VII Simpósio Brasileiro de Geoinformática, Campos do Jordão, Brasil, 20-23 novembro 2005, INPE, p. 223-235.

Um Ambiente para Monitoramento da Morte Súbita dos Citrus

Elias Teixeira Krainski[1], Paulo Justiniano Ribeiro Jr[1], Pedro Ribeiro Andrade Neto[1], Renato Beozzo Bassanezi[2]

[1]Laboratório de Estatística e Geoinformação (LEG)
Departamento de Estatística
Universidade Federal do Paraná (UFPR)
Caixa Postal 19.081 CEP 81.531-990 Curitiba – PR – Brasil

[2]Fundo de Defesa da Citricultura (Fundecitrus) – Araraquara, SP – Brasil

{elias, paulojus, pedro}@est.ufpr.br, rbbassanezi@fundecitrus.com.br

Abstract. *This article describes the implementation and applications of computational-statistical surveillance system for the Citrus Sudden Death Disease. The data is stored in a spatio-temporal TerraLib database and statistical analysis are performed using functions written as a add-on package for the R language called Rcitrus which implements some specialized statistical methods and also interfaces with other packages such as geoR, geoRglm and splancs. The interaction between the statistical environment and the database is provided by the package aRT.*

Resumo. *Este artigo descreve a implementação e aplicações de um sistema estatístico-computacional para monitoramento da Morte Súbita dos Citrus. Os dados são armazenados em um banco de dados espaço-temporal em formato TerraLib e as análises estatísticas são executadas usando funções escritas na forma de um pacote da linguagem R chamado Rcitrus que implementa métodos especializados e também provê interfaces para funcionalidades disponíveis em outros pacotes como geoR, geoRglm, e splancs. A interação do ambiente estatístico e o banco de dados ocorre utilizando-se o pacote aRT.*

1. Introdução

Segundo[BASSANEZI et al. 2003], a Morte Súbita dos Citrus (MSC) é uma nova doença dos Citrus que provoca rápido definhamento e morte de variedades de laranjas enxertadas em limoeiro *Cravo*. O primeiro registro oficial da doença foi realizado em fevereiro de 2001 no município de Comendador Gomes, estado de Minas Gerais. Em 2002 a MSC atingiu o estado de São Paulo. Este estado responde por 80% da produção citricola nacional e 85% de seus pomares apresentando laranjeiras doces enxertadas sobre limoeiro *Cravo*. Portanto, há uma elevada vulnerabilidade da cultura à ocorrência de novas epidemias e particularmente para a MSC.

O Citrus Sudden Death Vírus (CSDV), novo vírus da família *Tymoviridae*, tem sido associado à MSC em pesquisas feitas pela empresa de biotecnologia Alellyx. Em abril de 2004, a mesma empresa anunciou a descoberta da presença do

patógeno da MSC. O CSDV foi encontrado em três insetos, dois deles sendo capazes de transmitir o vírus para as plantas, [Alellyx 2004].

Vários trabalhos têm sido conduzidos buscando a compreensão dos mecanismos e dinâmica da doença, incluindo o estudo de padrões espaciais. Tais trabalhos abrangem a coleta e análise de dados epidemiológicos provenientes de avaliações feitas em diferentes momentos em talhões de plantas de Citrus. Os talhões são localizados diferentes propriedades, abrangendo municípios do sul de Minas Gerais e norte-noroeste de São Paulo. Os dados começaram a ser coletados em 2002 e são atualizados periodicamente em levantamentos feitos pelo Fundo de Defesa da Citricultura (Fundecitrus). Nos talhões analisados até o momento o número de plantas varia de 700 a 5000 plantas e os levantamentos vão de 1 a 26 momentos no tempo.

A análise estatística desse volume de dados requer o desenvolvimento e/ou adaptação de metodologias para o estudo de padrões espaciais de doenças. Além disso, considerando o volume de dados e o fato de que levantamentos para o acompanhamento da doença são feitos regularmente ao longo do tempo, são necessários procedimentos de análise e geração de resultados e relatórios de forma automática. A estrutura dos dados e as análises estatísticas requerem que o banco de dados seja construído segundo um modelo espaço-temporal. Soma-se ainda a necessidade de ferramentas automáticas de leitura dos dados e validação com detecção e correção de inconsistências.

Este trabalho descreve a implementação usando a linguagem R [R Development Core Team 2005] de um ambiente para análise estatística de dados da MSC. Na Seção 2. descrevemos alguns dos métodos estatísticos especialistas para análise de dados de doenças de plantas, implementados no pacote Rcitrus[1]. Na Seção 3., apresentamos o modelo do banco de dados TerraLib [2] que se mostra particularmente adequado para o tratamento de dados com a estrutura da MSC. Na Seção 3. apresentamos a interação entre o ambiente R e o banco de dados, feita pela (API) R-TerraLib, desenvolvida para ser utilizada em forma de um pacote do R, chamado aRT [3]. A Seção 4. apresente conclusões e discussões.

