Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

DETERMINAÇÃO DO RISCO DE FOGO UTILIZANDO ALGORITMOS DE CLASSIFICAÇÃO

Rodrigo Smarzaro da Silva
smarzaro@ufv.br

RESUMO

Este trabalho tem como objetivo utilizar técnicas de mineração de dados em dados de ocorrências de incêndios florestais no estado de Minas Gerais para prever o risco de fogo de uma determinada região. O risco de fogo é uma estatística utilizada pelo INPE para determinar a probabilidade de ocorrência de um incêndio e precisa de uma série de informações para ser calculada. A proposta é utilizar uma abordagem baseada nos algoritmos de classificação para classificar os dados em relação ao risco de fogo de uma maneira mais simples.

1 Introdução

A quantidade de informação geográfica cresce em ritmo acelerado. O volume e cobertura de imagens de satélite, sistemas de sensoriamento remoto com alta resolução espectral, espacial e temporal, dispositivos de monitoramento ambiental, dispositivos que coletam dados geográficos usando tecnologias de posicionamento (*Global Positioning System* – GPS, Glonass, Galileo, Com- pass) recolhem uma quantidade enorme de dados geográficos todos os dias [1, 4, 5, 12, 14]. Nos últimos anos, a popularização de celulares, sistemas de navegação automotivos, internet sem fio tem permitido a captura de padrões de movimento de entidades individuais, aumentando a quantidade de dados espaciais e temporais. Apesar do grande volume de dados, é difícil identificar informação relevante [2, 14] o que gera, segundo Miller e Han [14], *"necessidade urgente de novos métodos e ferramentas que possam transformar dados geográficos, de forma inteligente e automática, em informação e, num passo posterior, em conhecimento geográfico"*.

O problema de identificar informação útil no meio deste grande volume de dados tem gerado vários desafios aos métodos tradicionais de extração de informação e conhecimento como a Descoberta de Conhecimento em Bancos de Dados *Knowledge Discovery from Databases* – KDD) e Mineração de Dados (*Data Mining* – DM) [12]. KDD pode ser definido como a " *descoberta de conhecimento implícito,* útil *e previamente desconhecido a partir de grandes bases de dados* " [5, 9, 14]. DM é uma das etapas do processo de KDD e envolve a aplicação de técnicas para obter informação e detectar padrões úteis a partir dos dados [5, 9, 14]. Devido às características dos bancos de dados espaciais, os processos tradicionais de KDD e DM tem se mostrado insuficientes [3, 6, 14, 17, 19]. Entre as razões específicas estão a natureza do espaço geográfico, complexidade dos objetos, seus relacionamentos espaciais e suas transformações temporais, a heterogeneidade e falta de estrutura dos dados georreferenciados e a natureza do conhecimento geográfico [9, 14]. Tais razões motivaram o surgimento de uma nova subárea do KDD chamada Descoberta de Conhecimento Geográfico (*Geographic Knowledge Discovery*– GKD). O GKD pode ser entendido como o processo de extração de informação e conhecimento a partir de bancos de dados geográficos [14]. De forma análoga ao KDD, o GKD possui uma fase de mineração do dados que recebe o nome de Mineração de Dados Geográficos (*Geographic Data Mining* – GDM).

Uma aplicação interessante para o processo de GKD é a análise de dados de fenômenos climáticos. Por exemplo, Pultar et al. [16] propõe

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

um framework para GIS dinâmico utilizado na modelagem de evacuação em casos de incêndios. Cheng e Wang [2] propõe um *framework* para mineração de dados espaço-temporais para predição de incêndios. Lee et al. [13] utilizou técnicas de *clustering* para analisar a trajetória de furações.

2 Referências Básicas

O processo de GDM (ou *Spatial Data Mining*) possui diversas técnicas que podem ser aplicadas de acordo com o tipo de informação que se queira minerar. Dentre estas técnicas estão: classificação espacial, associação espacial, classificação e predição espacial, agrupamento espacial e análise de *outliers* espaciais [14].

A *classificação espacial* é uma técnica que mapeia objetos espaciais em categorias que consideram distância, direção, relacionamentos de conectividade, e morfologia. Um exemplo de utilização desta técnica é a classificação automática de objetos geográficos em categorias específicas (ex. shopping center, supermercado, hospital) ou classificação de ocupação de terras (plantações, florestas, lagos, etc).

