiro intermediário em diversas bacias hidrográficas da região . A área deste estudo foi representada por 23 municípios do Sul do estado do Espírito Santo. A unidade de análise foram propriedades que tem como objetivo econômico a pecuária bovina, totalizando 115 propriedades, 5 em cada município. O inquérito epidemiológico nessas propriedades foi feito durante o período compreendido entre 2009 e 2011. Os objetivos deste trabalho são: verificar a distribuição geográfica da fascíola no sul do estado do Espírito Santo verificando a existência de aglomerados, analisar os possíveis fatores que possam estar associados com a patologia (fatores de risco) e propor e discutir algumas técnicas de modelagem matemática baseadas em heurística na detecção de aglomerados espaciais nas propriedades acometidas com fascíola hepática. Foi verificado que a prevalência média de fascíola nas propriedades estudadas foi de 19,52%[13,41%;27,35%] e o coeficiente de variação foi estimado em de 8,24%. Foi verificado através da estatística de *kernel* que a maior intensidade de ocorrência da fascíola está na região central do estudo. Através do cálculo do índice de correlação de Moran, entre os municípios da área de estudo, foi observado o valor de 0,443 *(p-valor* < 0,001) indicando uma autorrelação espacial significativa entre as áreas estudadas, garantindo a formação de aglomerados geográficos. Após análise dos modelos e da revisão da literatura , é possível afirmar que o modelo GAM logístico multivariado é o modelo mais parcimonioso para o estudo da fascíola e seus fatores de risco. Desta forma é possível observar que as variáveis outros hospedeiros e casos anteriores são caracterizadas como fatores de risco epidemiológico para a fascíola. Para detecção dos aglomerados geográficos foram utilizadas duas metodogias, *k-means* com *elbow* e algoritmo genético com função silhueta, ambas foram significas mas para estes dados a segunda se mostrou mais precisa que a primeira, retornando o valor de 5 (cinco) clusters (aglomerados) e mostrando que o cluster de maior concentração é o localizado na região central do estudo. Tal resultado valida o que foi encontrado na estatística de *kernel.* Assim concluímos que as propriedades pertencentes ao cluster 1 (um) necessitam de um atendimento prioritário.

Palavras-chave: Epidemiologia Veterinária. Fasciolose. Modelagem Matemática.

GENERAL ABSTRACT

The fascioliasis is a liver disease caused by the parasite Fasciola, trematoda that parasite liver and biliary tract of cattle and sheep, but also diagnosed in goats, horses, buffalo, human and wild animals such as capybara and eagle rays bathed . Currently, fascioliasis has become a serious economic problem for producers from southern state of Espirito Santo and possibly for the entire land mass of the state, in view of the possibility of contamination of cattle herds, sheep, goats and buffaloes for shipping of infected animals and the intermediate host presence in several river basins in the region. The study area was represented by 23 municipalities in the southern state of Espirito Santo. The unit of analysis were property whose economic goal bovine livestock, totaling 115 properties, five in each municipality. The epidemiological survey in these properties was made during the period between 2009 and 2011. The objectives of this study are: to determine the geographical distribution of the fluke in the southern state of Espirito Santo checking for clusters, analyze the possible factors that may be associated with pathology (risk factors) and to propose and discuss some techniques of mathematical modeling based on heuristics to detect spatial clusters in the affected properties with liver fluke. It was found that the average prevalence of the fluke properties studied was 19.52% [13.41%, 27.35%] and the variation coefficient was estimated at 8.24%. It was verified by the kernel statistics of occurrence of the highest intensity in the central fluke is the area under study. By calculating the Moran's correlation coefficient between the municipalities of the study area, the value of 0.443 (p-value <0.001) was observed indicating a significant spatial autocorrelation between the areas studied, ensuring the formation of geographical clusters. After analysis of the models and the literature review, it can be mean that the GAM multivariate logistic model is the most parsimonious model for the study of fluke and its risk factors. This way you can see that variables other hosts and previous cases are characterized as epidemiological risk factors for fluke. For detection of geographic clusters were used two approachs, k-means with elbow and genetic algorithm with silhouette function, both were significant but to this data the second proved more accurate than the first, returning the value of 5 (five) clusters and showing the cluster with the highest concentration is located in the center of the study area. This result validates that was found in the kernel statistics. Thus we conclude that the properties belonging to the cluster 1 (one) require priority attention.

Keywords: Veterinary Epidemiology. Fasciolosis. Mathematical modeling.

SUMÁRIO

INDRODUÇÃO GERAL

Estudos a respeito da pecuária bovina de corte/leiteira vem sinalizando a ocorrência de fascíola hepática no Brasil e no mundo (BORAY, 1966; SERRA-FREIRE, 1995; PILE et al.,1999; LESSA et al., 2000; GOMES et al., 2002).

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda (Figura1) que parasita o fígado e as vias biliares de bovinos e ovinos, mas diagnosticada também em caprinos, equinos, búfalos (PILE et al., 2001), humanos (CORAL et al., 2007; DITTIMAR et al., 2005) e animais silvestres, como a capivara e ratão do banhado (EL KOUBA, 2005). Causa em seus hospedeiros condenação do fígado, perda de peso, anemia e outros sinais inespecíficos, o que torna o diagnóstico clínico da doença difícil fazendo-se necessária a realização do diagnóstico laboratorial (ECHEVARRIA, 1985), baseado na observação de ovos de *Fascíola hepática* nas fezes dos animais (BORAY, 1977; KLEIMAN et al., 2005) e de humanos.

Figura1: Parasito fascíola hepática. (Fonte: Wikipedia.org)

Em humanos, os sinais e sintomas são vários, diferindo conforme a fase e duração e o número de parasitas. Na fase adulta pode ocorrer dor abdominal, febre, vômito, diarréia, urticária, má digestão e absorção, icterícia, hepatomegalia e alterações de enzimas hepáticas, leucocitose e eosinofilia. Na fase crônica, os sinais e sintomas mais evidentes são os relacionados com a obstrução biliar intermitente e inflamação (ACHA e SZYFRES, 1986; GUIMARÃES, 2007).

A ocorrência desta parasitose está ligada a presença de moluscos do gênero *Lymnaea*, hospedeiro intermediário, bem como de hospedeiros definitivos parasitados (ovinos e bovinos principalmente), os quais são disseminadores de ovos (MATTOS et al., 1997). Além disso, presença de áreas alagadas favorece a manutenção do ciclo do parasito por ser ambiente adequado para o desenvolvimento do molusco (MUNGUÍA-XÓCHIHUA et al., 2007) (Figura 2).

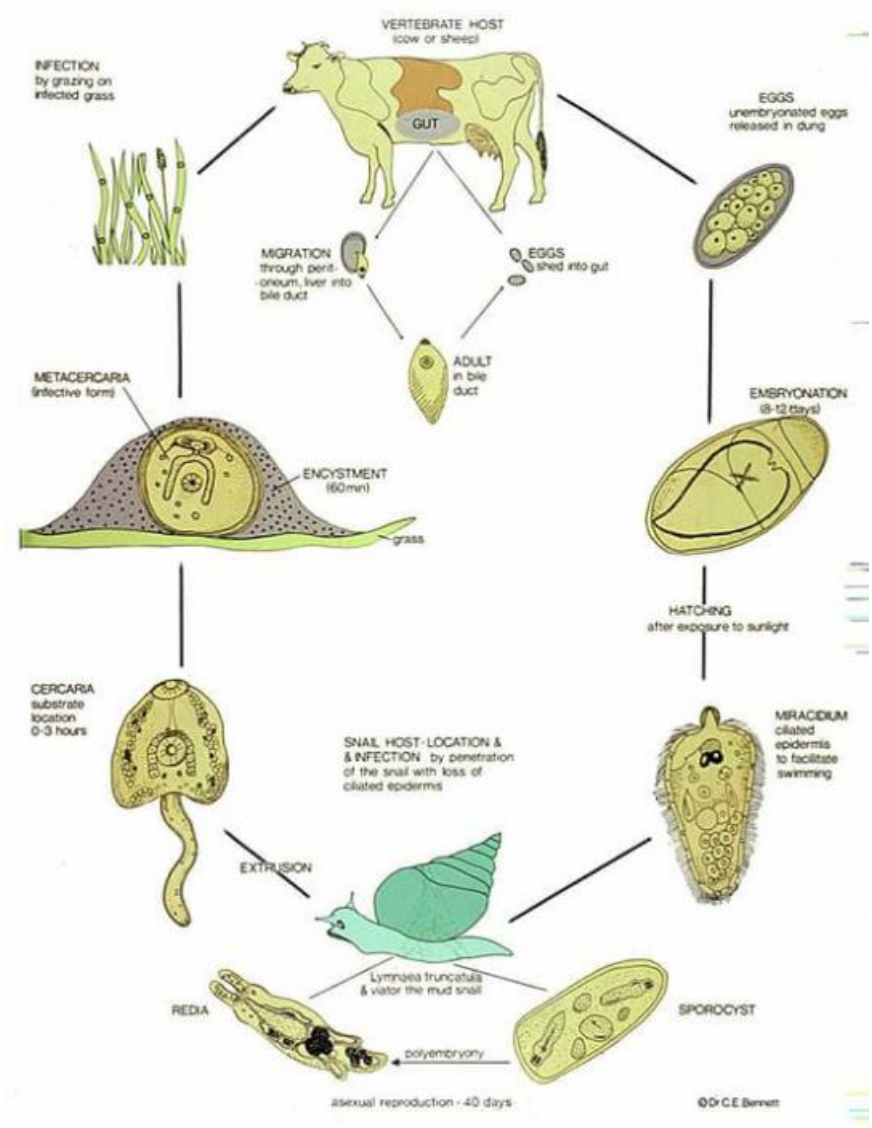


Figura 2: Ciclo biológico da fascíola hepática.

A fasciolose é uma doença de difícil controle, seja pelos medicamentos disponíveis que, na grande maioria, atingem apenas as formas adultas do parasito, seja pela difícil aceitação dos proprietários no uso de drenagens em áreas alagadas e de molusquicidas, hoje práticas também condenadas por leis ambientais. Assim, o estudo epidemiológico da doença é essencial para o planejamento do controle da fasciolose em cada região (MARTINS, 2007).

A fasciolose encontra-se em ampla distribuição mundial e a sua dispersão geográfica vêm aumentando com o passar dos anos devido à transferência de animais parasitados de locais onde a doença é enzoótica para localidades indenes (REID e DARGIE, 1995).

No Brasil a enfermidade relaciona-se historicamente aos Estados do Rio Grande do Sul, Santa Catarina, Paraná, São Paulo, Rio de Janeiro e Minas Gerais, sendo reportados como incidentes de fasciolose bovina. Nos estados das regiões Sudeste e Sul do Brasil foram notificados casos de doença crônica, com exceção do Estado do Espírito Santo (SERRA-FREIRE, 1995). Serra-Freire (1995), afirma que as áreas epidemiologicamente reconhecidas quanto à presença da F. hepática continuam a representar importantes regiões endêmicas e que o parasito está se difundindo sem as devidas notificações de sua existência.

Ao longo dos anos, em conjunto com a epidemiologia humana, a epidemiologia veterinária vem se destacando como uma área do conhecimento humano de bastante importância na saúde pública, na economia e no que diz respeito ao bem estar animal. Na maioria das vezes, os agravos na saúde animal ocorrem devido as consequências do manejo humano e devido aos fatores ambientais e genéticos. Para estudar mais a fundo estes possíveis fatores, dentro da área da epidemiologia são usadas ferramentas matemáticas, estatísticas e computacionais de análise de dados. Tais ferramentais são baseadas em técnicas de modelagem, pois tenta representar o cenário do fenômeno de maneira mais realística o possível (DYM e IVEY, 1980)

Quando o fenômeno analisado apresenta alguma realção com o espaço geográfico, pode ser utilizado a modelagem espacial (GAETAN e GUVON, 2010). Tal modelagem utiliza métodos científicos para a coleta, descrição, visualização e análise de dados que possam ser modelados como processos estocásticos, onde o espaço índice é um conjunto de dimensão maior que um. Usualmente os processos estocásticos são coleções de variáveis aleatórias $\{X_t ; t \epsilon T\}$ onde o espaço do índice t é um subconjunto da reta discreto ou contínuo. É a área que estuda os fenômenos ao longo do espaço.

Para verificar ou alocar agrupamentos de indivíduos, utilizaremos modelagem matemática baseada em heurística (RAYWARD-SMITH, 1996). Heurística é um método ou processo criado com o objetivo de encontrar soluções para um problema, é a denominação para o algoritmo que fornece soluções sem um limite formal de qualidade, tipicamente avaliado empiricamente em termos de complexidade (média) e qualidade das soluções.

JUSTIFICATIVA

Atualmente, a fasciolose tem se tornado um grave problema econômico para os produtores da região Sul do estado do Espírito Santo e possivelmente para toda a extensão territorial do estado, tendo em vista a possibilidade de contaminação de rebanhos bovinos, ovinos, caprinos e bubalinos pelo transporte de animais parasitados e pela presença do hospedeiro intermediário em diversas bacias hidrográficas da região (FRAGA, 2008).

Nos Estados de Minas Gerais, Goiás, Bahia e Espírito Santo estão surgindo novas áreas de ocorrência de F. hepática, caracterizadas pela presença dos hospedeiros intermediários e vertebrados naturalmente infectados. Bernardo et al. (2007) registraram prevalências de fígados de bovinos condenados por fasciolose no sul do Espírito Santo e encontraram percentuais médios variando de 15,24% a 28,57% entre 2006 e 2009, tendo as perdas econômicas devido à condenação de fígados sido consideradas altas.

Alguns trabalhos (AVELAR et al., 2014; FREITAS, 2013; AVELAR et al., 2012) sugerem que alguns municípios no sul do estado do Espirito Santo, são consideradas regiões endêmicas quanto à presença da F. hepática devido a difusão desses parasitas sem as devidas notificações de sua existência. Acarretando um grande problema na área financeira para a região, com a perda de rebanhos e de grande risco a saúde pública humana.

Esta dissertação será apresentada em dois capítulos, cada um abordando um objetivo que veremos a seguir.

OBJETIVOS GERAIS

Verificar a distribuição geográfica da fascíola no sul do estado do Espírito Santo, analisar os possíveis fatores que possam estar associados com a patologia (fatores de risco) e

propor métodos baseados em modelagem matemática para verificar a ocorrência de possíveis aglomerados (surtos) de fascíola hepática bovina no sul do estado do Espírito Santo.

CAPÍTULO I

ANÁLISE ESPACIAL DA DISTRIBUIÇAO DA FASCÍOLA HEPÁTICA BOVINA NO SUL DO ESPÍRITO SANTO

1 RESUMO

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda que parasita o fígado e as vias biliares de bovinos e ovinos, mas diagnosticada também em caprinos, equinos, búfalos, humanos e animais silvestres, como a capivara e ratão do banhado . Atualmente, a fasciolose tem se tornado um grave problema econômico para os produtores da região Sul do estado do Espírito Santo e possivelmente para toda a extensão territorial do estado, tendo em vista a possibilidade de contaminação de rebanhos bovinos, ovinos, caprinos e bubalinos pelo transporte de animais parasitados e pela presença do hospedeiro intermediário em diversas bacias hidrográficas da região. A área deste estudo foi representada por 23 municípios do Sul do estado do Espírito Santo. A unidade de análise foram propriedades que tem como objetivo econômico a pecuária bovina, totalizando 115 propriedades, 5 em cada município. O inquérito epidemiológico nessas propriedades foi feito durante o período compreendido entre 2009 e 2011. O objetivo deste trabalho é verificar a distribuição geográfica da fascíola no sul do estado do Espírito Santo verificando a existência de aglomerados, e analisar os possíveis fatores que possam estar associados com a patologia (fatores de risco). Foi verificado que a prevalência média de fascíola nas propriedades estudadas foi de 19,52%[13,41%;27,35%] e o coeficiente de variação foi estimado em de 8,24%. Foi verificado através da estatística de kernel que a maior intensidade de ocorrência da fascíola está na região central do estudo. Através do cálculo do índice de correlação de Moran, entre os municípios da área de estudo, foi observado o valor de 0,443 *(p-valor* < 0,001) indicando uma autorrelação espacial significativa entre as áreas estudadas. Após análise dos modelos e da revisão da literatura, é possível afirmar que o modelo GAM logístico multivariado é o modelo mais parcimonioso para o estudo da fascíola e seus fatores de risco. Desta forma é possível observar que as variáveis outros hospedeiros e casos anteriores são caracterizadas como fatores de risco epidemiológico para a fascíola.

