Trabalhando com dados de Citrus em R

Elias T. Krainski & Paulo J. Ribeiro Jr.

Última Atualização: 3 de agosto de 2006

Os dados de *Citrus* que motivam este trabalho são tipicamente armazenados em planilhas, onde cada linha corresponde às linhas de plantas nos talhões e cada coluna corresponde às plantas nas linhas. Nessa forma, os dados são chamados de mapas de doença.

Algumas doenças são acompanhadas ao longo do tempo e com isso vários mapas são obtidos para um mesmo talhão. Isso faz com que os dados tenham várias dimensões, que devem ser consideradas na análise.

Neste trabalho, mostramos como trabalhar com esse tipo de dados em R. Isto é feito utilizando o **Rcitrus**, pacote desenvolvido para análise de dados de doenças de plantas.

1 Importado dados de planilhas

Em R (**?**), é possível ler um arquivo do Excel diretamente, utilizando o pacote **gdata** (**?**) ou **RODBC** (**?**). Também é comum salvar o arquivo como texto separado por vírgulas, com extensão `.csv`. O arquivo `.csv` pode ser lido sem pacote adicional, utilizando a função `read.csv()` ou `read.csv2()`.

1.1 Dados com uma única avaliação

Um exemplo de dados com uma única avaliação, são dados de *Pinta Preta dos Citrus* (PPC) em um talhão no município de Itajobi (MG). Os dados foram dispostos em planilha e salvos em arquivo Excel. Uma cópia deste arquivo está disponível em `http://www.est.ufpr.br/~elias/Itajobi.xls`

A função `read.csv2()` será usada para ler o arquivo `Itajobi.csv`, pois neste arquivo as colunas são separadas por ";" e os números decimais usam "." como separador decimal. Esse arquivo não possui cabeçalho com nome das colunas, então usamos o argumento `header=FALSE`.

Lendo o arquivo:

```
> ita <- read.csv2("Itajobi.csv", header = FALSE)
```

O objeto `ita`, que contem os dados no ambiente R, é da classe `data.frame`.

```
> class(ita)
```

```
[1] "data.frame"
```

Nesse talhão há 62 linhas de plantas e 58 plantas na linha com o maior número de plantas.

```
> dim(ita)
```

```
[1] 62 58
```

Esta é uma representação espacial simplista, em linhas e colunas, das plantas de um talhão.

Pode-se visualizar as 5 primeiras linhas e 25 primeiras plantas nestas linhas, fazendo:

```
> ita[1:5, 1:20]
```

```
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
1  0  0  0  0  0 NA NA NA NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
2  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0  NA  NA  NA  NA
3  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
4  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
5  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0
```

O código 0 (zero) corresponde a plantas sadias e o código 1 (um) corresponde às plantas doentes. No caso de falha ou irregularidade no talhão, as posições sem plantas são identificadas com `NA`. Nesse talhão a primeira linha tem apenas 5 plantas e desta forma ao colocar os dados em formato `data.frame` usa-se o código `NA` a partir da sexta linha.

1.2 Dados com mais de uma avaliação no tempo

Alguns conjuntos de dados contém mais de uma avaliação do estágio da doença, feitas em diferentes momentos do tempo. Assim, é necessário incorporar a estrutura temporal, além da espacial. Nesse caso, os dados de cada avaliação são armazenados um abaixo do outro, na mesma planilha de dados. Um exemplo é o conjunto de dados da incidência de MSC no talhão 303, localizado no município de Comendador Gomes, estado de Minas Gerais. Uma cópia desses dados está disponível em `http://www.est.ufpr.br/~elias/original.xls`. Para simplificar, esse arquivo foi salvo em arquivo texto com extensão `.csv`.