A *associação espacial* procura regras de associação entre objetos de forma que o antecedente ou precedente possua um predicado espacial. Um tipo de informação que pode ser descoberto com esta técnica são objetos espaciais que são frequentemente encontrados juntos (ou próximos) como, por exemplo, hospitais e farmácias, consultórios oftalmológicos e óticas, etc.

A *classificação e predição espacial* utilizam algoritmos e técnicas de aprendizado para extrair regras a partir de um conjunto de dados (conjunto de treinamento), e em seguida utilizar estas mesmas regras para verificar um novo conjunto de dados. Shekhar et al. [20] fornece um exemplo ao criar um modelo de predição de onde encontrar ninhos de pássaros tordo-sargento (*Agelaius phoeniceus*) em regiões pantanosas. Um resultado interessante foi a comparação entre as técnicas modernas (que consideram os aspectos espaciais) e as técnicas tradicionais (sem considerar os aspectos espaciais) que mostrou vantagem para as primeiras.

Agrupamento espacial explora relacionamentos espaciais entre objetos para determinar agrupamentos entre eles. Como encontrar o conjunto ótimo de *k* agrupamentos é intratável [14] (onde *k* é um número inteiro menor do que a cardinalidade do banco de dados) várias heurísticas são utilizadas dentro dessa técnica. Exemplos de heurísticas incluem métodos de particionamento (*k-means*, *expectation maximization*), métodos hierárquicos (*splitting* e *aggregation*), métodos baseados em densidade(definem agrupamentos como regiões com grande número de objetos), métodos baseados em grid(dividem o espaço em uma tesselação e usam essa estrutura para os agrupamentos), métodos baseados em modelos(procuram o melhor agrupamento dos dados em relação a formas funcionais específicas) e métodos baseados em restrições(impõem restrições para os agrupamentos ou para os relacionamentos que os definem).

A *análise de* outliers *espaciais* permite verificar objetos que, de alguma forma, são inconsistentes com outros objetos semelhantes. Um exemplo de aplicação desta técnica, mostrado por Ng [15] usa medidas baseadas em distância para identificar trajetórias incomuns baseadas em pontos de entrada e saída, velocidade e geometria para detectar comportamentos indesejados, como roubo de carros.

As técnicas citadas envolvem apenas os aspectos espaciais dos dados. Em vários problemas a dimensão temporal não pode ser ignorada e técnicas de mineração de dados que consideram dados espaciais e temporais têm

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

aparecido na literatura [2, 7, 10, 17]. Essas técnicas envolvem previsão e análise de tendências, mineração de regras de associação, mineração de padrões sequenciais, classificação e agrupamento.

A *previsão e análise de tendências* em dados espaço-temporais teve seu desenvolvimento a partir de algoritmos que levavam em conta dados espaciais ou temporais de forma separada. Várias ferramentas de análise foram estendidas para considerar aspectos espaciais e temporais, tais como séries temporais e geoestatística [2]. Esta técnica pode ser empregada em vários cenários, como por exemplo, previsão de locais de incêndios (e sua possível extensão) a partir de dados históricos e análise da vizinhança das ocorrências.

A *mineração de regras de associação* envolve descoberta de padrões do tipo $X \rightarrow Y$, onde X e Y são conjuntos de atributos espaço-temporais ou atributos convencionais. A ideia é encontrar este tipo de regra onde um conjunto X de atributos, com determinados valores, em um dado instante de tempo levam, com uma certa probabilidade, a ocorrência de determinados valores no conjunto Y, no mesmo instante de tempo. Um exemplo de regra de associação seria descobrir que dado a ocorrência de alagamento em uma região A, uma outra região B também estará alagada. Uma grande dificuldade em encontrar algoritmos eficientes está na explosão exponencial quando são consideradas as dimensões espaciais e temporais juntas.

A *mineração de padrões sequenciais* é parecida com a mineração de regras de associação, mas considera que em uma regra $X \rightarrow Y$, X e Y ocorrem em instantes diferentes de tempo. Esta técnica poderia ser empregada, por exemplo, para descobrir associações interessantes entre dados históricos de ocorrências de epidemias de forma a identificar fatores que contribuem para o alastramento das ocorrências.