Palavras-chave: Epidemiologia veterinária. Fasciolose. Modelagem Matemática.

2 ABSTRACT

The fascioliasis is a liver disease caused by the parasite Fasciola, trematoda that parasite liver and biliary tract of cattle and sheep, but also diagnosed in goats, horses, buffalo, human and wild animals such as capybara and eagle rays bathed. Currently, fasciolose has become a serious economic problem for producers from southern state of Espirito Santo and possibly for the entire land mass of the state, in view of the possibility of contamination of cattle herds, sheep, goats and buffaloes for shipping of infected animals and the intermediate host presence in several river basins in the region. The study area was represented by 23 municipalities in the southern state of Espirito Santo. The unit of analysis were property whose economic goal bovine livestock, totaling 115 properties, five in each municipality. The epidemiological survey in these properties was made during the period between 2009 and 2011. The objective of this study is to assess the geographical distribution of fluke in the southern state of Espirito Santo checking for clusters, and analyze the possible factors that may be associated with pathology (risk factors). It was found that the average prevalence of the fluke properties studied was 19.52% [13.41%, 27.35%] and the variation coefficient was estimated at 8.24%. It was verified by the kernel statistics of occurrence of the highest intensity in the central fascioliasis is the area under study. By calculating the Moran's correlation coefficient between the municipalities of the study area, we saw a value of 0.443 (p-value <0.001) indicating a significant spatial autocorrelation between the areas studied. After analysis of the models and the literature review, it can be main that the GAM multivariate logistic model is the most parsimonious model for the study of fascioliasis and its risk factors. This way you can see that variables other hosts and previous cases are characterized as epidemiological risk factors for fascioliasis.

Keywords: Veterinary Epidemiology. Fasciolosis. Mathematical Modeling.

3 INTRODUÇÃO

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda que parasita o fígado e as vias biliares de bovinos e ovinos, mas diagnosticada também em caprinos, equinos, búfalos (PILE et al., 2001), humanos (CORAL et al., 2007) e animais silvestres, como a capivara e ratão do banhado (EL KOUBA, 2005).

Atualmente, a fasciolose tem se tornado um sério problema para a pecuária no Brasil e no mundo. Estudos apontam altos índices de perdas econômicas devido à condenação de fígados, perda de peso e anemia em decorrência da fasciolose .

Agravos de saúde e do bem estar animal, vem sendo estudado ao longo dos anos, tais agravos podem ser ocasionados por fatores ambientais, pelo manejo humano e ultimamente vem se cogitando a tal interferência nesse processo saúde/doença, através mudanças climáticas globais. A ciência que estuda a distribuição temporal e/ou geográfica e possíveis fatores que possam estar influenciando a saúde do animal é a Epidemiologia Veterinária. Como esta ciência tem um caráter quantitativo, para que seja possível trabalhar com ela, é necessário a utilização e até mesmo o desenvolvimento de ferramentas estatísticas e matemáticas. Uma das técnicas mais utilizadas são os modelos estatísticos de regressão (MEDRONHO et al, 2009; SELVIN, 2004; BROEMELING,2013).

Uma das linhas de pesquisa na área da estatística que vem contribuindo e muito com pesquisas epidemiológicas é a análise estatística espacial. A Estatística Espacial é uma área da estatística que estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possuem coordenadas geográficas, tendo como característica o uso implícito ou explicito dessas coordenadas na modelagem. Para BAILEY e GATRELL (1995) a análise estatística é espacial quando os dados são espacialmente localizados e se considera explicitamente a possível importância de sua disposição espacial na interpretação e análise dos resultados.

No contexto epidemiológico, tais técnicas de análise de dados são vitais para descrever a distribuição de uma epidemia no espaço e no tempo a estatística espacial é capaz de sugerir formas de controle e combate a tal doença.

4 OBJETIVOS

Verificar a distribuição geográfica da fascíola no sul do estado do Espírito Santo, e estudar seus possíveis fatores de risco.

5 MATERIAIS E MÉTODOS

5.1 Área de Estudo

A área deste estudo foi representada por 23 municípios do Sul do estado do Espírito Santo, Rio Novo do Sul e os 22 municípios que formam a mesorregião do Sul Espírito-santense : Jeronimo Monteiro, Cachoeiro de Itapemirim, Presidente Kennedy, Castelo, Muniz Freire, Muqui, Guaçuí, Atílio Vivacqua, Alegre, Mimoso do Sul, Bom Jesus do Norte, Itapemirim, Vargem Alta, Ibitirama, Apiacá, Divino São Lourenço, Marataízes, São José dos Calçados, Ibatiba, Iúna, Irupi, Dores do Rio Preto (Figura1.1), abrangeu uma área de aproximadamente 9047,018 km². A unidade de análise foram propriedades que tem como objetivo econômico a pecuária bovina. Foram analisadas cinco propriedades em cada município, totalizando 113 propriedades. O inquérito epidemiológico nessas propriedades foi feito durante o período compreendido entre 2009 e 2011, ver Martins, et al (2014) e Avelar, et al (2012).

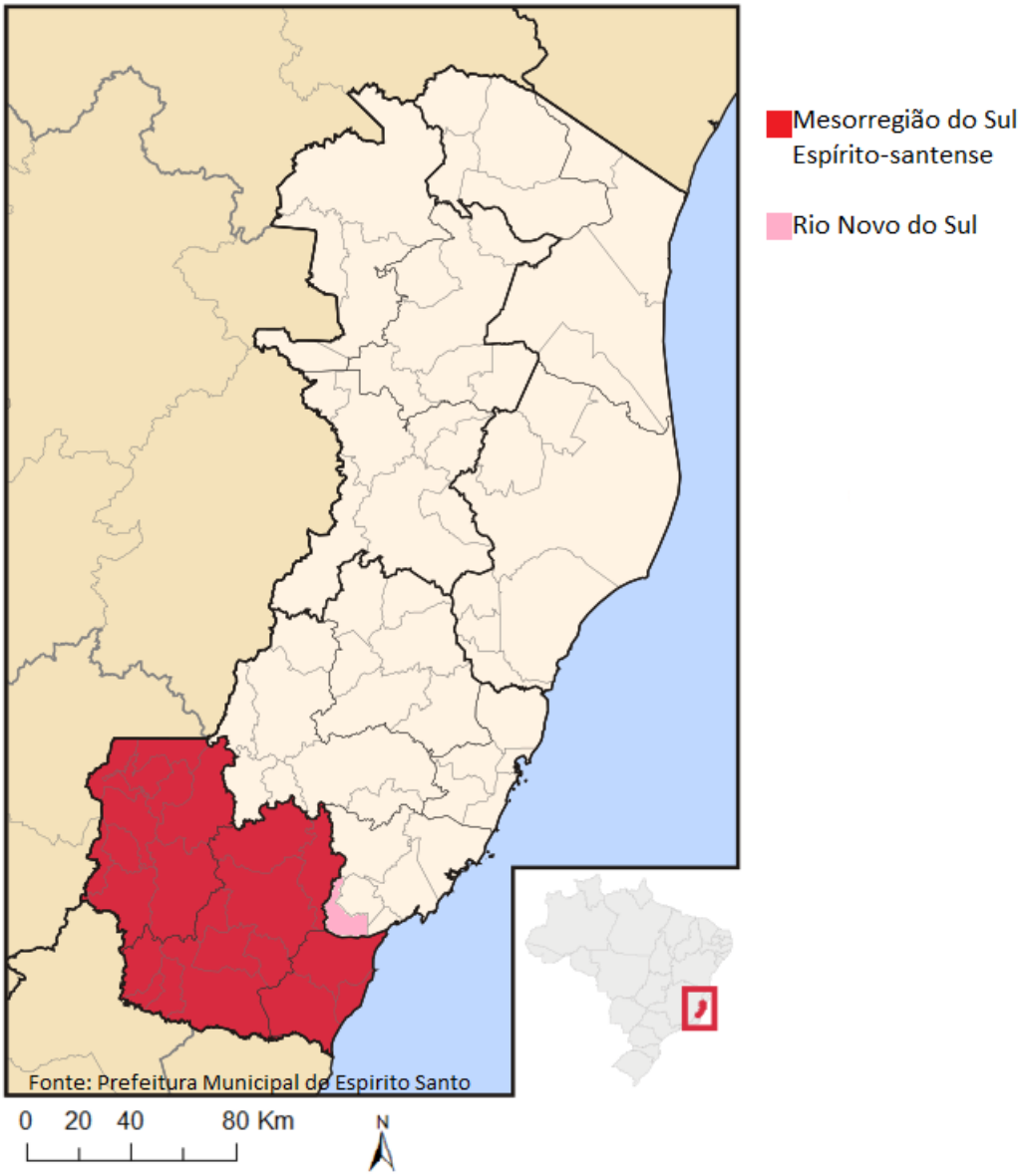


Figura 1.1: Mapa do Espírito Santo. Em destaque os municípios que abrangeram a área deste estudo, em vermelho os 22 que compõe a mesorregião do Sul Espírito-santense e em rosa o município de Rio Novo do Sul.

5.2 Variáveis de estudo

As variáveis analisadas foram: ocorrência de fascíola, presença do caramujo lymneae (hospedeiro intermediário da fascíola), presença de áreas alagadas, existência de outros hospedeiros finais, ocorrência de casos anteriores da doença na propriedade, e prevalência de fascíola. A prevalência foi obtida da divisão do número de animais positivos para a fascíola pelo número de animais examinados na propriedade.

$$prev = \frac{animais\ positivos}{animais\ examinados} \qquad (1.1)$$

5.3 Metodologia

5.3.1 Analise Exploratória de Dados Espaciais

5.3.1.1 Estimador de Kernel

Uma análise exploratória de um processo pontual de dados espaciais começa pela estimação da intensidade de ocorrências do processo em toda a região em estudo. Com isso, gera-se uma superfície cujo valor é proporcional à intensidade de eventos por unidade de área (DRUCK et al.,2004). O estimador *Kernel* é um interpolador, que possibilita a estimação da intensidade do evento em toda a área, mesmo nas regiões onde o processo não tenha gerado nenhuma ocorrência real (DRUCK et al.,2004).

Portanto, suponha que $s_1, \dots, s_n$ são localizações de n eventos observados em uma região A e que *s* represente uma localização genérica cujo valor queremos estimar. O estimador de intensidade é calculado considerando os m eventos $s_1, \dots, s_{m-1}$ contidos num raio de tamanho *t* em torno de *s* e da distância *d* entre a posição e a *i*-ésima amostra (figura1.2), a partir de funções cuja a forma geral é:

$$f_\tau(s) = \frac{1}{\tau^2} \sum_{i=1}^{n} k\left(\frac{d(s_i, s)}{\tau}\right) \ , d(s_i,\ s) < \tau \qquad (1.2)$$

sendo:

- $s_1, \dots, s_n$ são localizações dos n eventos de uma região;
- s é uma localização que será estimada;
- τ é o raio em torno de u;
- $d(s_i,\ s)$ é a distância entre a posição e a *i*-ésima amostra.

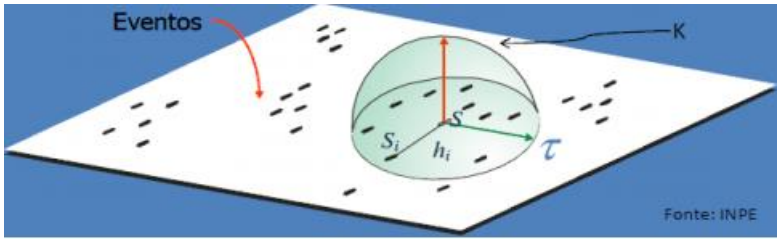


Figura 1.2: Caracterização do estimador de kernel.

5.3.1.1.1 Razão de Kernel

Em situações onde os dados estão distribuídos no espaço de forma heterogênea, o mapa do estimador de densidade de kernel de um determinado fenômeno, pode não refletir da melhor forma a distribuição espacial do risco, podendo indicar de forma errônea as áreas de maior risco. Neste caso, recomenda-se calcular a razão de kernel, que é a razão entre o kernel dos eventos pelo kernel da população (DRUCK et al,2004).

$$RK = \frac{Densidade\ estimada\ para\ os\ eventos}{Densidade\ estimada\ para\ a\ população} \quad (1.3)$$

5.3.1.2 Índice de Moran

A autocorrelação espacial é a correlação de uma certa variável (atributo) *z* numa área *i* com os valores com os valores dessa mesma variável em áreas vizinhas (DRUCK et al., 2004).

O Índice de Moran é um coeficiente muito útil para medir a correlação espacial. Ele mede a relação do desvio padronizado de uma variável *Z* numa área *i* com o desvio padronizado das áreas vizinhas para a mesma variável *Z*. Ele fornece uma medida geral da associação espacial existente no conjunto dos dados. Seu valor varia de –1 a 1. Valores próximos de zero, indicam a inexistência de autocorrelação espacial significativa entre os valores dos objetos e seus vizinhos. Valores positivos para o índice, indicam autocorrelação espacial positiva, ou seja, o valor do atributo de um objeto tende a ser semelhante aos valores dos seus vizinhos. Valores negativos para o índice, por sua vez, indicam autocorrelação negativa. O Índice de Moran é dado por:

$$I = \frac{Z^t W_Z}{Z^t Z} \quad (1.4)$$

sendo:

- *Z*, a variável que representa o atributo de interesse;
- W, a matriz de correlação espacial, ou de vizinhança.

5.3.2. Modelos de Regressão

5.3.2.1. Regressão Linear

Para estudar a relação entre um desfecho (variável dependente ou resposta) e um conjunto de potenciais fatores de risco (variáveis independentes ou explicativas), utiliza-se modelos estatísticos de regressão, com o objetivo de determinar um modelo matemático que descreve essa relação (GUNST e MASON, 1980).

Na maior parte das situações pode-se pensar na variável de desfecho consistindo de duas partes distintas: um componente sistemático (μ) e um componente aleatório (ϵ). Tem-se então um modelo linear clássico de regressão: $Y = \mu + \epsilon$, onde Y é o vetor de dimensões n x 1, da variável de desfecho, $\mu = E(Y) = X\beta$, o componente sistemático, X a matriz de dimensões n x p do modelo, $\beta = (\beta_1,..., \beta_p)^T$ o vetor dos parâmetros, $\epsilon = (\epsilon_1,..., \epsilon_n)^T$, o componente aleatório com $\epsilon_i \rightarrow N(0,\sigma^2)$, i = 1,...,n. O método de estimação mais comumente usado neste caso é baseado na minimização dos quadrados do componente aleatório ϵ^2, e por isso chamado de mínimos quadrados (MMQ). É importante ressaltar alguns de seus pressupostos básicos para o ajuste de modelos de regressão linear:

1. A ausência de autocorrelação entre os erros (componentes aleatórios), cor $(\epsilon_i, \epsilon_j) \rightarrow 0$;
2. Variáveis independentes não correlacionadas (colinearidade), cor $(x_1,..., x_p) \rightarrow 0$;
3. A existência de homocedasticidade, ou seja, variância constante dos resíduos, var $\epsilon_i = \sigma^2$.