2. Rcitrus

O Rcitrus é um pacote desenvolvido em R que implementa e adapta metodologias de análise estatística para dados de incidência de doenças em plantas. Até o momento, foram implementadas funções para manipulação e validação dos dados, alguns métodos estatísticos de análise para dados de doenças em plantas e adequação de métodos já implementados em outros pacotes do R. Estes pacotes incluem o splancs [Rowlingson et al. 2005] para análise de processos pontuais, os pacotes geoR [RIBEIRO JR. and DIGGLE 2001] e geoRglm [Christensen and Jr 2002] para análise geoestatística e o pacote survival [original by Terry Therneau and ported by Thomas Lumley 2005] para análise de sobrevivência.

O Rcitrus implementa funções para **ler** dados de planilhas para o ambiente R

[1] http://www.est.ufpr.br/Rcitrus
[2] http://www.terralib.org
[3] http://www.est.ufpr.br/aRT

e **escrever** estes dados para formato texto, **manipular** dados de doenças de plantas, **validar** dados considerando as características da MSC, **analisar** o padrão espacial da incidência de doenças dentro de talhões utilizando diferentes técnicas e **simular** dados utilizando diferentes modelos espaciais. Para a maioria das saídas das funções implementadas foram implementados métodos de vizualização das saídas e gráficos. O pacote pode ser instalado como usual para pacotes do R e carregado de forma usual com o comando do R require(Rcitrus).

```
[1] TRUE
[1] TRUE
```

2.1. Manipulação dos dados

Os dados recebidos do Fundecitrus estão disponíveis no formato de planilhas, onde cada célula representa uma planta. Foram implementadas funções para ler dados de planilhas em arquivos texto e conversores para formatos de representação espaço-temporal com classes definidas no Rcitrus. As classes implementadas facilitam a validação, a extração de estatísticas descritivas, gráficos e analise estatística dos dados. Também foram implementados conversores para classes definidas em outros pacotes, o que facilita a aplicação de técnicas estatísticas disponíveis nestes outros pacotes tais como métodos de análise espacial e de análise de sobrevivência. Os dados são convertidos em classes implementadas pelo pacote sp e exportados através do aRT para o banco de dados TerraLib, Os detalhes da interação como banco de dados estão na Seção 3..

```
> d.arr1 <- read.citrus("vv303.csv", find.form = "array", nrow = 20,
+     row.id = 1, n.att = 14, sep = ";", dec = ",")
> table(d.arr1)
d.arr1
    0     1     2     3     F     G     O     R
11090  6897  1363  4292    94     3     8   253
> d.geo1 <- citrus.conv(d.arr1, find.form = "geodata")
x = 7.5 y = 4
```

2.2. Validação dos dados

Os dados de MSC são codificados segundo o estágio do desenvolvimento da doença na planta: 0 (zero) para plantas sadias, 1 (um) para plantas em estágio inicial da doença, 2 (dois) para plantas em estágio avançado e 3 (três) para plantas mortas. Os dados brutos podem apresentar erros de digitação ou inconsistências. Diante disso, forem implementados alguns procedimentos de validação dos dados.

```
> d.geo2 <- change.code(d.geo1, ori = c("O", "G"), mod = c(0,
+     0))
> d.geo3 <- select.code(d.geo1, unselect.cods = c("F", "R"))
```

Visando as especificidades da análise de dados da MSC, também foi implementado um procedimento de validação temporal de inconsistências, respeitando a premissa biológica de que estágio da doença nunca regride.

```
> d.geo4 <- valid.time.citrus(d.geo3)
9  inconsistences in  25 evaluations of 945 plants.
```

2.3. Análise por *quadrats*

A análise por *quadrats* é um procedimento simples utilizado para caracterizar o padrão espacial da doença como agregado, regular ou aleatório. Para esta análise é feita uma subdivisão do talhão em N regiões denominadas *quadrats*. O procedimento de seleção dos *quadrats*, a forma e o tamanho são discutíveis e portanto a implementação deixa livre a escolha para o usuário. Estes podem ser tomados de forma "fixa" o que consiste em uma subdivisão regular do talhão ou de forma "aleatória" a posição do *quadrat* é tomada por sorteio no talhão. A função implementada contempla as duas possibilidades.

```
> args(disp.quadrats)
function (data, dx, dy = dx, counts.return = FALSE, by.evaluations = TRUE,
    suces = 1, unsuces = 0, model = c("binomial", "Poisson",
        "beta-binomial"), alpha = 0.05, random = FALSE, N = NULL,
    p.quadrats = 1, p.quadrats.random = FALSE, complet = TRUE,
    evaluation = "all", digits = 5, verbose = FALSE, bb.args = list(ini.p = NULL,
        ini.theta = NULL, usage = c("fitdistr", "mle")), ...)
NULL
```

A análise do padrão espacial é feita a partir do número de plantas doentes dentre as plantas de cada quadrat. Sendo aleatório o padrão da doença, espera-se que a distribuição binomial ajuste-se bem aos dados. Para quadrats com grande número de plantas pode-se utilizar a distribuição de Poisson. Para análise do padrão espacial, o teste do ajuste pode ser feito pelo índice de dispersão D, que é a razão entre a variância observada nos dados e a variância teórica, segundo a distribuição assumida. A hipótese de aleatoriedade espacial é testada considerando que, sob essa hipótese, $D(N-1)$ tem distribuição $\chi^2_{(N-1)}$.