As técnicas de *classificação e agrupamento* de dados espaço-temporais funcionam com o mesmo objetivo que as técnicas de agrupamento espacial, ou seja, classificar objetos em grupos que possuem algum determinado conjunto de atributos, com uma certa probabilidade [2]. Dentre as possíveis aplicações estão a visualização de dados de Censo, a exploração de levantamento de dados de saúde, entre outras [8]. Vários algoritmos de agrupamento utilizados amplamente em dados espaciais podem ser estendidos para trabalhar também com dados espaço-temporais, como *K*-means, *K*-medoid e CLARANS [9].

3 Proposta do Trabalho

O risco de fogo [18] é uma estatística levantada pelo INPE que leva em conta diferentes variáveis como ciclo natural de desfolhamento da vegetação, temperatura máxima, umidade relativa mínima do ar, assim como a presença de fogo na região de interesse para calcular a probabilidade de uma região sofrer alguma queimada.

O risco de fogo é calculado da seguinte maneira:

1. Determina diariamente para a área geográfica de abrangência, o valor da precipitação em mm acumulada para onze períodos imediatamente anteriores, de 1, 2, 3, 4, 5, 6 a 10, 11 a 15, 16 a 30, 31 a 60, 61 a 90, e 91 a 120 dias. Os dados pontuais das estações de superfície são interpolados para toda área, ou são determinados a partir das estimativas do hidroestimador.

2. Calcula os “Fatores de Precipitação”, (FP), com valor de 0 a 1, para

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

cada um dos onze períodos, por meio de uma função exponencial empírica da precipitação em milímetros de chuva para cada um deles. As equações são respectivamente:

- $FP1 = exp(-0.14 * precip)$;
- $FP2 = exp(-0.07 * precip)$;
- $FP3 = exp(-0.04 * precip)$;
- $FP4 = exp(-0.03 * precip)$;
- $FP5 = exp(-0.02 * precip)$;
- $FP6a10 = exp(-0.01 * precip)$;
- $FP11a15 = exp(-0.008 * precip)$;
- $FP16a30 = exp(-0.004 * precip)$;
- $FP31a60 = exp(-0.002 * precip)$;
- $FP61a90 = exp(-0.001 * precip)$, e;
- $FP91a120 = exp(-0.0007 * precip)$

3. Calcula os "Dias de Secura", (S), pela multiplicação dos FP conforme a equação:

$$S = 105 \quad FP1 \quad FP2.....FP91a120$$

4. Determina o risco de fogo "básico"p/ cada um dos cinco tipos de vegetação considerada, conforme equação e valores mostrados na Figura 1:

$$RB^{n=1,5} = 0.9\left[1_{+seno}\left(A_{n=1,5} * PSE\right)\right]/2$$

CLASSE DE VEGETAÇÃO	1	2	3	4	5
Tipo Vegetação	Ombrofila Densa	Ombrofila Aberta	Contato + Campinarana	Estacional + Decídua + Semi-Deciadual	Não Floresta
Constante "A"	1.715	2	2.4	3	4

Figura 1: Determinação do Risco de Fogo Básico

5. Corrige o risco de fogo para a umidade relativa mínima do ar;

6. Corrige o risco de fogo para a temperatura máxima do ar;

7. Gera o Risco Observado, RF;

8. Quando verifica-se que em áreas com RF Mínimo e Baixo ocorre algum foco de queima detectado pelos satélites, altera-se o valor do RF para Alto

Neste trabalho serão utilizadas técnicas de mineração de dados para, dadas as variáveis disponíveis (Vegetação, Suscetibilidade, Precipitação, Nun- DiasSemChuva) verificar o poder de previsão dos modelos gerados pelos algoritmos para classificação em relação ao Risco de Fogo calculado.

4 Dados Coletados

Foram utilizados dados disponíveis no banco de dados do INPE

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

chamado BDQueimadas [11]. Foram utilizados dados de focos de incêndios detectados no estado de Minas Gerais no ano de 2009. A tabela possui 356 registros e 17 campos. Cada registro representa um foco detectado de incêndio a partir de imagens de satélite.