5.3.2.2 Modelo Linear Generalizado

É possível utilizar métodos análogos àqueles desenvolvidos para o modelo de regressão linear, em situações em que a variável resposta obedece a outras distribuições que não a Normal, ou em que a relação entre a variável resposta e as variáveis explicativas não é linear. Isto se deve, em parte, ao conhecimento de que muitas das boas propriedades da distribuição Normal são partilhadas por uma larga classe de distribuições denominado de família exponencial (DOBSON, 1990).

Nelderand e Wedderburn (1972) propuseram uma extensão dos modelos lineares clássicos, denominado Modelos Lineares Generalizados (GLM). Este modelo tem como características principais:

- A variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica: distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;
- As variáveis explicativas, entram na forma de um modelo linear (componente sistemático);
- A ligação entre os componentes aleatório e sistemático é feita através de uma função de ligação (por exemplo, logarítmica para modelos log-lineares), conforme a fórmula abaixo:

$$f(y;\theta,\emptyset) = \exp\left(\frac{(y\theta - b(\theta))}{a(\emptyset)} + c(y,\emptyset)\right) \qquad (1.5)$$

sendo θ o parâmetro natural e $a(\emptyset)$ o fator de dispersão. Tendo como componentes básicas:

- A variável desfecho y, cuja distribuição de probabilidade pertence a família exponencial, com valores esperados $E(y_i) = \mu_i$;
- Um preditor linear baseado nas variáveis explicativas $x_{i1}, \dots, x_{i(p-1)}$ denotado por $x_i\beta = \eta_i$;
- A função de ligação g relacionada ao preditor linear do desfecho: $\eta = g(\mu_i)$.

Entre estes modelos os mais usados na área de epidemiologia são a regressão logística, tendo uma variável binaria como desfecho, e a regressão de Poisson, tendo como variável desfecho contagens de casos ou óbitos de uma determinada patologia.Tradicionalmente o ajuste desses modelos é baseado no método de estimação da máxima verossimilhança, pelo qual os estimadores são obtidos a partir da maximização da função de verossimilhança, e os cálculos envolvem um procedimento interativo (BOLFARINE e SANDOVAL, 2001).

5.3.2.3 Modelos aditivos generalizados

Uma extensão dos modelos lineares generalizados são os modelos aditivos generalizados (GAM). Neste, Hastie & Thibshirani (1990) propuseram a utilização de funções, usualmente não paramétricas, sobre as variáveis independentes de forma a linearizar a relação com a variável resposta. O parâmetro estimado, neste caso, não relaciona diretamente a quantidade x à quantidade y, mas uma função de x a y. na verdade, esta ideia é uma extensão da transformação de variáveis já muito utilizada, que tem sua maior aplicação quando o tipo de relação entre as variáveis é de forma complexa. Uma particularidade das funções não paramétricas é a capacidade de ajuste mesmo nos extremos. Temos então:

$$\eta = f_1(x_1) + \cdots + f_k(x_k) + \varepsilon \qquad (1.6)$$

sendo:

- k = 1, ... , p.
- f_k são as funções de alisamento (suavização) das covariáveis x_k.

Para descrever os fenômenos geograficamente distribuídos pode ser usado este modelo com a estrutura espacial incorporada no mesmo. Daí temos:

$$\eta = f_1(x_1) + \cdots + f_k(x_k) + f_{GEO}(latitude, longitude) + \varepsilon \qquad (1.7)$$

Sabendo que:

- $f_{GEO}(latitude, longitude)$ representa uma função bidimensional das coordenadas geográficas das observações em estudo.

Tais modelos apresentados podem representar as covariáveis não somente através de funções de alisamento, mas também pode representar as variáveis de forma original.

$$\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{w} f(x_j) + f_{GEO}(latitude, longitude) + \varepsilon \qquad (1.8)$$

5.3.3 Estratégia de Análise

Primeiramente serão calculadas as prevalências por município de estudo (fórmula1.1), e seus respectivos intervalos de confiança, supondo 5% de significância ($\alpha = 5\%$). Posteriormente será feita uma análise exploratória de dados com mapa da distribuição de ocorrência de casos de fascíola por propriedade e o mapa temático da distribuição da prevalência de fascíola no sul do estado do Espirito Santo. Será utilizada função de kernel para verificar existência de padrão espacial da fascíola.

Para verificar a existência de autorrorelação espacial dos dados, será estimado o valor do índice de Moran. Logo em seguida serão ajustados modelos de regressão logisticos, bivariados e multivariados, clássicos e com estrutura espacial. E para verificar o modelo mais parcimonioso será utilizado o critério de informação de "*akaike*" (AIC).

5.3.4 Softwares utilizados

No desenvolvimento desse trabalho foram utilizadas ferramentas computacionais livres, definida como aquela na qual "os usuários tem total liberdade de executar, copiar, distribuir, estudar, modificar e aperfeiçoar o software" como são o R e o TerraView.

R é uma linguagem e um ambiente de desenvolvimento integrado, para cálculos estatísticos e gráficos. Foi criada originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi desenvolvido por um esforço colaborativo de pessoas em vários locais do mundo (JOHN & ANDERSEN, 2005). É gratuito. O nome R provêm em parte das iniciais dos criadores e também de um jogo figurado com a linguagem S (da Bell Laboratories, antiga AT&T). O código fonte do R está disponível sob a licença GNU GPL e as versões binárias pré-compiladas são fornecidas para Windows, Macintosh, e muitos sistemas operacionais Unix/Linux. R é também altamente expansível com o uso dos pacotes, que são bibliotecas para funções específicas ou áreas de estudo específicas. Um conjunto de pacotes é incluído com a instalação de R, com muito outros disponíveis na rede de distribuição do R . A linguagem R é largamente usada entre estatísticos e data miners para desenvolver software de estatística e análise de dados (SMITH, 2012). Inquéritos e levantamentos de data miners mostram que a popularidade do R aumentou substancialmente nos últimos anos.

O TerraView é um sistema de informações geográficas desenvolvido pela Divisão de Processamento de Imagens (DPI) do Instituto Nacional de Pesquisas Espaciais INPE (TerraView 4.1.0. São José dos Campos, SP: INPE, 2010). Sua principal característica é a manipulação de dados vetoriais e matriciais. Todos os dados são armazenados em um banco de dados relacional ou geo-relacional, como MySQL ou PostGreSQL. O TerraView permite a criação de mapas temáticos com os mais diferentes tipos de legendas, além de ser compatível com dados nos formatos MID/MIF, Shapefile e Tab/Geo.

6 RESULTADOS E DISCUSSÃO

Foi verificado que a prevalência média de fascíola nas propriedades estudadas foi de 19,52%[13,41%;27,35%] e o coeficiente de variação foi estimado em de 8,24%, concluindo-se uma variabilidade baixa na região de estudo (Tabela1). Verifica-se que as propriedades Jerônimo Monteiro 50,72%[42,10%;59,30%], Muqui 51,24%[42,61%;59,80%] e Vargem Alta 49,00%[40,43%;57,63] apresentaram os maiores índices de prevalência. Enquanto os municípios Ibitirama, Divino São Lourenço, Ibatiba, Iuna e Irupi apresentaram prevalência nula.

Tabela1.1: Distribuição da Prevalência Média [IC 90%] de fascíola bovina entre as propriedades estudadas por município no Estado do Espirito Santo, 2009-2011.

Municípios	Prev[IC 90%]
Jeronimo Monteiro	50,72%[42,10%;59,30%]
Cachoeiro de Itapemirim	37,33%[29,34%;46,03%]
Presidente Kennedy	35,67%[27,80%;44,34%]
Castelo	29,50%[22,17%;37,98%]
Muniz Freire	4,39%[1,74%;9,77%]
Muqui	51,24%[42,61%;59,80%]
Guaçui	15,22%[9,82%;22,60%]
Atilio Vivacqua	51,33%[42,69%;59,89%]
Alegre	24,33%[17,57%;32,54%]
Mimoso do Sul	10,00%[5,69%;16,62%]
Bom Jesus do Norte	12,12%[7,33%;19,08%]
Rio Novo do Sul	8,00%[4,20%;14,25%]
Itapemirim	5,00%[2,13%;10,55%]
Vargem Alta	49,00%[40,43%;57,63%]
Ibitirama	0,00%[0,00%;0,00%]
Apiacá	18,33%[12,41%;25,05%]
Divino Sao Lourenço	0,00%[0,00%;0,00%]
Marataizes	9,27%[5,14%;15,76]
São José dos Calçados	10,00%[5,69%;16,62%]
Ibatiba	0,00%[0,00%;0,00%]
Iuna	0,00%[0,00%;0,00%]
Irupi	0,00%[0,00%;0,00%]
Dores do Rio Preto	24,44%[17,67%;32,65%]

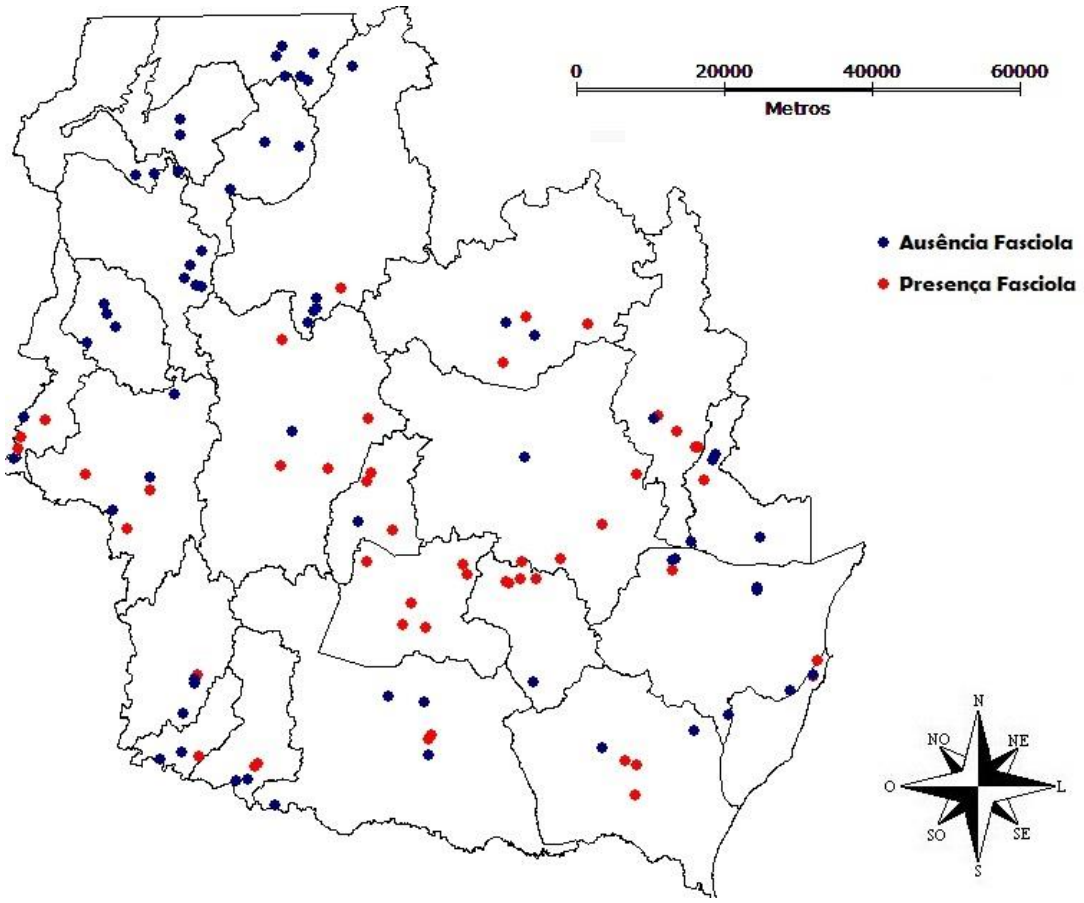


Figura1.3: Distribuição das propriedades estudadas com e sem registro de casos de ocorrência de fascíola hepática bovina no sul do estado do Espírito Santo, 2009-2011.

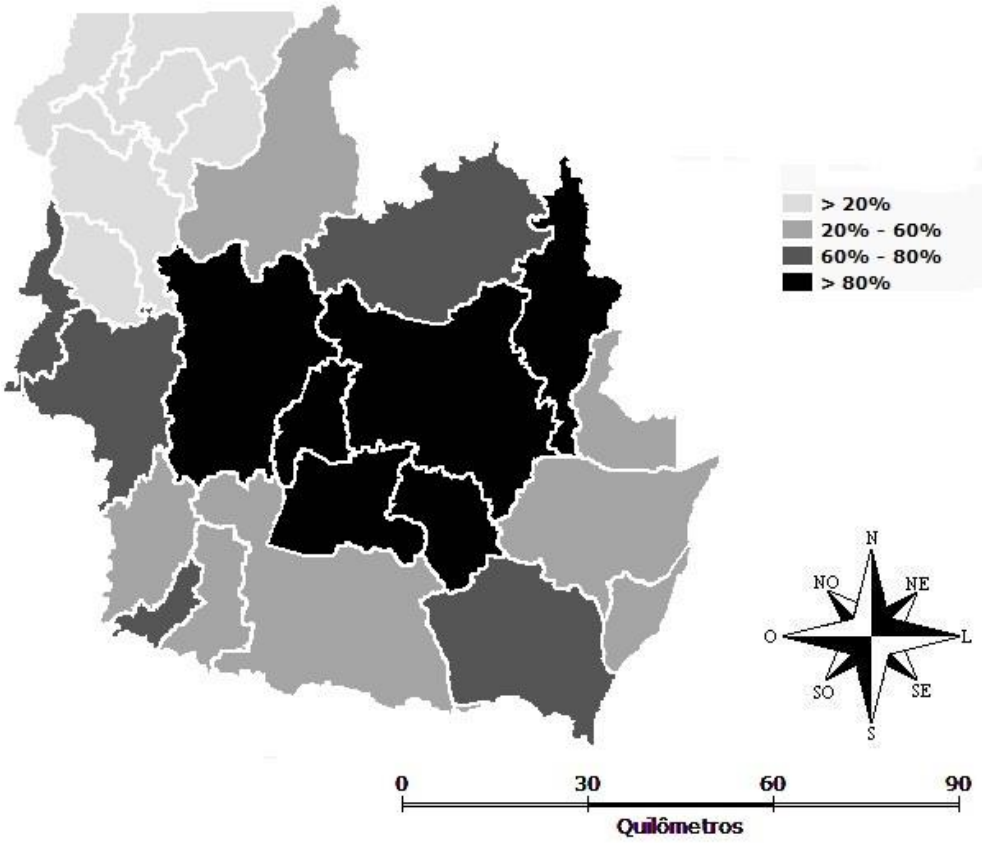


Figura 1.4: Distribuição espacial da prevalência média por municípios de fascíola entre as propriedades estudadas no estado do Espirito Santo, 2009-2011.

Observando as figuras 1.3 e 1.4 nota-se que os dados estão distribuídos de maneira heterogênea, municípios com maior prevalência, estão situados na região central do estudo, são eles Atílio Vivacqua, Cachoeiro de Itapimirim e Muqui.

Observando a figura 1.5 verifica-se através do mapa de kernel da região de estudo que a fascíola está concentrada de forma mais intensa na região central do mapa, mas especificamente nos municípios Atílio Vivacqua, Cachoeiro de Itapimirim e Muqui. Esta concentração ocorre no limite desses três municípios.