Lendo esse arquivo com a função `read.csv2()` e usando a opção `header=FALSE`.

```
> ori303 <- read.csv2("original303.csv", header = FALSE)
```

Inspencionando o arquivo:

```
> class(ori303)
```

```
[1] "data.frame"
```

```
> dim(ori303)
```

```
[1] 612  49
```

Inspecionando as primeiras linhas e colunas dos dados:

```
> ori303[1:42, 1:5]
```

```
              V1                V2 V3 V4 V5
1       Municipio Comendador Gomes
2     Propriedade       Vale Verde
3    Proprietario
4          Talhao              303
5       Variedade         Valencia
6    PortaEnxerto            Cravo
7   LinhasPlantas               20
```

```
8  PlantasLinhas              48
9    EntreLinhas             7.5
10  DentroLinhas               4
11       Plantio            1991
12      Latitude
13     Longitude
14
15    08/01/2001
16                             1  2  3  4
17             1               0  0  0  0
18             2               0  0  0  0
19             3               0  0  0  0
20             4               0  0  0  0
21             5               0  1  0  0
22             6               0  0  1  0
23             7               0  0  0  0
24             8               0  0  0  0
25             9               0  0  0  0
26            10               0  0  0  0
27            11               R  0  0  0
28            12               0  R  R  F
29            13               0  0  0  0
30            14               0  0  0  0
31            15               R  0  0  0
32            16               0  0  0  0
33            17               0  0  0  0
34            18               0  0  0  0
35            19               0  0  0  0
36            20               0  0  F  0
37
38
39    08/08/2001
40                             1  2  3  4
41             1               0  0  0  1
42             2               0  0  0  0
```

Nesse `data.frame`, observa-se que as 13 primeiras linhas, contém atributos do talhão: Proprietário, Talhão, Variedade, etc. Na 15^a linha observa-se a data da avaliação. Na linhs 16, está a numeração das colunas. Os dados da 1^a avaliação iniciam na linha 17, sendo a primeira coluna, a numeração da linha. Os dados dessa avaliação terminam na linha 36. Na linha 39 está a data da 2^a avaliação e a seguir inicia o dados dessa avaliação dispostos da mesma forma que os dados da primeira avaliação.

2 Classes de representação espaço temporal

No **Rcitrus**, foram implementadas outras classes para possibilitar a representação espaço-temporal. Essas classes facilitam a manipulação e as análises estatísticas. As classes utilizadas são adaptações das classes `array` do pacote básico, `geodata` do pacote **geoR** (**?**) e `Surv` do

pacote **Survival** (**?**). Para particularizar essas classes usadas para dados de plantas, o objeto é identificado como sendo da classe `citrus`, além dessas.

2.1 Classes `citrus` e `array`

É utilizado um `array` com três dimensões, em que a primeira dimensão indica a linha de plantas, a segunda indica as plantas na linha e a terceira indica as avaliações feitas em tempos diferentes. Este formato de dados é particularmente útil na análise por *quadrats*. Os dados de *Pinta Preta* em Itajobi, estão em um `data.frame`. Usando a função `as.citrus` para podemos converte-lo para `array`.

```
> fram.ita <- as.citrus(ita, find.form = "array")
```

Inspencionando o objeto:

```
> fram.ita

Disease plant data in 1 evaluations of
62 rows of plants and 58 plants in each row.

> class(fram.ita)

[1] "citrus" "array"

> dim(fram.ita)

 Rows  Cols Evals
   62    58     1
```

Também podemos usar a função `as.citrus()` para converter os dados do talhão 303, considerando os dados de mais de uma avaliação e os atributos do talhão. Basta usar os demais argumentos da função:

```
> args(as.citrus)

function (data, find.form = c("array", "geodata", "Surv"), nrow = NULL,
    row.id = NULL, col.id = 1, n.att = NULL, col.nam.att = 1,
    col.val.att = 2, pos.date = -2, x = NULL, y = NULL, cod.start = 0:2,
    cod.event = 1:3, order = "dmy", ref.date = "01/01/2001")
NULL
```

Deve-se informar: o número de ruas no talhão (`nrow`), a identificação da primeira rua do talhão (`row.id`), coluna com os códigos identificadores (`col.id`), número de atributos do talhão (`n.att`), coluna onde estão os nomes dos atributos (`col.nam.att`), coluna onde estão os valores dos atributos (`col.val.att`) e posição onde estão as datas de avaliações em relação aos dados (`pos.date`)

```
> o303.array <- as.citrus(ori303, nrow = 20, row.id = 1, n.att = 13)
```