5 Pré-Processamento dos Dados

Os 17 campos da tabela inicial foram reduzidos à 5 campos: Vegetação, Suscetibilidade, Precipitação, NunDiasSemChuva e Risco. Os campos Precipitação, NunDiasSemChuva e Risco foram discretizados para aplicação dos algoritmos de mineração de dados. Os campos Nr, Lat, Long, LatGMS, LongGMS, Data, Hora, Satelite, Municipio, Estado, Pais e Persistencia foram apagados.

6 Algoritmos de Mineração de Dados Utiliza- dos

Com os 5 campos restantes da tabela, o interesse era verificar se os algoritmos de classificação poderiam gerar modelos que classificassem os registros adequadamente em relação ao risco de fogo. Os algoritmos utilizados no trabalho foram *C&R Tree*, CHAID, QUEST, C5, rede neural e SVM.

7 Resultados Encontrados

Dentre os algoritmos de classificação, o que obteve melhor resultado foi o SVM com 88,39% de acertos, seguido da rede neural com 86,97% de acertos, como pode ser visto na Tabela 1.

Tabela 1: Resultados dos algoritmos de classificação

(%)			*CR Tree*	C5	rede neural	
	82,44	82,44	85,27		86,97	
	17,56	17,56	14,73		13,03	

7.1 Conjunto de Regras

Dentre os algoritmos que geram um conjunto de regras, o C5 foi o que obteve melhores resultados na classificação. O conjunto de regras gerados está reproduzido na Figura 2 e a árvore de decisão está na Figura 3

Interessante notar que o cálculo do Risco de Fogo tem como princípio é o de que quanto mais dias sem chuva, maior o risco de queima da vegetação, entretanto no algoritmo de classificação C5, a variável mais importante foi a precipitação, como pode ser observado na Figura 4.

8 Conclusão

O cálculo do risco de fogo é um processo de várias etapas e bastante complexo no que diz respeito aos dados necessários para seu funcionamento. Utilizando técnicas de mineração de dados foi possível atingir uma taxa de acerto de quase 90% com SVM utilizando como base os dados de focos de incêndios de 2009. Como o próprio cálculo do risco de fogo não é preciso no que diz respeito a previsão de focos de incêndios, utilizar os modelos gerados pode ser uma alternativa viável quando não for possível conseguir todos os dados necessários para o cálculo do Risco de Fogo.

```
Rules for Alto - contains 3
rule(s) Rule 1 for    Alto
```

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

```
 if Suscetibilidade in [ “ALTA” “MEDIA” ]
 and Precipitacao in [ “Entre 50 e 100” “Menor que 50” ]
 and NunDiasSemChuva in [ “Entre 10 e 20” “Entre 5 e 10” “Maior que 30” ]
 then Alto
Rule 2 for  Alto
 if Suscetibilidade = AGRICULTURA
 and Precipitacao in [ “Entre 50 e 100” “Menor que
 50” ]  and  NunDiasSemChuva  in  [ “Maior  que  30”
 “Menor que 5” ] then Alto
Rule 3 for  Alto
 if Vegetacao in [ “EstacionalDecidual” “EstacionalSemidecidual”
 “OmbrofilaDensa” ]
 and Precipitacao in [ “Entre 50 e 100” “Menor que 50” ]
 then Alto
Rules for Baixo - contains 2
rule(s) Rule 1 for    Baixo
 if Precipitacao in [ “Entre 100 e 150” “Entre 150 e 200” “Maior que 200” ]
 then Baixo
Rule 2 for  Baixo
 if Vegetacao in [ “Contato”
 “NaoFloresta” ] and Suscetibilidade
 in [ “ALTA” “MEDIA” ] and
 NunDiasSemChuva = Menor que 5
 then Baixo
Rules for Medio - contains 1
rule(s) Rule 1 for    Medio
 if Suscetibilidade = AGRICULTURA
 and NunDiasSemChuva in [ “Entre 10 e 20” “Entre 5 e 10” ]
 then Medio
Default: Alto
```