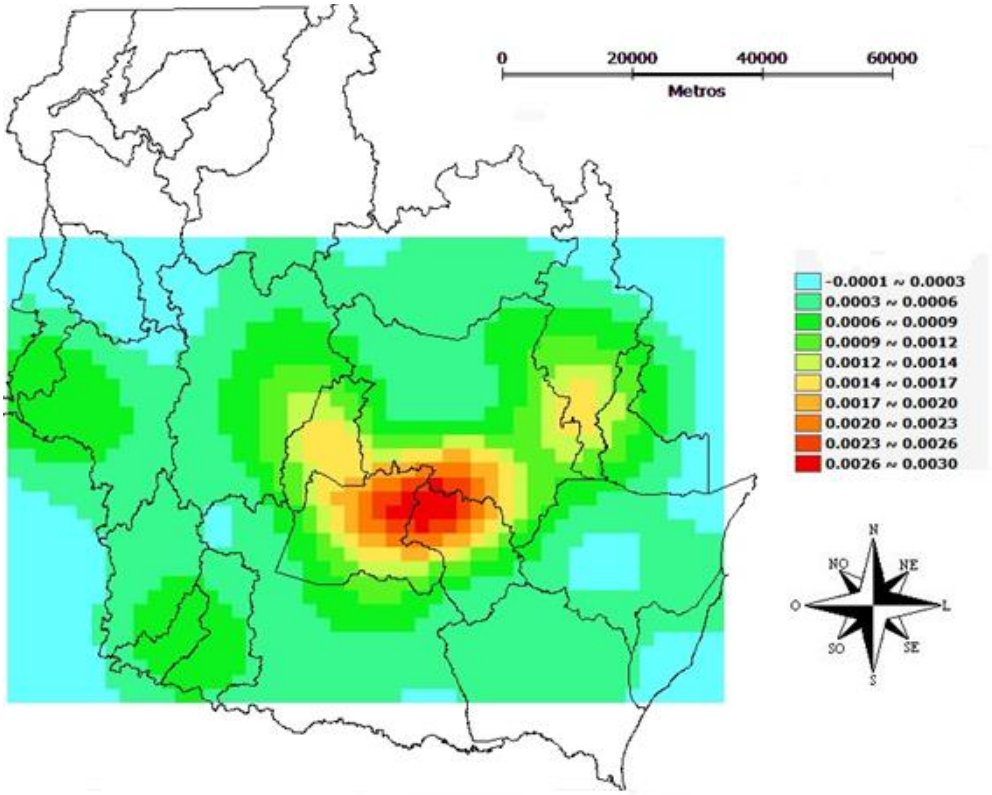


Figura 1.5: Mapa de kernel baseado na probabilidade de ocorrência de fascíola entre as propriedades estudadas.

Ainda com os dados agregados por municípios, ao observar as prevalências, verifica-se que através do cálculo do índice de correlação de Moran, entre os municípios da área de estudo, foi observado o valor de 0,443 *(p-valor* < 0,001) indicando uma autorrelação espacial significativa entre as áreas estudadas, tal tipo de inferência corrobora com os achados da estatística de kernel, sugerindo a ocorrência de aglomerados espaciais.

Como os dados de caso de fascíola estão distribuídos de forma heterogênea na área de estudo, foi utilizada a razão de kernel entre as propriedades onde foram diagnosticados casos de fascíola em relação a todas as propriedades do estudo. Tal análise constata-se cada vez mais uma intensidade de probabilidade de ocorrência de fascíola na região central do estudo (Figura 1.6).

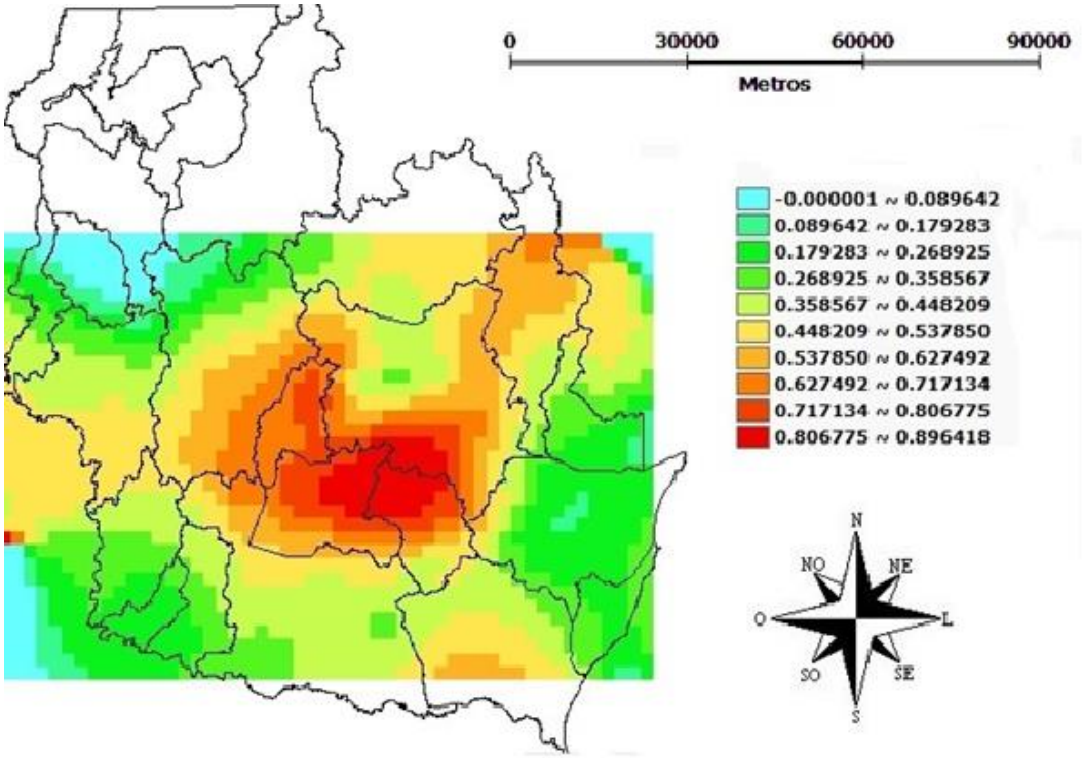


Figura 1.6: Mapa da razão de kernel baseado na probabilidade entre as propriedades com fascíola e todas as propriedades.

Tabela 1.2: Estimativa dos coeficientes de regressão dos modelos não espaciais e seus respectivos IC 90%. Os modelos logísticos foram utilizadas estimativas Odds Ratio (OR).

	Logístico Bivariado	Logístico Multivariado
Lymneae	6,09 [3,04;12,18]*	__
Area Alagada	8,50 [2,37;30,43]*	__
Outros Hospedeiros	16,08 [4,56;56,63]*	8,50[2,11;34,36]*
Casos Anteriores	18 [5,01;64,62]*	10,83[2,67;43,90]*
AIC	**	121,33

* Variáveis significativas ao nível de 10% de significância.
** O AIC já foi estimado para cada modelo e ajustado em relação a cada variável.

Verificando a tabela 1.2 observa-se que de acordo com os modelos logísticos bivariados todas as variáveis também foram significativas. A probabilidade de ocorrência da fascíola é aumentada em 6,09 [3,04;12,18] vezes com a presença do Lymnae, 8,50 [2,37;30,43] vezes com a presença de áreas alagadas, 16,08 [4,56;56,63] vezes com a presença de outros hospedeiros e 18 [5,01;64,62] vezes com a presença de casos anteriores, e de acordo com o modelo logístico multivariado, apenas as variáveis outros hospedeiros e casos anteriores foram significativas, ao nível de 10% de significância. Logo estas são suficientes para explicar a ocorrência da fascíola.

Tabela 1.3: Estimativa dos coeficientes de regressão dos modelos e seus respectivos IC 90%. Para os modelos de regressão linear foram utilizadas estimativas dos coeficientes de regressão linear, e para os modelos logísticos, foram utilizadas estimativas Odds Ratio (OR).

	GAM Logístico Bivariado	GAM Logístico Multivariado
Lymneae	5,25[1,87;14,74]*	__
Área Alagada	6,50[1,15;36,92]*	__
Outros Hospedeiros	15,09[2,94;77,41]*	13,38[1,92;93,35]*
Casos Anteriores	15,03[2,43;93,07]*	9,61[1,17;79,05]*
AIC	**	108.25

* Variáveis significativas ao nível de 10% de significância.
** O AIC já foi estimado para cada modelo e ajustado em relação a cada variável.

Já a tabela 1.3, mostra que através do ajuste dos modelos que levam em conta a estrutura espacial, de acordo com os modelos GAM logísticos bivariados todas as variáveis foram significativas ao nível de 10% de significância, e mais, a probabilidade de ocorrência da fascíola é aumentada em 5,25[1,87;14,74] vezes com a presença do Lymnae, 6,50[1,15;36,92] vezes com a presença de área alagada, 15,09[2,94;77,41] vezes com a presença de outros hospedeiros e 15,03[2,43;93,07] vezes com a presença de casos anteriores, de acordo com o modelo GAM logístico multivariado, apenas as variáveis outros hospedeiros e casos anteriores foram significativas, ao nível de 10% de significância. Logo estas são suficientes para explicar a ocorrência da fascíola.

Foi verificado pelo critério de "*akaike*" (AIC = 108) que o modelo mais parcimonioso para o estudo da fascíola e seus fatores de risco foi o modelo GAM logístico multivariado A revisão da literatura (TASSINARI et al., 2013), ratifica esse resultado. Desta forma é possível observar que as variáveis outros hospedeiros e casos anteriores são suficientes para explicar a ocorrência da fascíola.

Apesar da presença de áreas alagadas não ter sido significativa, é sabido que esta variável é considerada como fator importante para a disseminação da fasciolose. As regiões alagadas, permitem que haja um ambiente favorável ao desenvolvimento do molusco, completando o ciclo do parasito. Busseti (1982) confirma este fato ao relatar em Curitiba, presença de sete bovinos contaminados em regiões de onde provieram casos humanos. Além disso, o autor relata que na área estudada foram encontrados moluscos da família Lymnaeidae, criados em áreas alagadiças onde também cresce o agrião, foco de contaminação humana.

7 CONCLUSÃO

- Foi observado uma prevalência de fascíola média de 19,52%[13,41%;27,35%] no sul do estado do Espirito Santo, podendo variar com cidades com 0,00%[0,00%;0,00%] até 51,33%[42,69%;59,89%] de prevalência.

- A região central do estudo, englobando os municípios de Atílio Vivacqua, Cachoeiro de Itapimirim e Muqui, bem como os municípios de Jeronimo Monteiro e Vargem alta são regiões com altos níveis de ocorrência da fascíola.

- A ocorrência de casos anteriores e outros hospedeiros nas propriedades são caracterizadas como fatores de risco epidemiológico para a fascíola.

- Em trabalhos futuros, é importante a utilização de modelos mais sofisticados que expliquem com maior precisão a complexidade do problema.

- O estímulo para utilização de softwares livres e gratuitos é necessária no meio acadêmico, já que se trata de uma poderosa ferramenta de análise de dados e desenvolvimento de rotinas operacionais.

- Diante deste panorama é cada vez mais eminente a necessidade de uma equipe multidisciplinar onde atuem matemáticos, médicos veterinários, epidemiologistas, geógrafos, etc, em prol de causas em comum.

CAPITULO II

PROPOSTAS DE TÉCNICAS BASEADAS EM HEURISTICA PARA A DETECÇÃO DE AGLOMERADOS ESPACIAIS DE FOCOS DE OCORRÊNCIA DE FASCÍOLA HEPÁTICA BOVINA

1 RESUMO

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda que parasita o fígado e as vias biliares de bovinos e ovinos, mas diagnosticada também em caprinos, equinos, búfalos, humanos e animais silvestres, como a capivara e ratão do banhado. Atualmente, a fasciolose tem se tornado um grave problema econômico para os produtores da região Sul do estado do Espírito Santo e possivelmente para toda a extensão territorial do estado, tendo em vista a possibilidade de contaminação de rebanhos bovinos, ovinos, caprinos e bubalinos pelo transporte de animais parasitados e pela presença do hospedeiro intermediário em diversas bacias hidrográficas da região. A área deste estudo foi representada por 23 municípios do Sul do estado do Espírito Santo. A unidade de análise foram as 51 propriedades que tem como objetivo econômico a pecuária bovina cujo resultado para fascíola foi positivo. O inquérito epidemiológico nessas propriedades foi feito durante o período compreendido entre 2009 e 2011. O objetivo deste trabalho é propor e discutir algumas técnicas de modelagem matemática baseadas em heurística na detecção de aglomerados espaciais nas propriedades acometidas com fascíola hepática. Foi verificado que a prevalência média de fascíola nas propriedades estudadas foi de 19,52%[13,41%;27,35%] e o coeficiente de variação foi estimado em de 8,24%. Para detecção dos aglomerados geográficos foram utilizadas duas heurísticas, *k-means* com *elbow* e algoritmo genético com função silhueta, ambas foram significas mas para estes dados a segunda se mostrou mais precisa que a primeira, retornando o valor de 5 (cinco) clusters (aglomerados) e mostrando que o cluster de maior concentração é o localizado na região central do estudo. Tal resultado valida o que foi encontrado via estimados de *kernel.* Assim concluímos que as propriedades pertencentes ao cluster 1 (um) necessitam de um atendimento prioritário.

Palavras-chave: Epidemiologia Veterinária. Fasciolose. Modelagem Matemática.

2 Abstract

The fascioliasis is a liver disease caused by the parasite Fasciola, trematoda that parasite liver and biliary tract of cattle and sheep, but also diagnosed in goats, horses, buffalo, human and wild animals such as capybara and eagle rays bathed. Currently, fascioliasis has become a serious economic problem for producers from southern state of Espirito Santo and possibly for the entire land mass of the state, in view of the possibility of contamination of cattle herds, sheep, goats and buffaloes for shipping of infected animals and the intermediate host presence in several river basins in the region. The study area was represented by 23 municipalities in the southern state of Espirito Santo. The unit of analysis was the 51 properties whose economic goal bovine cattle for fluke whose result was positive. The epidemiological survey in these properties was made during the period between 2009 and 2011. The objective of this work is to propose and discuss some techniques of mathematical modeling based on heuristics to detect spatial clusters in the affected properties with liver fascioliasis. It was found that the average prevalence of the fascioliasis properties studied was 19.52% [13.41%, 27.35%] and the variation coefficient was estimated at 8.24%. For detection of geographic clusters were used two heuristics, k-means with elbow and genetic algorithm with silhouette function, both were significas but to this data the second proved more accurate than the first, returning the value of 5 (five) clusters and showing that the cluster with the highest concentration is located in the center of the study area. This result validates what was found in the kernel statistics. Thus we conclude that the properties belonging to the cluster 1 (one) require priority attention.

Keywords: Veterinary Epidemiology. Fasciolosis. Mathematical Modeling.

3 INTRODUÇÃO

Atualmente, a fasciolose tem se tornado um sério problema para a pecuária bovina no Brasil e no mundo. Estudos apontam altos índices de perdas econômicas devido à condenação de fígados bovinos em decorrência da fasciolose.

A fasciolose é uma doença causada pelo parasito *Fasciola hepática*, trematoda que parasita o fígado e as vias biliares de bovinos e ovinos, mas também é diagnosticada em caprinos, equinos, búfalos (PILE et al., 2001), humanos (CALRETAS et al., 2003; SVS, 2005; CORAL et al., 2007) e animais silvestres, como a capivara e ratão do banhado (EL KOUBA, 2005). Causa em seus hospedeiros condenação do fígado, perda de peso, anemia e outros sinais inespecíficos, o que torna o diagnóstico clínico da doença difícil fazendo-se necessária a realização do diagnóstico laboratorial (ECHEVARRIA, 1985), baseado na observação de ovos de *Fascíola hepática* nas fezes dos animais (BORAY, 1985; KLEIMAN et al., 2005) e de humanos.