Uma família mais flexível para tais ajustes é dada pela distribuição Beta-Binomial que permite que o parâmetro de incidência da doença não seja assumido constante como nos casos anteriores, mas varie na região. O ajuste da distribuição beta-binomial também foi implementado no Rcitrus. Note-se que neste caso é utilizado para estimação dos parâmetros um procedimento iterativo de maximização numérica da função de verossimilhança. Os procedimentos que podem ser utilizados são algoritmos padrão de minimização disponíveis em R, tais como: "Nelder-Mead", "BFGS", "CG", "L-BFGS-B" ou "SANN". Inferências sobre os parâmetros da distribuição beta-binomial é feita aproximando-se a matriz de informação de Fisher pelo hessiano obtido numericamente. A hipótese de aleatoriedade espacial é testada pelo teste da nulidade do parâmetro de agregação.

```
> disp.quadrats(d.geo4, dx = 3, dy = 5, by = F, eval = 1:5,
+     mod = "bet", usag = "mle")
$"3x5"
     N        n    prob   theta p.value      conc
Av1 54 14.64815 0.01898 0.06183 0.04974  Agregado
Av2 54 14.64815 0.02148 0.04425 0.10032 Aleatorio
Av3 54 14.33333 0.05109 0.04932 0.06993 Aleatorio
Av4 54 14.29630 0.06984 0.06303 0.03488  Agregado
Av5 54 14.29630 0.07983 0.04061 0.13402 Aleatorio
```

2.4. Análise de processos pontuais

A análise de processos pontuais é feita a partir das coordenadas das plantas doentes. Foram implementados métodos para análise utilizando um treste Monte Carlo para distância mínima e para o número de vizinhos doentes. A distância mínima é a distância entre uma planta doente e a planta doente mais próxima. É razoável assumir que quanto menor a média das distâncias mínima das plantas doentes, dada uma incidência, mais agregado é o padrão espacial. O teste de Monte Carlo nesse caso consiste em comparar a distância mínima média observada com distribuição dessa estatística calculada para dados simulados sob a hipótese nula, com a mesma incidência e no grid definido pelas posições das plantas no talhão. Também foi implementado um teste de Monte Carlo para análise do número médio de vizinhos doentes dentro de um raio.

```
> summary(mmd <- mmdist.test(d.geo4))

Loading required package: splancs

Spatial Point Pattern Analysis Code in S-Plus

 Version 2 - Spatial and Space-Time analysis
test avaluation: 1
Monte Carlo test results!
Obs.vals: 14.37121
Rand.vals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.36   20.91   23.03   23.05   25.31   31.43
P-value: 0.01

> summary(nei <- neigh.test(d.geo4))

Monte Carlo test results!
Obs.vals: 1.882353
Rand.vals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5882  0.9412  1.1760  1.2370  1.4120  2.4710
P-value: 0.07

> par(mfrow = c(2, 2))
> hist(mmd, main = "Distancia Minima Media")
> plot(mmd, main = "Distancia Minima Media")
> hist(nei, main = "Numero de Vizinhos Proximos")
> plot(nei, main = "Numero de Vizinhos Proximos")
```

Conversores para disponibilizar os dados em formatos utilizados pelo splancs foram implementados para facilitar a análise por métodos de processos pontuais deiponíveis neste pacote. Para a geração de mapas ou filmes, estão implementados a suavização por kernel em 2 e em 3 dimensões e para análise do padrão espacial, o envelope simulado para a função K de Ripley, dentre outros métodos.