Inspensionando o objeto:

```
> o303.array
```

```
Disease plant data in 25 evaluations of
20 rows of plants and 48 plants in each row.
```

```
> dim(o303.array)
```

```
 Rows  Cols Evals
   20    48    25
```

```
> names(attributes(o303.array))
```

```
 [1] "dim"           "Municipio"     "Propriedade"   "Proprietario"
 [5] "Talhao"        "Variedade"     "PortaEnxerto"  "LinhasPlantas"
 [9] "PlantasLinhas" "EntreLinhas"   "DentroLinhas"  "Plantio"
[13] "Latitude"      "Longitude"     "dimnames"      "class"
```

```
> dimnames(o303.array)
```

```
$Rows
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
[15] "15" "16" "17" "18" "19" "20"

$Cols
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
[15] "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28"
[29] "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42"
[43] "43" "44" "45" "46" "47" "48"

$Evals
 [1] "08/01/2001" "08/08/2001" "08/16/2001" "08/22/2001" "08/29/2001"
 [6] "09/06/2001" "09/12/2001" "09/25/2001" "10/06/2001" "10/10/2001"
[11] "11/07/2001" "12/06/2001" "01/08/2002" "02/12/2002" "03/15/2002"
[16] "04/05/2002" "04/24/2002" "05/08/2002" "06/03/2002" "08/05/2002"
[21] "01/10/2003" "02/10/2003" "03/06/2003" "04/14/2003" "05/09/2003"
```

Pode-se ver facilmente a evolução do *status* de uma planta ao longo das avaliações. Por exemplo, a planta 3 da linha 10:

```
> o303.array[3, 10, ]
```

```
08/01/2001 08/08/2001 08/16/2001 08/22/2001 08/29/2001 09/06/2001
       "0"        "0"        "0"        "0"        "0"        "0"
09/12/2001 09/25/2001 10/06/2001 10/10/2001 11/07/2001 12/06/2001
       "0"        "0"        "0"        "0"        "0"        "0"
01/08/2002 02/12/2002 03/15/2002 04/05/2002 04/24/2002 05/08/2002
       "0"        "0"        "0"        "0"        "0"        "1"
06/03/2002 08/05/2002 01/10/2003 02/10/2003 03/06/2003 04/14/2003
       "1"        "1"        "3"        "3"        "3"        "3"
05/09/2003
       "3"
attr(,"Municipio")
[1] "Comendador Gomes"
```

```
attr(,"Propriedade")
[1] "Vale Verde"
attr(,"Proprietario")
[1] ""
attr(,"Talhao")
[1] "303"
attr(,"Variedade")
[1] "Valencia"
attr(,"PortaEnxerto")
[1] "Cravo"
attr(,"LinhasPlantas")
[1] "20"
attr(,"PlantasLinhas")
[1] "48"
attr(,"EntreLinhas")
[1] "7.5"
attr(,"DentroLinhas")
[1] "4"
attr(,"Plantio")
[1] "1991"
attr(,"Latitude")
[1] ""
attr(,"Longitude")
[1] ""
```

2.2 Classe `citrus` e `geodata`

A classe `geodata` no **geoR** é uma lista contendo basicamente dois elementos: `coords`, da classe `matrix` com duas colunas representando as coordenadas das plantas no talhão, e `data`, da classe `numeric` com os atributos observados em cada planta. No **Rcitrus**, o elemento `data` é um `data.frame` onde cada coluna é uma avaliação no tempo e cada linha contém atributos e uma planta no tempo. Além disso há um elemento adicional, `date`, da classe `character` com as datas das avaliações. Este formato é utilizado para aplicação de análise de processos pontuais.

Pode-se converter dados de uma ou mais avaliações para o formato `geodata`, usando a função `as.citrus()` e colocando a opção `find.form='geodata'`.

No caso de converter dados de uma única avaliação, dispostos em um simples `data.frame`, os dados de Itajobi por exemplo, deve-se informar o espaçamento entre linhas `x` e dentro das linhas `y`:

```
> geo.ita <- as.citrus(ita, find.form = "geodata", x = 8, y = 4)

Carregando pacotes exigidos: geoR
x = 8 y = 4
```

Ao digitar o nome do objeto temos a descrição sucinta dos dados:

```
> geo.ita

Disease plant data in 1 evaluations of
62 rows of plants and 58 plants in each row.
```

Inspencionando o objeto:

```
> class(geo.ita)

[1] "citrus"  "geodata"

> names(geo.ita)

$coords
NULL

$data
[1] "e1"

$other
[1] "dates"
```