Ao longo dos anos vem se estudando vários agravos na saúde animal. Tais agravos são provocados pelo manejo humano e ultimamente vem se cogitando a interferência nesse processo saúde/doença, pelas mudanças climáticas globais. Uma das ciências que estuda esse processo é a Epidemiologia Veterinária. Ela é responsável pelo estudo dos fatores que possam vir a influenciar determinadas patologias, e também fazer previsões temporais e/ou espaciais destas enfermidades.

Como ação para um retorno rápido de um estudo para um diagnóstico coletivo a nível populacional, o inquérito epidemiológico é importante ferramenta analítica para a investigação dos fatores de risco presentes na causalidade das doenças e na caracterização de sua distribuição, no espaço e no tempo no âmbito populacional. Desde a década de 80, surge um renovado interesse nos estudos de padrões espaciais e temporais de doenças, conforme salienta a extensa literatura publicada em periódicos de diferentes áreas, incluindo importantes revisões (KNOX, 1991; WERNECK e STRUCHINER, 1997). Dentre os desenhos epidemiológicos utilizados neste contexto, destacam-se os estudos de aglomerados (clusters, na língua inglesa). De maneira geral, aglomerados espaciais de doenças podem ser atribuídos aos fatores demográficos, genéticos, ambientais ou, socioculturais superpostos geograficamente ao padrão de ocorrência observado. O estudo de técnicas para detecção de aglomerados espaciais no campo da Epidemiologia recebe importante contribuição científica (WERNECK e STRUCHINER, 1997).

Modelagem matemática e computacional é uma área de conhecimento multidisciplinar que trata da aplicação de modelos matemáticos e técnicas da computação à análise, compreensão e estudo da fenomenologia de problemas complexos em áreas tão abrangentes quanto as engenharias, ciências exatas, biológicas, humanas, economia e ciências ambientais. Dentre as várias técnicas utilizadas para resolver os modelos matemáticos destacam-se as heurísticas.

Um algoritmo é considerado um método heurístico quando não há conhecimentos matemáticos completos sobre seu comportamento, ou seja, quando, sem oferecer garantias, o algoritmo objetiva resolver problemas complexos utilizando uma quantidade não muito grande de recursos - especialmente no que diz respeito ao consumo de tempo - para encontrar soluções de boa qualidade. Uma metaheurística é um conjunto de conceitos que pode ser utilizado para definir métodos heurísticos aplicáveis a um extenso conjunto de diferentes problemas. Em outras palavras, uma metaheurística pode ser vista como uma estrutura algorítmica geral que pode ser aplicada a diferentes problemas de otimização com relativamente poucas modificações que possam adaptá-la a um problema específico. Alguns

exemplos de metaheurísticas são: simulated annealing, busca tabu, iterated local search, algoritmos genéticos e ant colony optimization" . (GLOVER e KOCHENBERGER, 2003)

O problema de encontrar aglomerados espaciais de doenças pode ser definido na área de Modelagem Matemática como um problema de Clusterização ou Agrupamento (CRUZ, 2010; BASTOS et al., 2014). Clusterização é o termo genérico para um processo que une objetos similares em um mesmo grupo. Cada grupo é denominado um cluster. O número de clusters pode ser conhecido a priori ou não. Quando o número de clusters é conhecido, a priori, tratamos Problema de K-Clusterização ou, simplesmente Problema de Clustrização (PC). Caso contrário, ou seja, quando não se conhece o número de grupos, o problema é denominado Problema de Clusterização Automática (PCA). O problema de clusterização pode ser resolvido utilizando heurística baseada ou não em metaheurística.

O objetivo deste capítulo é investigar a existência de aglomerados espaciais de fascíola hepática em propriedades do sul do estado do Espirito Santo, utilizando heurísticas, com o intuito de verificar quais são as regiões de maior prioridade para a intervenção.

4 OBJETIVOS

Em Epidemiologia, o método utilizado para resolver o PCA é denominado *k-means*, que é uma heurística simples e apresenta bons resultados.

Porém para verificar a sua qualidade, o *k-means* foi comparado com um algoritmo genético proposto por (CRUZ e OCHI, 2011). Para isso, os dois métodos foram executados em um conjunto de instancias propostas em (CRUZ e OCHI, 2011). O objetivo foi escolher o melhor método para utilizar nos dados deste trabalho a fim de detectar o número de aglomerados espaciais formados.

5 MATERIAIS E MÉTODOS

5.1 Dados

A área deste estudo foi representada por 23 municípios do Sul do estado do Espírito Santo: Jeronimo Monteiro, Cachoeiro de Itapemirim, Presidente Kennedy, Castelo, Muniz Freire, Muqui, Guaçuí, Atílio Vivacqua, Alegre, Mimoso do Sul, Bom Jesus do Norte, Rio Novo do Sul, Itapemirim, Vargem Alta, Ibitirama, Apiacá, Divino São Lourenço, Marataízes, São José dos Calçados, Ibatiba, Iúna, Irupi, Dores do Rio Preto. A unidade de análise foram propriedades que tem como objetivo econômico a pecuária bovina. Foram analisadas cinco propriedades em cada município, totalizando 115 propriedades. O inquérito epidemiológico nessas propriedades foi feito durante o período compreendido entre 2009 e 2011.

Foram utilizados dados de 51 propriedades em que foram verificadas pelo menos um caso positivo da fascíola hepática bovina. Além dos dados originais, para validação dos métodos de clusterização propostos neste trabalho, foram utilizados quarenta e oito bancos de dados já testados em outros trabalhos (CRUZ e OCHI, 2011), cujo valor "ótimo de aglomerados" foram previamente conhecidos.

5.2 Metodologia

5.2.1 Estratégia de Análise

O método *k-means* conjugado com a técnica do "cotovelo" é o método mais utilizado na área epidemiológica, para encontrar aglomerados espaciais. Isto acontece devido a sua simplicidade e a sua disponibilidade nos pacotes estatísticos. Porém, para verificar a sua qualidade, o k-means foi comparado com outro método proposto em (CRUZ e OCHI, 2011), onde o código fonte estava disponível para a utilização.

Estes métodos foram comparados utilizando um conjunto de quarenta e oito instâncias das quais já sabíamos qual seria o numero de grupos. O método que obteve o maior número de acertos em relação ao numero de grupos foi o escolhido para a utilização dos dados deste trabalho.

5.2.2 O problema de clusterização automática

O problema estudado pode ser definido como o problema de clusterização automática na Pesquisa Operacional. O problema de clusterização automática é definido da seguinte forma: dado X um conjunto de N objetos $X=\{x_1, x_2, x_3,..., x_n\}$, onde cada objeto xi é uma tupla $(x_{i1},x_{i2},...,x_{ip})$, e cada coordenada x_{ij} está relacionada com um atributo *j* do objeto *i*. Cada objeto pode ser considerado um ponto no espaço R^p. Como objetivo, devemos encontrar o conjunto $C=\{C_1,C_2,...,C_k\}$ de grupos ou clusters, *k* não conhecido previamente, tal que a similaridade entre os objetos de um mesmo cluster seja maximizada e a similaridade entre objetos de diferentes clusters seja minimizada, sujeito as seguintes condições:

$C_i \neq \{\}$, *para i = 1, ... , k* (2.1)

$C_i \cap C_j \neq \{\}$, *para i,j = 1,...,k e* $i \neq j$ (2.2)

$\cup C_i = X$, *para i = 1, ... , k* (2.3)

Outra definição necessária é o conceito de centróide de um cluster Ci. A substituição de um conjunto *Ci*com *t* pontos similares por um único ponto $v_m = (v_{m1}, v_{m2},...,v_{mp})$ que os represente, pode ser feito considerando-se v_m o centróide de *Ci*, onde cada coordenada de v_m é dada pela equação (2.4):

$$v_{mi} = \frac{1}{t}\sum_{j=1}^{t} x_{mj} \text{ , } i = 1, \ldots, p. \quad (2.4)$$

5.2.3 O método K-means

K-means (MACQUENN,1967) é um método de clusterização que objetiva particionar um conjunto com n objetos em um número *k* de clusters, fixado a priori. O algoritmo visa minimizar uma função objetivo, neste caso uma função do erro quadrado. A função objetivo

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 , \quad (2.5)$$

onde $||x_i^{(j)} - c_j||^2$ é uma medida de distância escolhida entre um objeto $x_i^{(j)}$ e o centroide do aglomerado c_j.

O algoritmo consiste em:

1) Escolher aleatoriamente um número k de centros para os clusters;
2) Atribuir cada objeto para o cluster de centro mais próximo (ex. usando a distância Euclidiana)
3) Mover cada centro para a média dos objetos atribuídos;
4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido (número de iterações, tolerância em relação às mudanças nos centroides).

5.2.4 O Método Elbow

O Método Elbow olha para a percentagem de variância explicada como uma função do número de agrupamentos: Deve-se escolher um certo número de aglomerados de modo que a adição de outro agrupamento não melhore muito a modelagem dos dados. Mais precisamente, se traça a percentagem de variância explicada pelos grupos contra o número de clusters, os primeiros agrupamentos vão acrescentar muita informação, mas em algum momento o ganho marginal vai cair, dando um ângulo no grafo. O número de clusters é escolhido, neste ponto, portanto, o "critério cotovelo". Este "*cotovelo*" pode nem sempre ser identificado sem ambiguidade (GOUTTE et al., 1999).

5.2.5 Heuríticas e Metaheurísticas

A palavra heurística vem do grego "*heuristikê*", cujo significado é "arte de descobrir". Ou seja, a heurística é um processo utilizado para a resolução de um problema, e consiste em métodos e regras que levam à invenção, à descoberta e à resolução de uma questão mediante o uso da criatividade. Seu método é responsável por proporcionar uma rápida e simples solução com o menor gasto de energia e esforço. É um método eficiente para buscar soluções em áreas difíceis.

Segundo a definição original, metaheurísticas são métodos de solução que coordenam procedimentos de busca locais com estratégias de mais alto nível, de modo a criar um processo capaz de escapar de mínimos locais e realizar uma busca robusta no espaço de soluções de um problema (GLOVER e KOCHENBERGER, 2003). Posteriormente, a definição passou a abranger quaisquer procedimentos que empregassem estratégias para escapar de mínimos locais em espaços de busca de soluções complexas. Em especial, foram incorporados procedimentos que utilizam o conceito de vizinhança para estabelecer meios de fugir dos mínimos locais. Uma metaheurística, portanto, visa produzir um resultado satisfatório para um problema, porém sem qualquer garantia de otimalidade. Metaheurísticas são aplicadas para encontrar respostas a problemas sobre os quais há poucas informações: não se sabe como é a aparência de uma solução ótima, há pouca informação heurística disponível e força-bruta é desconsiderada devido ao espaço de solução ser muito grande. Porém, dada uma solução candidata ao problema, esta pode ser testada e sua otimalidade, averiguada. Algumas metaheurísticas conhecidas são o algoritmo genético, GRASP, busca tabu (LOPES, 2013)

5.2.6 Algoritmo Genético

Os Algoritmos genéticos foram inspirados no mecanismo da evolução das espécies, tendo como base os trabalhos de Darwin e Mendel. Tais algoritmos vêm sendo utilizados com sucesso para a resolução dos mais variados e complexos tipos de problemas.

Os Algoritmos Genéticos (AG) vêm sendo usados com sucesso para encontrar boas soluções para uma ampla variedade de problemas de otimização (GEN e CHENG, 1997) desde sua introdução por Holland na década de 1970 (HOLLAND, 1975).

O AG é um método computacional de busca baseado em mecanismos de evolução natural e da genética. Em um AG, uma população de possíveis soluções para um dado problema evolui de acordo com operadores probabilísticos concebidos a partir de conceitos biológicos, de modo que há uma tendência de que os indivíduos representem soluções cada vez melhores à medida que o processo avança.

Desde sua criação há um interesse crescente na utilização dos AG como uma ferramenta para resolver problemas complexos de otimização (GEN e CHENG, 1997). E embora sejam mais gerais e abstratos do que outros métodos de otimização e, nem sempre ofereçam a solução ideal, eles são considerados flexíveis e aplicáveis a uma ampla variedade de problemas (ASLLANI e LARI, 2007).

Os algoritmos genéticos simulam processos naturais de sobrevivência e reprodução das populações, essenciais em sua evolução. Na natureza, indivíduos de uma mesma população competem entre si, buscando principalmente a sobrevivência, seja através da busca de recursos como alimento, ou visando a reprodução. Os indivíduos mais aptos terão um maior número de descendentes, ao contrário dos indivíduos menos aptos. São requisitos para a implementação de um AG:

- Representar das possíveis soluções do problema no formato de um código genético;
- Uma população inicial que contenha diversidade suficiente para que o algoritmo possa combinar características e gerar novas soluções;
- Um método para medir a qualidade de uma potencial solução;
- Um procedimento de combinação de soluções para gerar novos indivíduos na população;
- Um critério de escolha das soluções que permanecerão na população ou que serão retirados desta;
- Um procedimento para introduzir periodicamente alterações em algumas soluções da população. Desse modo mantém-se a diversidade da população e a possibilidade de se produzir soluções inovadoras para serem avaliadas pelo critério de seleção dos mais aptos.

O algoritmo genético possui como principais componentes a população, função aptidão (que neste trabalho será a função silhueta), seleção, cruzamento e mutação.

5.2.6.1 Buscas Locais

A finalidade de empregar busca local numa num algoritmo heurístico, e em particular num algoritmo genético (AG), é com o intuito de refinar as soluções obtidas pelo AG. Em testes empíricos efetuados, foi observado que as buscas locais tendem a melhorar bastante as soluções encontradas pelo genético. Neste trabalho usamos a busca local por *inversão individual* e busca local *troca entre pares*.

A busca local por *Inversão Individual* tenta melhorar a solução corrente analisando soluções próximas a ela. Para isso, essa busca permuta o valor de cada elemento do indivíduo (1 por 0, ou 0 por 1), um por vez, e calcula o novo valor da função de aptidão. Porém, o algoritmo só aceita a mudança, se o novo valor da função de aptidão for maior (melhor) que o valor anterior.

A busca local *Troca Entre Pares* é uma busca intensiva que troca o status de dois elementos do indivíduo com valores diferentes. A idéia desta busca local, ao contrário da anterior, é tentar encontrar indivíduos diferentes sem alterar o número de clusters pais encontrados pelas melhores soluções geradas pelo AG. Devido ao elevado tempo computacional exigido por este módulo, esta busca é feita somente ao final do algoritmo e somente nos três melhores indivíduos do CE final. O objetivo da busca local *Troca Entre Pares* é investigar possíveis soluções diferentes com o mesmo número de clusters pais.

5.2.6.2 O algoritmo utilizado

O Algoritmo Genético para o Problema de Clusterização Automática (AGPCA) aqui utilizado, é um método composto de duas fases: Fase de Construção e o Módulo Evolutivo. A fase de Construção tenta reduzir as dimensões dos dados de entrada do problema e ao mesmo tempo gerar soluções iniciais de boa qualidade. Nos Algoritmos Genéticos tradicionais isto não ocorre, pois a população inicial é normalmente gerada de forma aleatória. Estas duas metas da primeira fase são obtidas através de um algoritmo construtivo baseado em conceitos de componentes conexas. A segunda fase do AGPCA é composta por um Algoritmo Genético com buscas locais e que utiliza conceitos de memória adaptativa, cujo objetivo é buscar a melhor configuração de uma solução possível.

A Fase de Construção do AGPCA é uma etapa inicial que tem por objetivos tentar reduzir a cardinalidade dos dados de entrada do problema, bem como facilitar a geração de soluções iniciais de boa qualidade para o Algoritmo Genético. O procedimento é baseado no critério de densidade (GARAI e CHAUDHURI, 2004; TSENG e YAN, 2001). A ideia é substituir grupos de objetos da base de dados cuja similaridade é considerada alta, por um único objeto (cluster inicial) que represente o grupo.