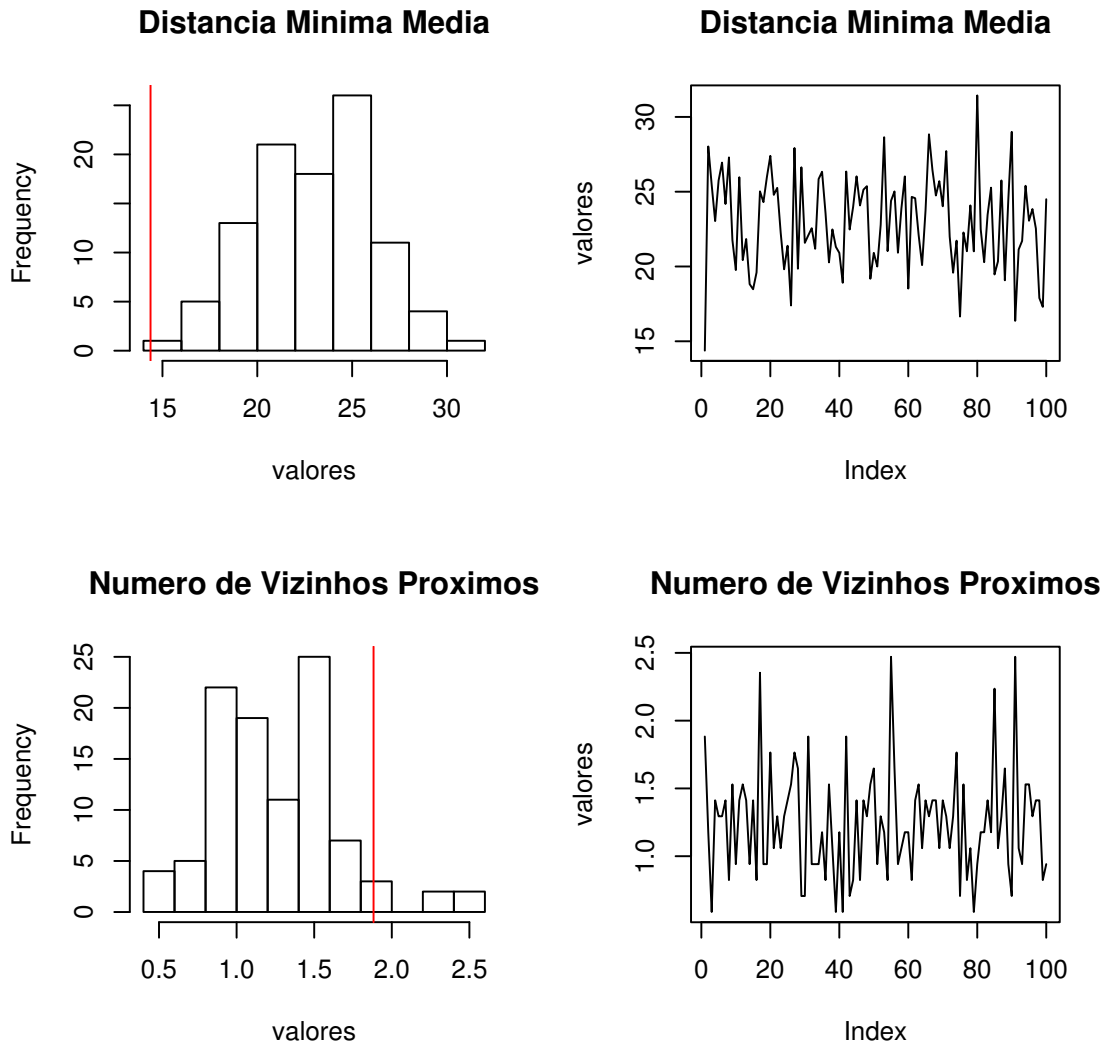

Figure 1. Vizualização gráfica dos testes de Monte Carlo.

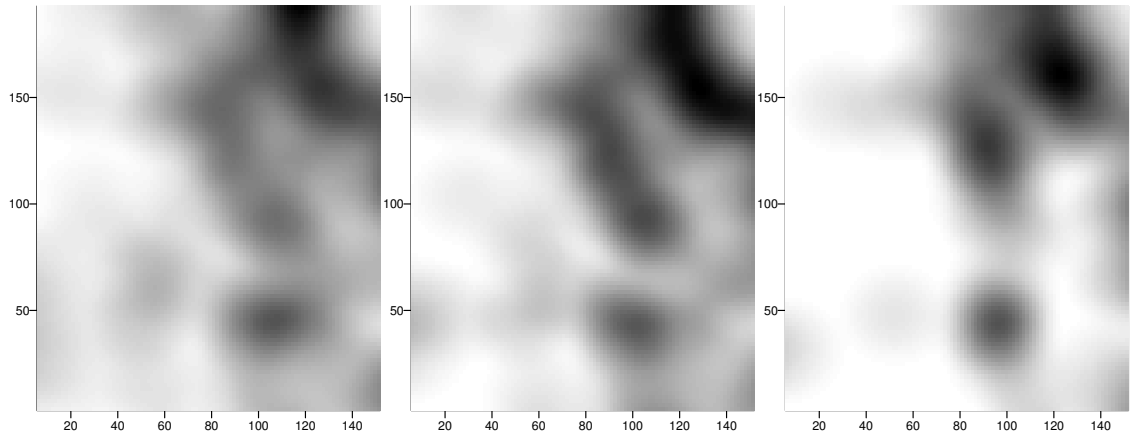

Figure 2. Kernel com escala de cores individual para cada avaliação.

```
> ker2d <- kernel2d.citrus(d.geo4, h0 = 20, eval = 19:21, death = 1:3)
> p3d <- as.Tpoints(d.geo4, death = 1:3, eval = 19:21, ref.time = "01/01/2001",
+     form.date = "dd/mm/yyyy")
> ker3d <- kernel3d(p3d[, 1:2], p3d[, 3], ker2d[[1]]$x, ker2d[[1]]$y,
+     seq(510, 750, by = 30), 20, 100)

> par(mfrow = c(1, 3), mar = c(2, 2, 1, 0.1), mgp = c(1, 0.5,
+     0), las = 1)
> image(ker2d, zlim = "individual", col = gray(seq(0, 1, 0.01)))

> par(mfrow = c(3, 3), mar = c(2, 2, 1, 0.1), mgp = c(1, 0.5,
+     0), las = 1)
> for (i in 1:9) {
+     image(ker3d$xgr, ker3d$ygr, ker3d$v[, , i], asp = 1,
+         xlab = "", ylab = "", main = 0 + i, col = gray(seq(0,
+             1, 0.01)), xlim = range(ker3d$xgr) + c(0, 20))
+     legend.krige(max(ker3d$xgr) + c(1, 11), range(ker3d$ygr),
+         ker3d$v[, , i], vert = TRUE, col = gray(seq(0, 1,
+             0.01)))
+ }
```