Figura 2: Conjunto de Regras geradas pelo algoritmo C5

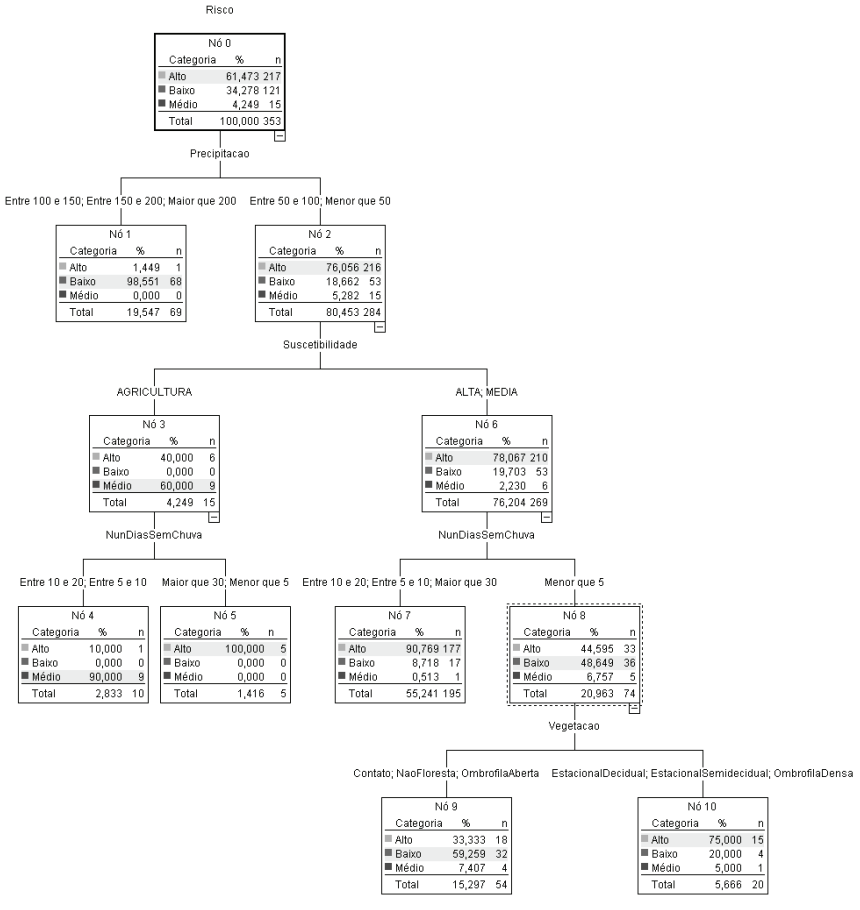

Figura 3: Árvore de decisão resultante do algoritmo C5

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

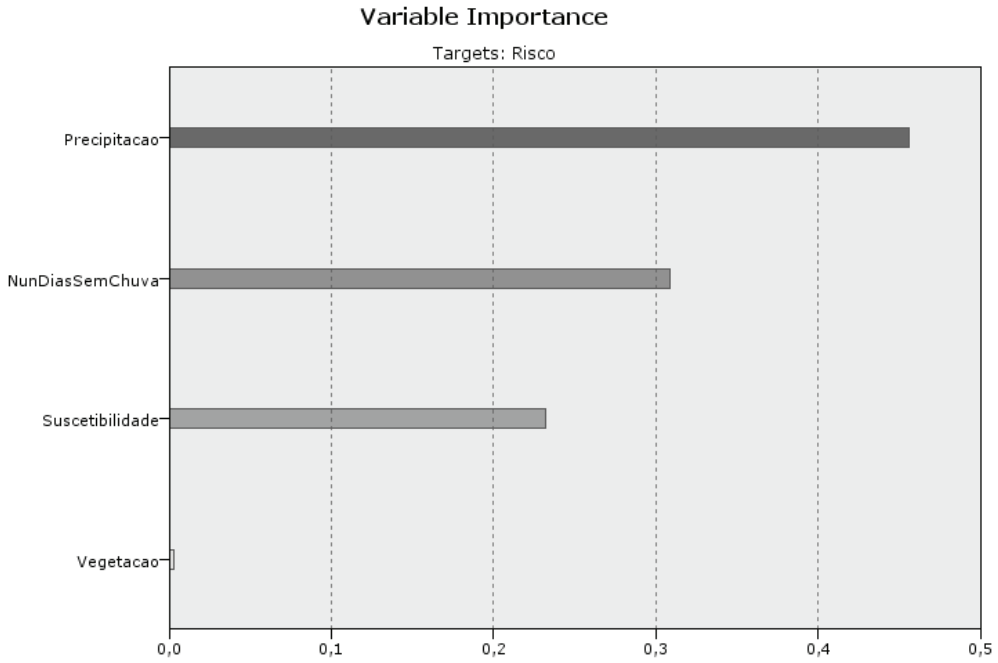

Figura 4: Importância das variáveis no algoritmo C5

Referências

[1] Buttenfield, B., Gahegan, M., Miller, H., and Yuan, M. (2000). Geospatial data mining and knowledge discovery. In *UCGIS White Paper on Emergent Research Themes*.

[2] Cheng, T. and Wang, J. (2008). Integrated spatio-temporal data mining for forest fire prediction. *Transactions in GIS*, 12(5):591–611.

[3] Compieta, P., Martino, S. D., Bertolotto, M., Ferrucci, F., and Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *J. Vis. Lang. Comput.*, 18(3):255–279.

[4] Devillers, R., Bédard, Y., Jeansoulin, R., and Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical In- formation Science*, 21(3):261–282.

[5] Foody, G. M. (2003). Uncertainty, knowledge discovery and data mining in GIS. *Progress in Physical Geography*, 27(1):113 – 121.

[6] Gidófalvi, G. and Pedersen, T. (2009). Mining long, sharable patterns in trajectories of moving objects. *GeoInformatica*, 13(1):27–55.

[7] Goodall, J. L. and Maidment, D. R. (2009). A spatiotemporal data mo- del for river basin-scale hydrologic systems. *International Journal of Ge- ographical Information Science*, 23(2):233–247.

[8] Guo, D. (2009). Multivariate spatial clustering and geovisualization. In Miller, H. J. and Han, J., editors, *Geographic Data Mining and Knowledge Discovery*, chapter 12, pages 325–346. Taylor & Francis, 2 edition.

[9] Han, J. and Kamber, M. (2005). *Data Mining: Concepts and Techniques*. Morgan Kauffman, 2 edition.