1. **Procedimento Construtivo (X, u)**
2. **PARA** $i = 1$ **até** n **FAÇA**
3. $d_{\min}(x_i) = \min \| x_i - x_j \|, i \neq j, j = 1,...n$
4. **FIM PARA**
5. $d_{medio} = \frac{1}{n} \sum_{i=1}^{n} d_{\min}(x_i)$
6. $r = u * d_{medio}$
7. **PARA** $i = 1$ **até** n **FAÇA**
8. $N_i = circulo\ (x_i, r)$
9. $T = T \bigcup N_i$
10. **FIM PARA**
11. *ordenar T em ordem decrescente*
12. $i = 1$
13. **ENQUANTO** ($T \neq \emptyset$) **FAÇA**
14. $B_i = próximo(N_j \in T)$
15. $T = T - \{N_j\}$
16. $i = i + 1$
17. **FIM ENQUANTO**
18. *retornar* $B = \{B_1, B_2, ...B_t\}$, *os clusters parciais.*
19. **FIM construtivo**

Figura2.1: Descrição do procedimento construtivo.
Fonte: : Dib e Ochi (2010).

Neste trabalho usamos como medida de similaridade a distância euclidiana entre dois pontos. A redução do tamanho da entrada (pré-processamento) é realizada agrupando-se em um mesmo cluster os pontos pertencentes a uma região densa, como mostra a figura 2.1. Inicialmente, nas linhas 2, 3 e 4, para cada ponto é definido a menor distância a outro ponto qualquer. Depois, na linha 5, é calculada a média destas distâncias, denominada d_{medio}. Então, cada ponto $x_i \in X$ é considerado o centro de um círculo cujo valor do raio é $r = u * d_{médio}$, onde u é um parâmetro de entrada. Logo após, na linha 8, é calculado o conjunto de pontos contidos no círculo de centro x_i e raio r (N_i = circulo (x_i, r)). Estes valores são colocados em uma lista T que é ordenada em ordem decrescente. Os elemento de T são considerados os clusters parciais $B = \{B_1, B_2,...B_t\}$. Para que os clusters não possuam elementos em comum, toda vez que um círculo é selecionado, todos os pontos do seu interior não podem mais entrar em outro círculo. Com este procedimento as regiões mais densas são selecionadas.

Considere os clusters iniciais gerados na fase de construção como $B = \{ B_1, B_2,...,B_m\}$ e seja v_i , $i = 1,2,...,m$ o centroide (equação 2.4) de cada cluster B_i. Para representar uma solução é utilizada uma cadeia binária de m posições. Por exemplo, se m = 7, então uma cadeia binária poderia ser {0110010}. Se o valor correspondente ao B_i na cadeia binária for igual a 1, isso significa que o cluster inicial B_i faz parte da solução como cluster pai. Se o valor correspondente ao B_i na cadeia binária for igual a 0, B_i é considerado um cluster filho. Os clusters filhos são unidos aos clusters pais utilizando o critério de menor distância entre os centroides. A cada união, o valor do centroide do cluster pai é recalculado. No final, todos os clusters filhos são unidos aos clusters pais para gerar uma solução completa. Portanto, nesta representação, a cada novo individuo (solução) poderemos ter um número distinto de clusters pais, que não se alteram depois deste processo. Os clusters gerados após esse processo são denominados clusters finais $C = \{C_1, C_2,,..., C_k\}$.

Para avaliar a qualidade da solução encontrada, é utilizada uma função de Aptidão *F*, nesse caso foi usada a função silhueta. O índice silhueta define a qualidade dos agrupamentos com base na proximidade entre os objetos de um determinado grupo e na distância desses objetos ao grupo mais próximo. Publicado por Rousseeuw (1987), o critério silhueta original é calculado para cada objeto de um grupo, mostrando quais objetos estão bem situados no mesmo e quais seriam melhor situados em outro grupo. O critério silhueta pode ser calculado com qualquer medida de similaridade/dissimilaridade.

Seja um objeto x_i e o grupo C_A tal que $x_i \epsilon\ C_A$. Seja a(x_i) a dissimilaridade média do objeto x_i em relação a todos os demais objetos do grupo C_A e b(x_i) a menor dissimilaridade média de x_i dos objetos de um outro grupo C_B, $C_B \neq C_A$. A silhueta de um objeto x_i é obtida pela equação 6.

$$silhueta(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \qquad (2.6)$$

Se x_i for o único objeto do grupo a que pertence, então silhueta (x_i) se torna indefinido e uma escolha neutra seria assumir silhueta (x_i)= 0 (ROUSSEEUW 1987). O resultado obtido pela equação (2.6) estará no intervalo [-1,1]. Se um objeto está bem situado em seu grupo, sua silhueta será positiva, caso contrário, se o objeto estiver mais próximo de outro grupo, ela será negativa. Como a silhueta depende apenas do agrupamento resultante e não do algoritmo de agrupamento empregado, ela pode ser usada para executar uma análise dos grupos obtidos, para comparar partições geradas por diferentes algoritmos, ou diferentes execuções do mesmo algoritmo, em um mesmo conjunto de dados. O critério silhueta é mais apropriado para agrupamentos volumétricos com grupos compactos e separados.

O módulo evolutivo é composto de um Algoritmo Genético Híbrido incluindo módulo de Memória Adaptativa e de duas buscas locais. A racionalidade de utilizar Algoritmos Genéticos é gerar aleatoriamente um número qualquer de clusters pais. Como o número ideal de clusters é um dos objetivos do problema, o AG permite gerar a cada iteração, números diferentes de clusters pais.

Para construir a população inicial, o algoritmo gera um número bem maior de indivíduos (dez vezes o tamanho da população) e escolhe aqueles com os maiores valores da função de aptidão. Este procedimento permite começar o AG com uma população de melhor qualidade.

No Algoritmo Genético, são utilizados operadores clássicos como o operador de mutação e proposto uma forma alternativa de seleção por cruzamentos.

Para efetuar a seleção dos indivíduos para cruzamento, são utilizados dois operadores que se alternam. O primeiro operador escolhe aleatoriamente os dois indivíduos pais dentre os 60% com melhores valores da função de aptidão. E o segundo operador, escolhe um indivíduo aleatoriamente dentre os 60% com melhores valores da função de aptidão e o outro, entre os 40% restantes. O operador de cruzamento utilizado no AGPCA é o cruzamento de dois pontos, que atua sobre dois indivíduos com genótipos diferentes. Este operador funciona da seguinte maneira: pares de pontos de cruzamento são obtidos de forma aleatória e os valores dos indivíduos localizados entre cada par de pontos de cruzamento são trocados. Os indivíduos são submetidos ao cruzamento com probabilidade p_c. Os dois pontos de corte definem os segmentos dos vetores que serão trocados entre os mesmos para gerar novos indivíduos.

O operador de mutação realiza trocas aleatórias de alguns valores dos indivíduos com probabilidade p_m, com o intuito de pesquisar novas áreas do espaço de busca, a partir de indivíduos selecionados.

Após a aplicação dos operadores de cruzamento e mutação, os descendentes que obtiverem valores da função de aptidão melhores que os valores da população atual são inseridos na nova população.

A cada t iterações (onde *t* é um parâmetro de entrada), os melhores indivíduos da população passam pela busca local *Inversão Individual*. O objetivo é intensificar a procura de soluções diferentes no conjunto de soluções existentes. A cada iteração, a melhor solução é armazenada no conjunto elite (CE).

No final das iterações do AG, o CE passa pela busca local *Troca entre Pares*.

6 RESULTADOS E DISCUSSÃO

Nesta parte do trabalho foram utilizadas as cinquenta e uma propriedades positivas para a fascíola. A localização destas propriedades consta na figura 2.2.

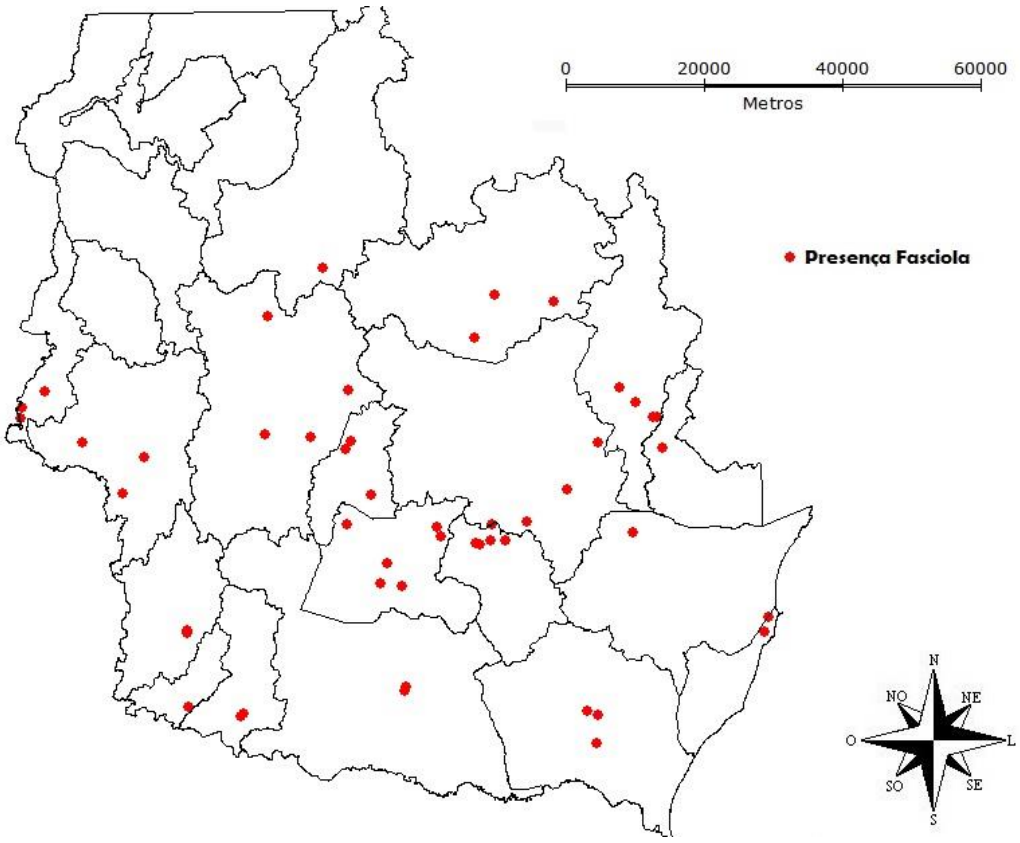


Figura 2.2: Distribuição das propriedades estudadas com registro de casos de ocorrência de fascíola hepática bovina no sul do estado do Espirito Santo.

Para verificar qual técnica seria a mais apropriada para encontrar o número de *Clusters*, o *K-means* com *Elbow* e o AGPCA foram executados e comparados, utilizando as quarenta e oito instâncias já mencionadas (Anexo A) (Tabela 2.1).

Na tabela 2.2 tem-se a comparação entre as heurísticas utilizadas nas instâncias. Na primeira coluna foram colocadas as instâncias, na segunda o número de clusters previamente conhecidos, na terceira o número de clusters obtido pelos métodos *k-means* e *elbow*, na quarta a numeração da figura 2.2 onde está representado o gráfico dos métodos *k-means* e *elbow* (Anexo A) e na quinta coluna o número de clusters obtido pelo algoritmo com a função silhueta.

Tabela 2.1: Comparação entre os métodos *K-means* com *elbow* e algoritmo genético com silhueta, com o padrão ouro**.**

Instância	Nº conhecido de clusters	K-means	Gráficos Anexo A	Silhueta
100p2c	2	2 ou 3	A.1	2
100p2c1	2	2	A.2	2
100p3c	3	2 ou 4	A.3	3
100p3c1	3	3	A.4	3
100p5c1	5	4	A.5	5
100p7c	7	4	A.6	7
100p7c1	7	6	A.7	7
100p10c	10	3	A.8	10
200p2c1	2	2	A.9	2
200p3c1	3	5	A.10	3
200p4c	4	4	A.11	4
200p4c1	4	5	A.12	3
200p7c1	7	3	A.13	6
200p12c1	12	3	A.14	12
300p2c1	2	2	A.15	2
300p3c	3	3	A.16	3
300p3c1	3	3	A.17	4
300p4c1	4	4	A.18	4
300p4c2	4	4	A.19	4
300p6c1	6	4	A.20	5
300p10c1	10	7	A.21	9
300p13c1	13	7	A.22	12
400p3c	3	3	A.23	3
400p4c1	4	5	A.24	4
400p17c1	17	3	A.25	16
500p3c	3	3	A.26	3
500p4c1	4	3	A.27	4
500p6c1	6	4	A.28	6
600p3c1	3	3	A.29	3
600p15c	15	5	A.30	15
700p4c	4	4	A.31	4
800p4c1	4	4	A.32	4
800p10c1	10	3	A.33	10
800p18c1	18	2	A.34	18
800p23c	23	5	A.35	23
900p5c	5	6	A.36	5
900p12c	12	7	A.37	12
1000p5c1	5	4	A.38	6
1000p6c	6	8	A.39	6
1000p14c	1	9	A.40	14
1000p27c1	27	8	A.41	27
1100p6c1	6	7	A.42	6
1300p17c	17	10	A.43	17
1500p6c	6	8	A.44	6
1500p6c1	6	7	A.45	7
1500p20c	20	11	A.46	20
2000p9c1	9	11	A.47	10
2000p11c	11	8	A.48	11

O que se observa é que para o nosso conjunto de dados o algoritmo genético com a função silhueta foi mais preciso.

Foi verificado através da execução do algoritmo que o número de Clusters formados na área de estudo são 5 (cinco). A figura 2.3 mostra a localização de cada um dos cinco clusters encontrados. Na tabela 2.2, constam os municípios que compõe cada cluster.

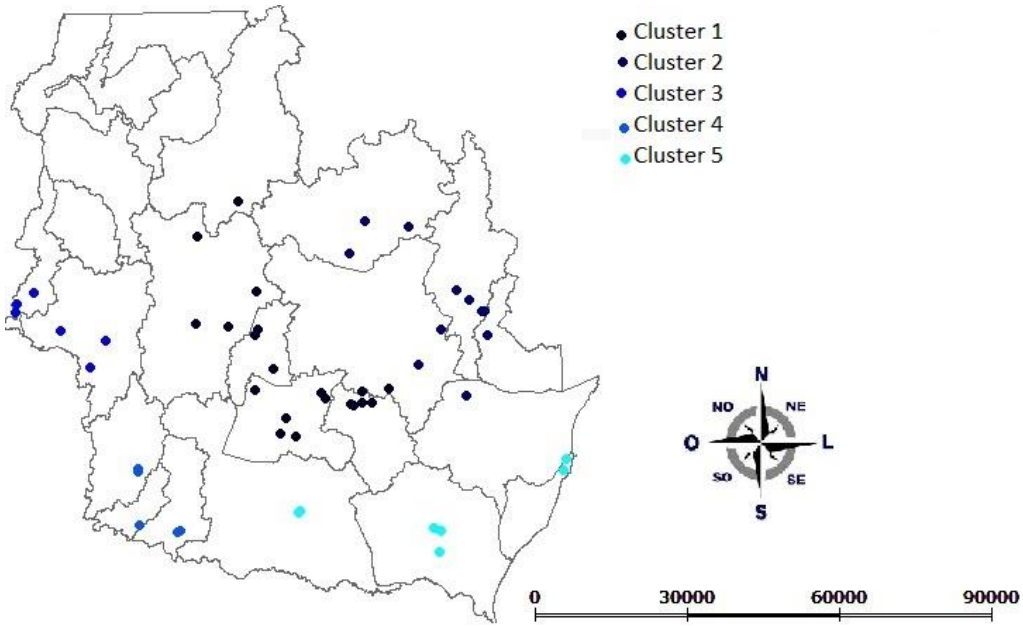


Figura2.3: Localização dos clusters obtidos pelo algoritmo genético.