2.5. Modelo autologístico

O modelo autologístico é um modelo de regressão logística aplicado a dados binários tendo o status das plantas vizinhas como covariáveis [GUMPERTZ and RISTAINO 1997]. Esse modelo pode ser utilizado para avaliar se a informação do status das plantas vizinhas influencia a probabilidade da presença da doença numa determinada planta. A existência de correlação espacial pode ser avaliada a partir da inferência sobre parâmetros da (auto)regressão. Diferentes correlações podem ser testadas considerando diferentes estruturas de vizinhança. No contexto de doenças de plantas é conveniente considerar separadamente a vizinhança entre linhas, colunas e diagonais devido ao fato de que usualmente o espaçamento entre linhas e colunas são diferentes. O

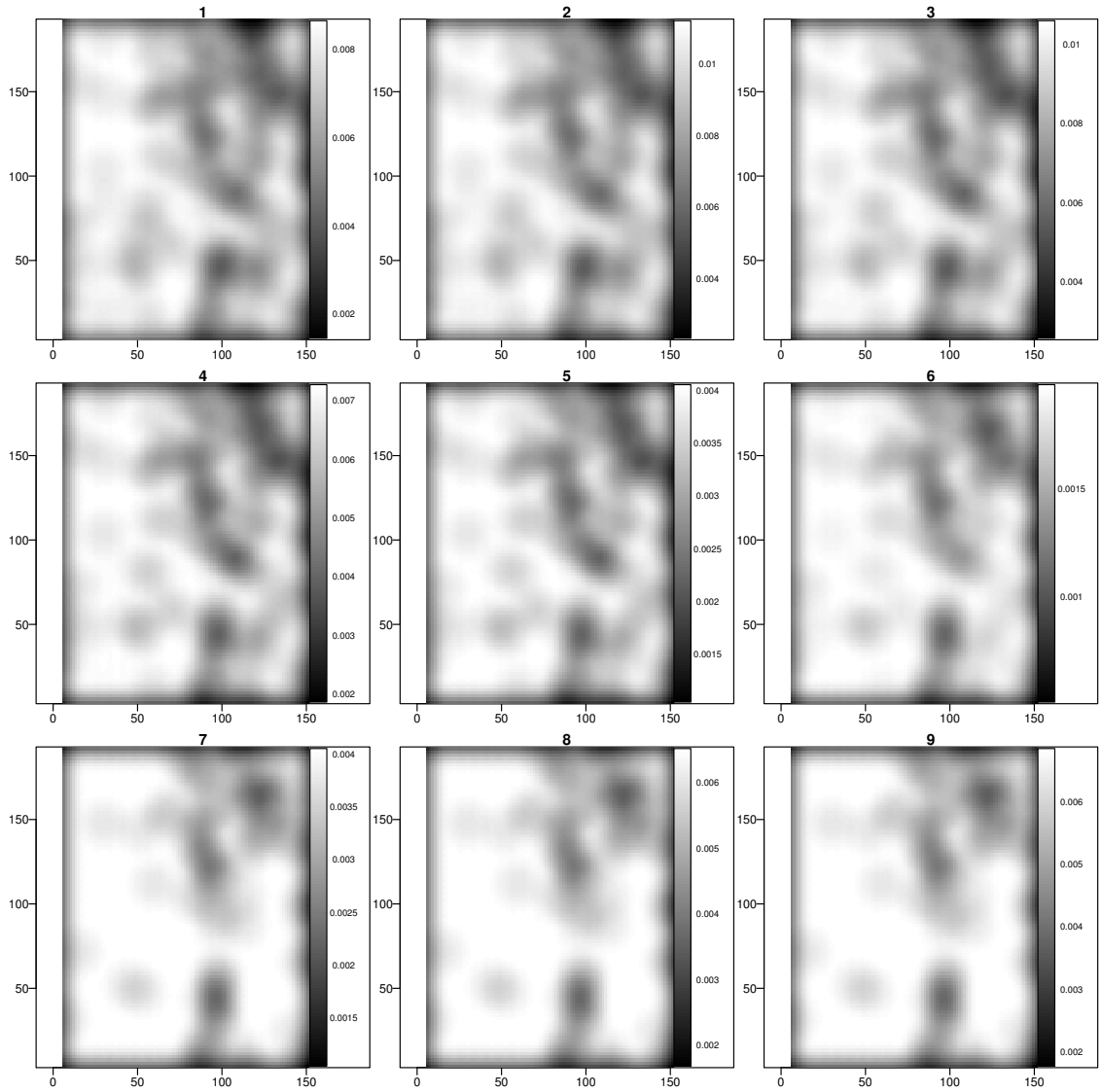

Figure 3. Kernel espaço-temporal

modelo autologístico foi implementado considerando tal estrutura e também termos de interação no modelo de regressão.