[10] Huang, B., Zhang, L., and Wu, B. (2009). Spatiotemporal analysis of rural–urban land conversion. *International Journal of Geographical Infor- mation Science*, 23(3):379–398.

[11] INPE (2010). Bdqueimadas. http://www.cptec.inpe.br/queimadas.

Revista Eletrônica de Iniciação Científica: Gestão Sustentável e Negócios - v. 1, ano 1, p. 29-36, Jan/Jul 2011

[12] Koperski, K., Han, J., and Adhikary, J. (1998). Mining knowledge in geographical data. *Communications of the ACM*, 26.

[13] Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: A partition- and -group framework. In *Intl. Conf. Management of Data (SIGMOD'07)*, pages 593—-604.

[14] Miller, H. J. and Han, J., editors (2009). *Geographic Data Mining and Geographic Discovery*. Taylor & Francis, London, 2 edition.

[15] Ng, R. (2001). *Geographic Data Mining and Knowledge Discovery*, chap- ter Detecting outliers from large datasets, pages 218–235. Taylor and Francis.

[16] Pultar, E., Raubal, M., Cova, T. J., and Goodchild, M. F. (2009). Dynamic GIS case studies: Wildfire evacuation and volunteered geographic information. *Transactions in GIS*, 13(s1):85–104.

[17] Roddick, J. F. and Lees, B. G. (2009). Spatial-temporal data mining paradigms and methodologies. In Miller, H. J. and Han, J., editors, *Ge- ographic Data Mining and Knowledge Discovery*, chapter 2, pages 27–44. Taylor & Francis.

[18] Setzer, A. W. and Sismanoglu, R. A. (2007). Risco de fogo – resumo do método de cálculo. Technical report, DSA / CPTEC / INPE.

[19] Shekhar, S., Lu, C. T., and Zhang, P. (2003). A unified approach to detecting spatial outliers. *Geoinformatica*, 7:139–166.

[20] Shekhar, S., Vatsavai, R. R., and Chawla, S. (2009). Spatial classification and prediction models for geospatial data mining. In Miller, H. J. and Han, J., editors, *Geographic Data Mining and Knowledge Discovery*, chapter 6, pages 117–148. Taylor & Francis.