Tabela 2.2: Distribuição dos *clusters* encontrados a partir do algoritmo genético e os respectivos municípios que pertencem a cada *cluster*.

Cluster	Municípios
1	Cachoeiro de Itapimirim, Muqui, Jeronimo Monteiro, Atílio Vivacqua, Alegre e Muniz Freire
2	Cachoeiro de Itapimirim, Rio Novo do Sul, Vargem Alta, Castelo e Itapimirim
3	Guaçui e Dores do Rio Preto
4	Bom Jesus do Norte, Apiacá e São José dos Calçados
5	Presidente Kennedy, Mimoso do Sul e Marataízes

Uma observação a ser feita é: validação dos dados foi realizada tendo como parâmetro as instâncias, chamadas de padrão de ouro na figura 2.4. Através do coeficiente de correlação de *spearman*, foi verificada que a correlação entre o método silhueta e o padrão ouro é de 0,99 [p-valor < 0,001], e que a correlação entre o método *k-means* e o padrão ouro é de 0,48 [p-valor < 0,001], ou seja, ambos os métodos são significativos.

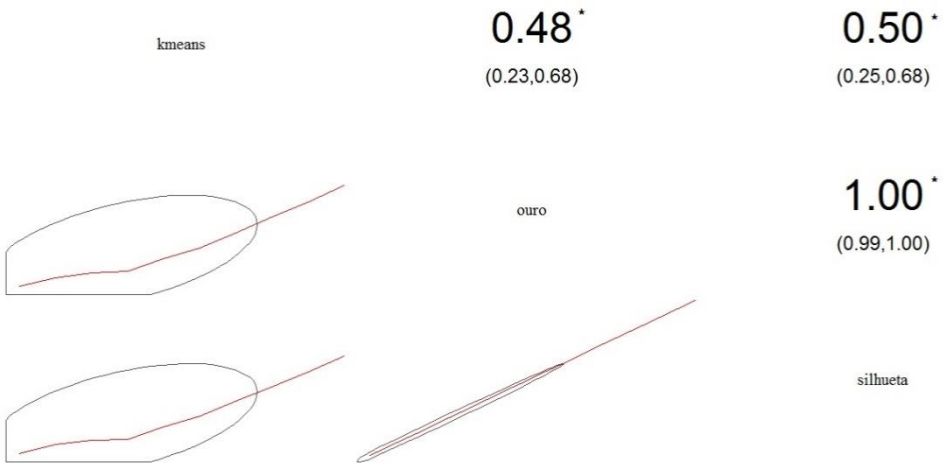


Figura 2.4: Correlação entre as metodologias utilizadas (métodos *k-means* com *elbow* e função silhueta) e o padrão ouro.

7 Conclusão

- A correlação garante que tanto o *k-means* quanto o algoritmo genético são satisfatórias para esse conjunto de dados, haja visto que estamos trabalhando em R^2, portanto não podemos afirmar que em R^n essas metodologias possam ser equiparadas.
- O método *k-means* juntamente ao método *elbow* é significativo e seu uso se justifica em diversas áreas como a epidemiologia em virtude de sua simplicidade.
- O algoritmo genético com a função silhueta é significativo e mais preciso, mas possui um custo computacional maior (quando comparado ao método *k-means* juntamente ao método *elbow)* e há necessidade de implementação, o que o torna menos viável a pesquisadores da área de epidemiologistas em geral.
- Para a escolha do melhor método deve-se levar em conta o custo benefício e a necessidade de precisão.

CONCLUSÃO GERAL

A fim de encontrar os fatores de risco para a fascíola hepática bovina no Sul do estado do Espírito Santo foram analisados os seguintes fatores em cada uma das propriedades: ocorrência de casos anteriores da patologia, existência de outros hospedeiros, presença do hospedeiro intermediário *lymnaea* e existência de áreas alagadas. A ocorrência de casos anteriores e outros hospedeiros nas propriedades são caracterizadas como fatores de risco epidemiológico para a fascíola.

A fascíola apresenta uma forte componente espacial, isto é, uma vez encontrada a doença em um lugar a probabilidade de um vizinho também ser acometido por ela é alta. Consequentemente quanto mais próximos os eventos (propriedades com fascíola) mais prioritária é uma intervenção sanitária.

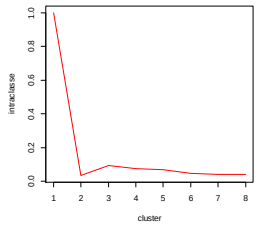
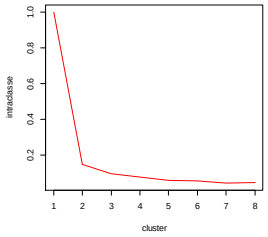
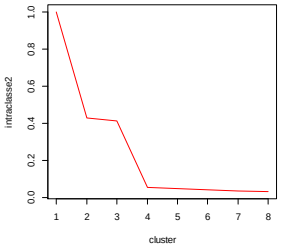
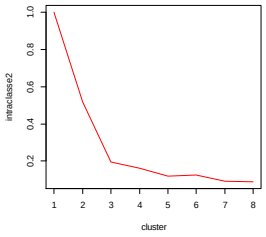
Para análise dos dados foi necessária a utilização de modelos de regressão espacial e métodos computacionais intensivos como heurísticas, em virtude da complexidade de modelagem do problema.

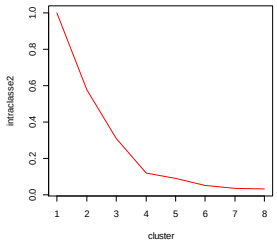
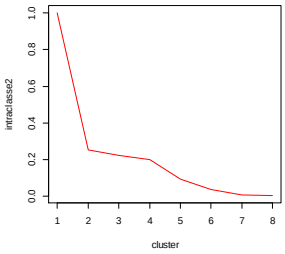
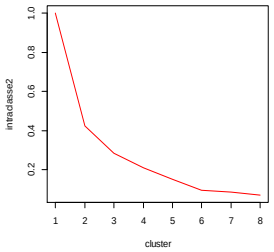
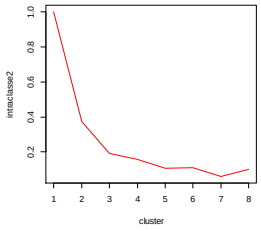
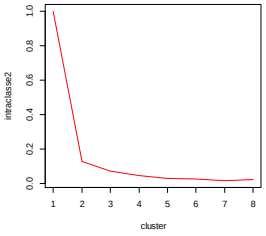
A estatística de kernel nos mostra que as propriedades pertencentes ao cluster 1 necessitam de um atendimento prioritário. Tal resultado é validado pelo algoritmo genético.

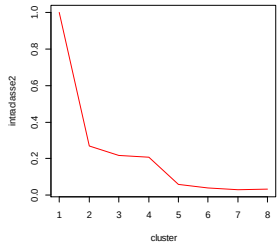
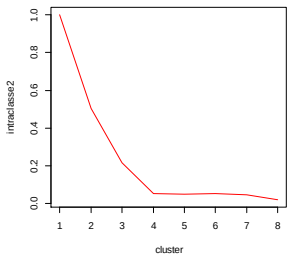
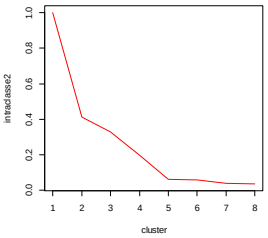
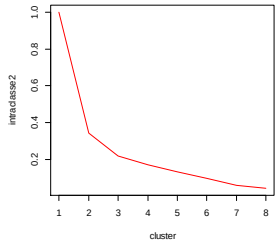
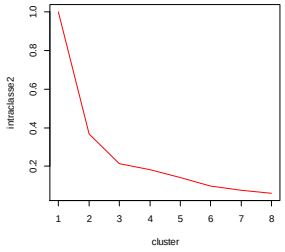
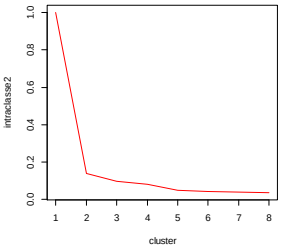
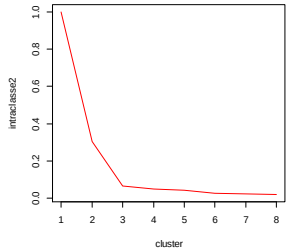
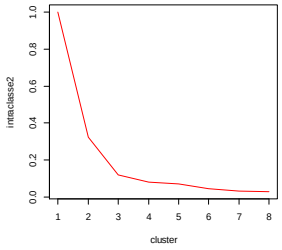
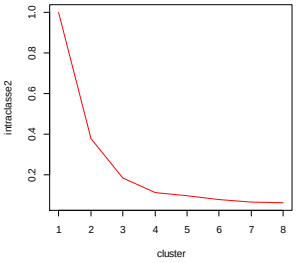
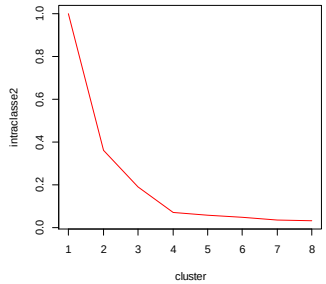
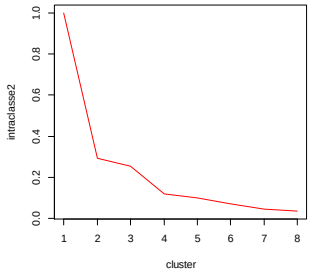
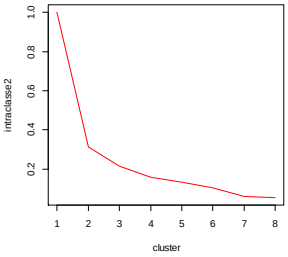
Diante deste panorama é cada vez mais eminente a necessidade de uma equipe multidisciplinar onde atuem matemáticos, médicos veterinários, geógrafos, etc, em prol de causas em comum.

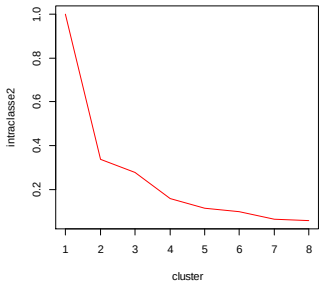
ANEXOS

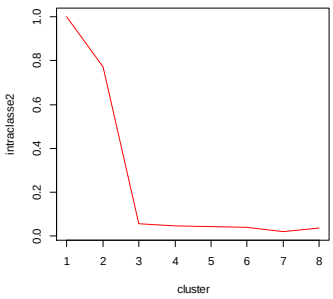
Anexo A - Gráficos com o número de agrupamentos em função das variâncias intra-grupos estimadas através do método *k-means* utilizando as instâncias.

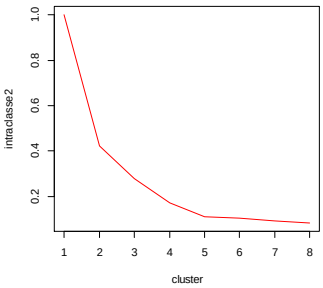
<table>
<tr><td>1)
 
</td><td>2)
 
</td><td>3)
 
</td></tr>
<tr><td> 

4)</td><td>5)
 
</td><td>6)
 
</td></tr>
<tr><td>7)
 
</td><td>8)
 
</td><td>9)
 
</td></tr>
<tr><td>10)</td><td>11)</td><td>12)</td></tr>
</table>

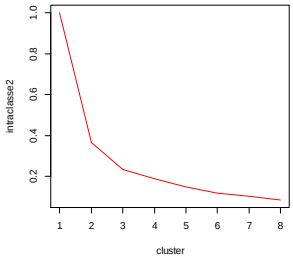
		
13) 	**14)** 	**15)** 
16) 	**17)** 	**18)** 
19) 	**20)** 	**21)** 
22)	**23)**	**24)**



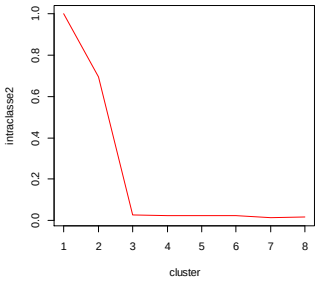




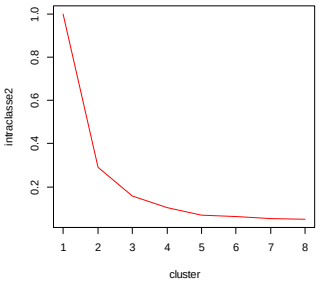
25)



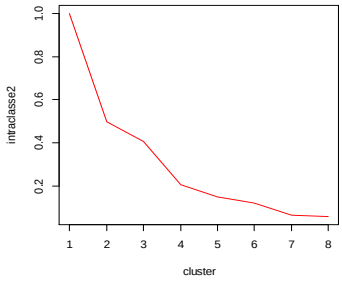
26)



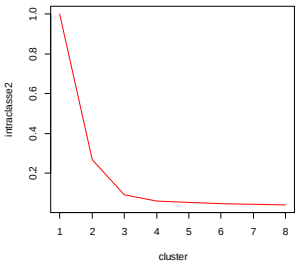
27)



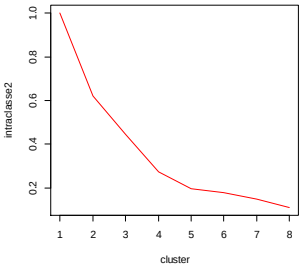
28)



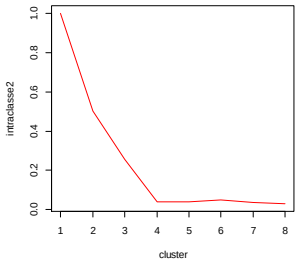
29)



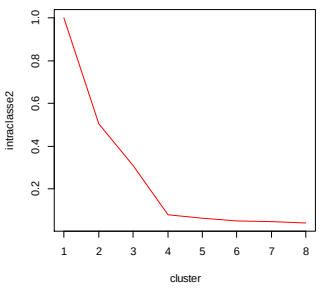
30)



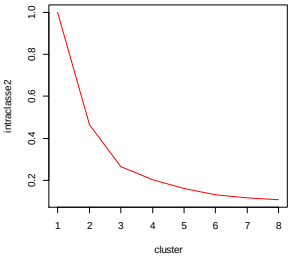
31)



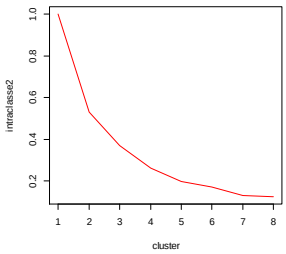
32)

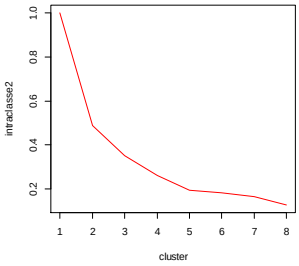


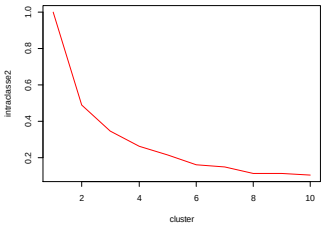
33)