Os erros padrão retornados por um modelo de regressão logística usual são inválidos devido ao fato de que cada dado é usado como resposta e também para contruir a covariável de outro ponto. A estimação dos erros padrões dos coeficientes de regressão pode entretanto ser feita utilizando de métodos computacionalmente intensivos de reamostragem. O procedimento implementado consiste em um procedimento de reamostragem paramétrica (bootstrap). Esse procedimento não é trivial devido a necessidade de preservar aspectos da configuração espacial dos dados nas reamostragens. O procedimento implementado de simulação das amostras, foi o algoritmo amostrador de Gibbs, tendo o modelo ajustado aos dados observados com o condicionais completas.

```
> data(bellPepper)
> (aut <- autologistic.citrus(bellPepper, N = 30))

(Intercept)           R           C          d1          d2
 -2.9423866   1.2488850  -0.1889378   0.5628347   1.0212967
sim 1 : ok : -3.438762 1.530194 -0.8336252 1.191749 1.676766
...
sim 30 : ok : -2.845457 0.5823241 0.66048 0.2104695 0.752867
Resultados da Pseudo-Verossimilhanca
Coeficientes:
(Intercept)           R           C          d1          d2
 -2.9423866   1.2488850  -0.1889378   0.5628347   1.0212967
Variancias:
(Intercept)           R           C          d1          d2
 0.08388588  0.07436137  0.13587450  0.11181274  0.07559117
Resultados da reamostragem bootstrap via Amostrador de Gibbs:
Coeficientes:
(Intercept)           R           C          d1          d2
 -3.0268124   1.2744492  -0.9268327   0.5686159   1.0362258
Variancias:
(Intercept)           R           C          d1          d2
  0.1281389   0.4213960   9.2726198   0.3879388   0.3638586

> par(mfrow = c(2, 5), mar = c(3, 3, 3, 1), mgp = c(2, 1, 0))
> plot(aut)
> density.autologistic(aut)
```

2.6. Simulação de dados com padrão espacial

O procedimento de simulação de dados com padrão espacial é importante para se comparar métodos e fazer inferência. A simulação de dados binários com dependência espacial pode ser feita de diferentes formas. Foram implementados 5 métodos: (1) modelo hierárquico [Y|S] onde S é um campo aleatório gaussiano. (2) grampeamento de um campo aleatório gaussiano de forma simples fazendo $Y = 1$ se $S < z$ e $Y = 0$ se $S > z$, onde z é um valor de corte. (3) transformação de um campo aleatório

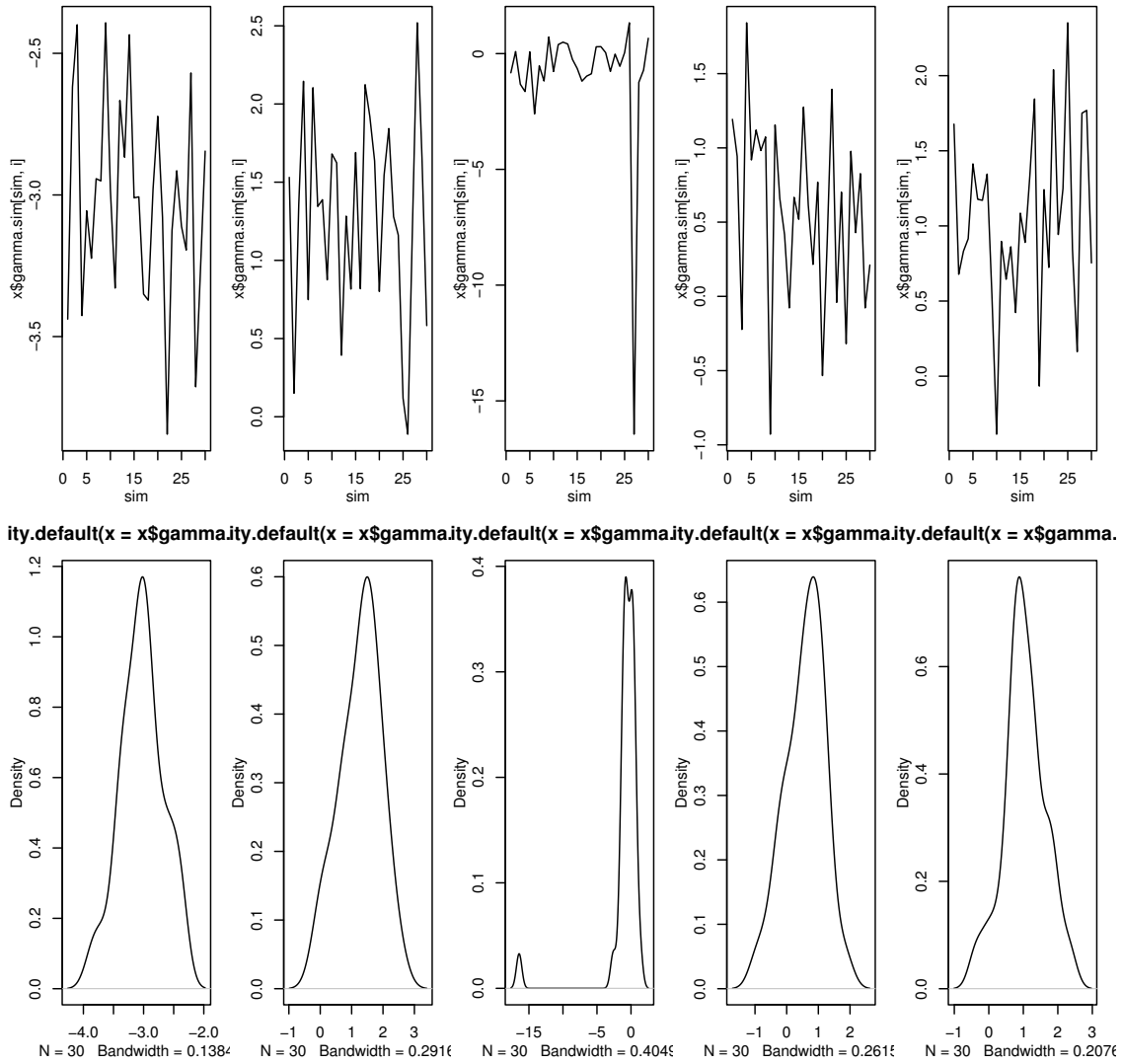

Figure 4. Vizualizacao dos valores obtidos nas simulações.

gaussiano por: $Y = fb(f - 1(S))$, onde fb é o quantil da distribuição bernoulli e f-1 é a densidade acumulada da distribuição Normal univariada com mesma média e variância de S. (4) O modelo de pressão infectiva exponencial. (5) O modelo de pressão infectiva e potencial.

```
> nr <- 20
> nc <- 20
> coords <- as.matrix(expand.grid(1:nr, 1:nc))
> set.seed(123)
> mpd <- sim.citrus(coords, 0.1, "pdist", a1 = 0, a2 = 2)
> set.seed(123)
> med <- sim.citrus(coords, 0.1, "edist", a1 = 0, a2 = 2)
> set.seed(123)
> mHgrf <- sim.citrus(coords, 0.1, "hGRF", cov.pars = c(1,
+     20))
> set.seed(123)
> mCgrf <- sim.citrus(coords, 0.1, "clipGRF", cov.pars = c(1,
+     20))
> set.seed(123)
> mTgrf <- sim.citrus(coords, 0.1, "transGRF", cov.pars = c(1,
+     20))

> par(mfrow = c(2, 3), mar = c(3, 3, 3, 1), mgp = c(2, 1, 0))
> points(mpd)
> points(med)
> points(mHgrf)
> points(mCgrf)
> points(mTgrf)
```

3. O pacote aRT

O aRT é um pacote R que possibilita o acesso a uma biblioteca de geoprocessamento chamada Terralib. A TerraLib define um modelo de banco de dados espaço-temporal utilizando um Sistema Gerenciador de Banco de Dados. Na TerraLib estão implementados vários procedimentos de consulta espaço-temporal o que é muito adequado para dados como os da MSC. Com o aRT, esses procedimentos podem ser acessados e incorporados na análise estatística, além de ser permitida a leitura e escrita de dados no banco de dados.

Os dados são colocados no banco de dados espaciais de forma a permitir que usuários leigos possam acessá-lo e visualizar o estado atual da doença de acordo com as análises realizadas, utilizando para algum visualizador tal como o programa TerraView. As diferentes classes de tabelas implementadas pelo TerraLib, permitem que não só os dados sejam armazenados em forma de geometria de pontos e de talhões com seus atributos, mas também mídias assossiadas.

As análises estatísticas podem gerar resultados ao nível do talhão, que não tenham necessidade de possuir atributo espacial ou temporal. Esse resultado pode ser armazenado, disponibilizado para consulta e o seu endereço guardado no banco