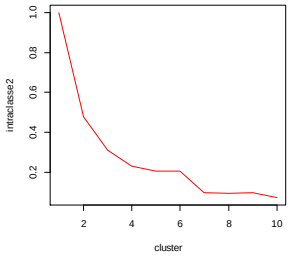
34) **35)** **36)**



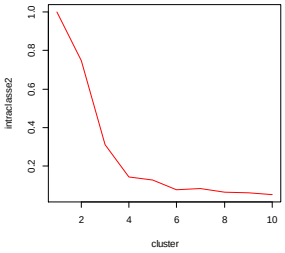




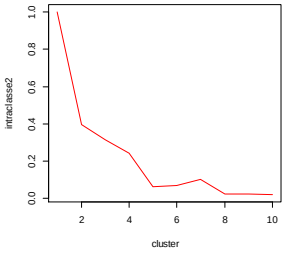
37)



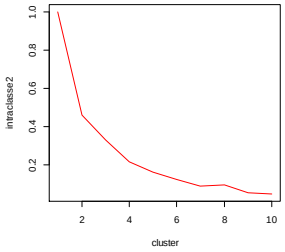
38)



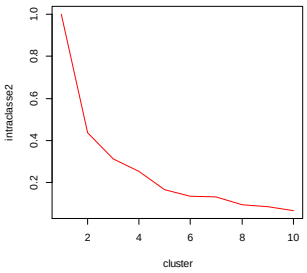
39)



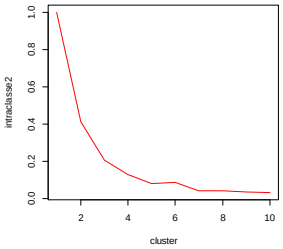
40)



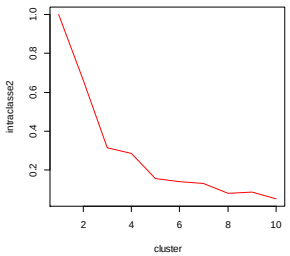
41)



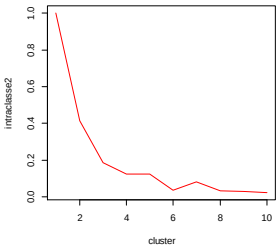
42)



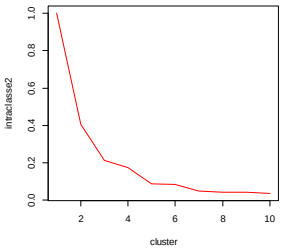
43)



44)



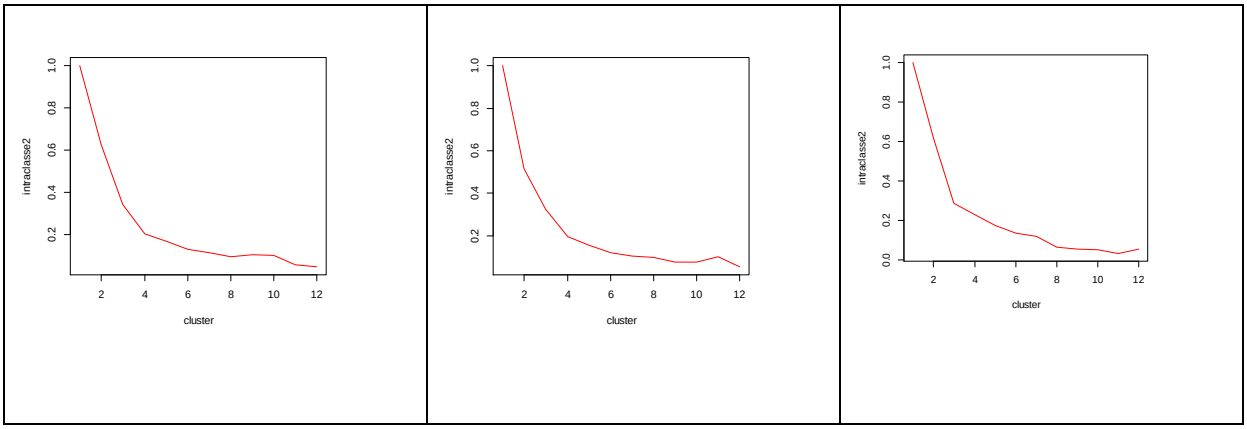
45)



46)

47)

48)


intraclasse2
cluster
intraclasse2
cluster
intraclasse2
cluster

REFERÊNCIAS

ACHA, P.N.; SZYFRES, B. Zoonosis y Enfermedades transmisibles comunes al hombre y a los animales. Washington, D.C. Organização Mundial da Saúde, pp 98-110. 1977.

ASLLANI, A.; LARI, A. Using genetic algorithm for dynamic and multiple criteria web-site optimizations, European Journal of Operational Research, 2007.

AVELAR, B. R. ; FREITAS, D. F. ; MARTINS, I. V. F. ; GOMES, DS ; DOS SANTOS, G.M.A.D.A. ; SANTOS, A. R. . Zoneamento bioclimatológico e mapas de prevalência para fasciola hepatica no estado do espírito santo, brasil. In: XVIII Congresso Brasileiro de Parasitologia Veterinaria, 2014, Gramado. Anais do XVIII Congresso Brasileiro de Parasitologia Veterinaria, v. 1. p. 1, 2014.

AVELAR, B. R. ; BERNARDO, C.C. ; DE LEÃO, A.G.C. ; CAMARGO, P.F. ; MARTINS, I. V. F. . Estudo epidemiológico da fasciolose em municipio do Espirito Santo. In: XVII Congresso Brasileiro de Parasitologia Veterinaria, 2012, Sao luis. Anais do XVII Congresso Brasileiro de Parasitologia Veterinaria, v. 1, 2012.

BAILEY, T. C. E GATRELL, A. C., Interactive spatial data analysis Longman Higher Education, Harlow, 1995.

BERNARDO, C.C.; DEMONER, L.C.; FRAGA, J.C.L.; GONÇALVES, M.F.; DONATELE, D.M.; MARTINS, I.V.F. Avaliação comparativa entre a técnica de sedimentação fecal e achados de *F. hepática* em fígados bovinos. Anais da XXXIV Semana Capixaba do Médico Veterinário. 2007 .

BOLFARINE, H. e SANDOVAL, C. Introdução a Inferencia Estatistica. Sociedade Brasileira de Matematica, 2001.

BORAY, J. C. (1966). Studies on the relative susceptibility of some Lymnaeids to infections with Fasciola hepatica and F. gigantica and on the adaptation of Fasciola spp. Ann. Trop. Med. Parasitol., 10: 114 – 124.

BORAY, J.C. Experimental fascioliasis in Australia. Advances in Parasitology. 7 ed., 1977, p.95–210.

BORAY, J. C. In: Gaafar, S.M.H.W.E.M.R.E. (Ed.). Flukes of Domestic Animals in parasite, pests and predators. *Elsevier*, New York, p. 179-218, 1985.

BUSETTI, E.T. Informações adicionais sobre a fasciolose hepática em Curitiba (estado do Paraná, Brasil). Revista do Instituto de Medicina Tropical. v.24, n.2, p.104-106, 1982.

CALRETAS, S.; LAIZ, M.; SIMÃO, A.; CARVALHO, A.; RODRIGUES, A.; SÁ, A.; SANTOS, A.; SANTOS, R.; Da SILVA, J.A.P.; REIS, C.; ALMIRO, E.; PORTO, A. Seis casos de fasciolíase hepatica. *Medicina Interna*, v.10, n.4, 2003.

CORAL, R.P.; MASTALIR, E.T.; MASTALIR, F.P. Retirada de *Fasciola hepatica* da via biliar principal por coledocoscopia – relato de caso. Revista do Colégio Brasileiro de

Cirurgiões. [periódico na Internet] 2007; v.34, n.1. Disponível em URL: http:/ /www.scielo.br/rcbc

CRUZ, M. D. ; OCHI, L. S. . Um Algoritmo Evolutivo com Memória Adaptativa para o Problema de Clusterização Automática. Learning and Nonlinear Models, v. 8, p. 227-239, 2011.

DITTIMAR, K.; TEEGEN, W.R. The presence of *Fasciola hepatica* (Liver-fluke) in humans and cattle from a 4.500-year old archaeological site in the Saale-Unstrut Valley, Germany. Memórias do Instituto Oswaldo Cruz. Rio de Janeiro, v.98, n.1, p. 141-143, 2005.

DYM, C.L,. IVEY, E. S - Principles of Mathematical Modeling, Academic Press, 1980.

DOBSON, A. Anintroductiontogeneralized linear models. Chapman and Hall, New York, 2 edition, 1990.

DRUCK, S. ;. CARVALHO, M.S; CÂMARA, G.; MONTEIRO, A.M.V. "Análise Espacial de Dados Geograficos", EMBRAPA, DF, 2004

ECHEVARRIA, F. A. M. Fasciolose: ocorrência, diagnóstico e controle. Agroquímica Santo Amaro. v. 27. 1985.

EL-KOUBA, M. M. A. N. Aspectos gerais da fasciolose e das endoparasitoses em capivaras (*Hidrochaeris hidrochaeris* - LINNAEUS, 1766) e ratões de banhado (*Myocastor coypus* – MOLINA, 1782) residentes em três parques do estado do Paraná. Dissertação (Mestrado) Universidade Federal do Paraná, Curitiba, 89p, 2005.

FRAGA, J.C.L. Incidência de fasciolose hepática em bovinos abatidos no sul do estado do Espírito Santo. Curso de pós-graduação - Instituto Qualittas, 2008.

FREITAS, D.F." Análise Espacial do Risco de Fasciolose Bovina no Estado do Espirito Santo por Meio dos Sistemas de Informações Geográficas", Dissertação de Mestrado, CCA-UFES, 2013.

GAETAN, C., & GUYon, X. *Spatial Statistics and Modeling*. New York: Springer. 2010.

GARAI, G. , CHAUDHURI, B. B. A novel genetic algorithm for automatic clustering. Pattern Recognition Letters, pp. 173-187. 2004.

GEN, M. e CHENG, R. Genetic Algorithms and Engineering. Design, 1997.

GLOVER, F. W. e KOCHENBERGER, G.A. Handbook of Metaheuristics. Springer Science & Business Media, 2003.

GOMES, F. F., SANTOS, J.A., OLIVEIRA, F. C. R.; LOPES, C.W.G. Fasciolose bovina na microrregião de Campos dos Goytacazes–RJ, Brasil. Estudos Preliminares. Anais do XII Congresso Brasileiro de Parasitologia Veterinária. 1: 167-168. 2002.

GOUTTE, C.; TOFT, P.; ROSTRUP, E; NIELSEN, F.A.; HANSEN, L. K. (March 1999). "OnClusteringfMRI Time Series".

GUIMARÃES, M.P. *Fasciola hepatica*. 2003. In: Oliveira, A.A., et al. Estudo da prevalência e fatores associados à fasciolose no Município de Canutama, Estado do Amazonas, Brasil. Epidemiologia e Serviços de Saúde. v.16, n.4, p.251-259. 2007.

GUNST, R.F., MASON, R.L., Regression Analysis and Its Application: A Data-Oriented Approach, Marcel Dekker, 1980.

HASTIE, T. and Tibshirani, R. Generalised Additive Models. 1990.

HOLLAND, J. H. Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press, 1975.

KLEIMAN, F.; PIETROKOVSKY, S.; GIL, S.; WISNIVESKY-COLLI, C. Comparação de dois métodos coprológicos para diagnóstico da fasciolose. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia,* v.57, n.2, p.181-185, 2005.

KNOX, E. G., 1991. Spatial and temporal studies in epidemiology. In: *Oxford Textbook of Public Health: Methods of Public Health* (W. W. Holland, R. Detels & G. E. Knox, eds.), vol. 2, pp. 95-105, Oxford: Oxford University Press.

LESSA, C. S. S., SCHERER, P. O., VASCONCELLOS, M. C., FREIRE, L. S., SANTOS, J. A. A., FREIRE, N. M. S. Registro de Fasciola hepatica em equinos (Equus caballus), caprinos (Capra hircus) e ovinos (Ovis aries), no município de Itaguaí, Rio de Janeiro Brasil. Revista Brasileira Ciência Veterinária, Niterói. 1:63-64. 2000.

LOPES, H.S.; RODRIGUES, L.C.A.; STEINER, M.T.A.; *Meta-Heuristica em Pesquisa Operacional*. Curitiba: Omnipax.473p. 2013.

MACQUEEN, J.B.; "Alguns métodos de classificação e análise das observações multivariadas, *Anais do 5 º Berkeley Simpósio de Estatística Matemática e Probabilidade "* , Berkeley, UniversityofCalifornia Press, 1: 281-297. 1967.

MARTINS, I.V.F. Novas tecnologias em Ciências Agrárias. Editores Waldir Cintra de Jesus et al. Alegre – ES. p. 245-251. 2007.

MARTINS,, I. V. F. ; AVELAR, B. R. ; BERNARDO, C.C. ; DE LEÃO, A.G.C. ; PEREIRA, M.J.S. Distribution of bovine fasciolosis and identification of associated factors in the south of the state of Espírito Santo, Brazil: an update. Revista Brasileira de Parasitologia Veterinária (Impresso), v. 23, p. 23-29, 2014.

MATTOS, M. J. T.; UENO, H.; GONÇALVES, P. C.; ALMEIDA, J. E. M. Ocorrência estacional e bioecologia de *Lymnaea columella* Say, 1817 em habitat natural no Rio Grande do Sul. Revista Brasileira de Medicina Veterinária. v.19, p.248-52, 1997.

MEDRONHO R; BLOCH KV; LUIZ RR; WERNECK GL (eds.). Epidemiologia. Atheneu, São Paulo, 2009, 2ª Edição.

MUNGUÍA-XÓCHIHUA, J.A.; IBARRA-VELARDE, F.; DUCOING-WATTY, A.; MONTENEGRO-CRISTINO, N.; QUIROZ-ROMERO, H. Prevalence of *Fasciola hepatica* (ELISA and fecal analysis) in ruminants from a semi-desert área in the northwest of Mexico. Parasitology Research. v.101, p.127-130, 2007.

NELDERAND, J.; WEDDERBURN, R. Generalized linear models. Journalofthe Royal StatisticalSociety A, 135:370,n.84, 1972.

PILE, E.; SANTOS, J. A. A.; PASTORELLO, T.; VASCONCELLOS M.C. *Fasciola hepatica* em búfalos (*Bubalus bubalis*) no município de Maricá, Rio de Janeiro, Brasil. Brazilian Journal of Veterinary Research and Animal Science, v. 38, n. 1, p. 42-43, 2001.

RAYWARD-SMITH, V. J. ; OSMAN, I. H.; REEVES, C. R.; SMITH, G. D. Modern Heuristic Search Methods, Wiley, Inglaterra, 1996.

REID, J. F. S.; DARGIE, J. D. Como os estágios adultos da Fasciola hepatica afetam a saúde e a produtividade do bovino. A Hora Veterinária, n.1, p.23-26, 1995.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics , vol. 20, 1987.

SERRA-FREIRE, N.M. Fasciolose hepática. A Hora Veterinária, v. 1, p.13-18, 1995.

SVS (Secretaria de Vigilância em Saúde). Detecção de casos humanos de *Fasciola hepatica* no estado do Amazonas**.** *Boletim eletrônico EPIDEMIOLÓGICO*, ano 05, n. 5, 2005.

TASSINARI, W. S. ; LORENZON, Maria Cristina ; PEIXOTO, Eduardo . Spatial regression methods to evaluate beekeeping production in the state of Rio de Janeiro. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, v. 65, p. 553-558, 2013.

TSENG, L. Y., YANG, S. B (2001). A genetic approach to the automatic clustering problem; Pattern Recognition 34, pp. 415- 424.

WERNECK G.L & STRUCHINER C.J. Estudos de agregados de doenças no espaço-tempo: conceitos, técnicas e desafios. *Cadernos de Saúde Pública 13*(4): 611-624, 1997.