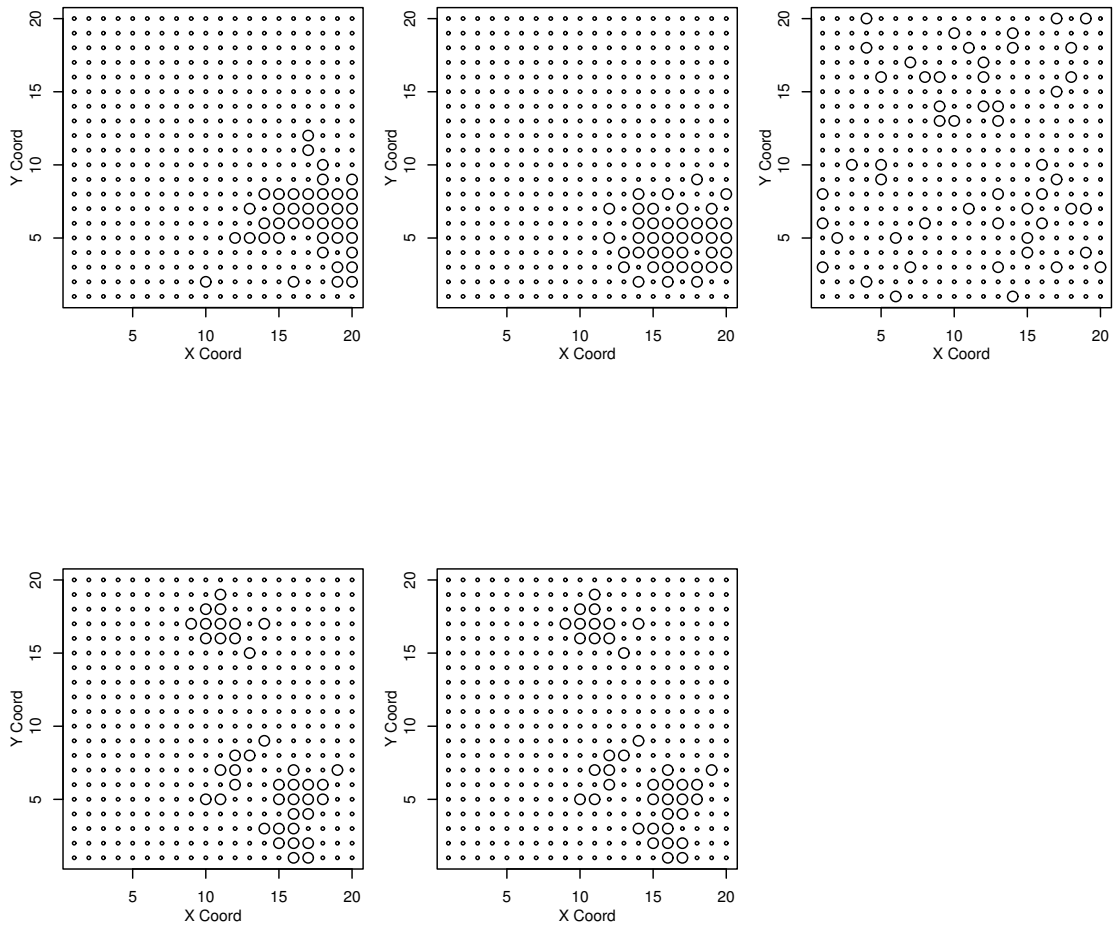

Figure 5. Vizualização de dados simulados.

de dados como tabela de mídia. Com a tabela de mídia, o pesquisador poderá acessar os resultados da análise estatística, quando estiver vizualizando o mapa dos dados. Dessa forma, um usuário do TerraView pode simplesmente clicando no mapa em um talhão vizualizado ter acesso ao relatório da análise.

Um exemplo dessa aplicação pode ser vizualizada no link `http://www.est.ufpr.br/Rcitrus/relatorios/VaVer303`.

4. Conclusões

O ambiente apresentado é capaz de manipular os dados de forma eficiente e de disponibilizar análises estatísticas modernas para aplicação automática em grandes quantidades de dados.

O ambiente torna possível uma integração dinâmica entre o trabalho de pesquisadores e diferentes analistas.

A partir dos recursos computacionais disponíveis, a análise estatística dos dados da MSC armazenados no banco de dados pode ser automatizada e feita integralmente em ambiente R. Os resultados das análises podem ser guardadas no mesmo banco de dados e acessados por quem tiver permissão, e em particular epidemiologistas de plantas que usam tais resultados para compreensão dos mecanismos da doença e definição de estra'egias de controle. Para a análise de dados novos inseridos no banco, basta que o "script" em R seja executado novamente o que pode ser automatizado.

References

Alellyx (2004). Alellyx identifica vetor da morte súbita dos citros. *Alellyx*. URL: http://www.gravena.com.br/dicas_alellyx_vetor_MSC.htm.

BASSANEZI, R. B., FERNANDES, N. G., and YAMMAMOTO, P. T. (2003). Morte súbita do citros. Technical report, Fundecitrus.

Christensen, O. and Jr, P. R. (2002). georglm - a package for generalised linear spatial models. *R-NEWS*, 2(2):26–28. ISSN 1609-3631.

GUMPERTZ, M. L. ; GRAHAM, J. M. and RISTAINO, J. B. (1997). Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics*.

original by Terry Therneau, S. and ported by Thomas Lumley (2005). *survival: Survival analysis, including penalised likelihood.* R package version 2.19.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RIBEIRO JR., P. and DIGGLE, P. (2001). geoR: a package from geostatistical analysis. *R-NEWS*, 1(2):15–18.

Rowlingson, B., Diggle, P., adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen (2005). *splancs: Spatial and Space-Time Point Pattern Analysis*. R package version 2.01-16.