A conversão dos dados de MSC do talhão 303, pode ser feita de forma semelhante à conversão para `array` e adicionando `find.form="geodata"`. Neste caso, não é necessário informar o espaçamento, pois este já é informado nos atributos do talhão:

```
> o303.geo <- as.citrus(ori303, find.form = "geodata", nrow = 20,
+     row.id = 1, n.att = 13)

x = 7.5 y = 4
```

Inspencionando o objeto:

```
> o303.geo

Disease plant data in 25 evaluations of
20 rows of plants and 48 plants in each row.

> class(o303.geo)

[1] "citrus"  "geodata"

> names(o303.geo)

$coords
NULL

$data
 [1] "e1"  "e2"  "e3"  "e4"  "e5"  "e6"  "e7"  "e8"  "e9"  "e10" "e11"
[12] "e12" "e13" "e14" "e15" "e16" "e17" "e18" "e19" "e20" "e21" "e22"
[23] "e23" "e24" "e25"

$other
[1] "dates"

> dim(o303.geo$data)

[1] 960  25
```

2.3 Classe `citrus` e `Surv`

A classe `Surv` é um objeto que contém informações do tempo de progresso da doença. No **Rcitrus**, é uma lista contendo: `coords` com as coordenadas das plantas no talhão e `data` com duas colunas: a primeira da classe `Surv` com tempo ate ocorrência do evento de interesse em formato padrão para análise de sobrevivência; e a segunda coluna contendo inteiros indicadores das linhas do elemento coords. Este formato pode ser usado na análise de sobrevivência com censura intervalar.

Para exemplo, tomaremos os dados validados do talhao 303, incluido no pacote **Rcitrus** sob nome `v303.geo`. Esses dados estão no mesmo formado do objeto `o303.geo`, porém já estão validados. Vamos considerar o evento morte como sendo o de interesse. Nesses dados as plantas sadias são representadas pelo código 0 (zero) e as mortas pelo código 3 (três).

Carregando os dados:

```
> data(v303.geo)
> class(v303.geo)

[1] "citrus"  "geodata"

> v303.geo

Disease plant data in 25 evaluations of
20 rows of plants and 48 plants in each row.
```

Basta usar a função `citrus.conv` para fazer a conversão:

```
> v303.surv <- citrus.conv(v303.geo, find.form = "Surv", cod.start = 0,
+     cod.event = 3)

Carregando pacotes exigidos: survival
Carregando pacotes exigidos: splines
```

Inspencionando os dados

```
> class(v303.surv)

[1] "citrus" "Surv"

> names(v303.surv)

[1] "coords"       "data"         "dates"        "n.subst.code"
[5] "unselect"     "invalids"
```

Os intervalos de tempo das dez primeiras plantas do talhão:

```
> v303.surv$data[1:10, ]

   left right
1   248   833
2   248    NA
3   282    NA
4   339    NA
5   310   770
6   267    NA
7   248    NA
8   282   833
9   267   739
10  339   833
```

Observa-se que a planta da posição 158 já tinha sintomas de MSC na primeira avaliação e não chegou à morte na última avaliação:

```
> v303.geo$data[155:160, ]
```

```
    Av1 Av2 Av3 Av4 Av5 Av6 Av7 Av8 Av9 Av10 Av11 Av12 Av13 Av14 Av15 Av16
155   0   0   0   0   0   0   0   0   0    0    1    1    1    1    1    1
156   0   0   0   0   0   0   0   0   0    0    1    1    1    1    1    1
157   0   0   0   0   0   1   1   1   1    1    1    1    1    1    1    1
158   1   1   2   2   2   2   2   2   2    2    2    2    2    2    2    2
159   0   0   0   0   0   0   0   0   0    0    1    1    1    1    1    1
160   0   0   0   0   0   0   0   0   1    1    1    1    1    1    1    1
    Av17 Av18 Av19 Av20 Av21 Av22 Av23 Av24 Av25
155    1    1    1    1    1    1    1    3    3
156    2    2    2    3    3    3    3    3    3
157    1    1    1    2    2    2    2    3    3
158    2    2    2    2    2    2    2    2    2
159    1    1    1    1    1    1    1    1    1
160    1    1    1    1    1    1    1    1    1
```

Então:

```
> v303.surv$data[155:160, ]
```

```
    left right
155  282   833
156  282   581
157  240   833
158   NA    NA
159  282    NA
160  267    NA
```

3 Escrevendo dados em arquivos texto

Os dados da classe `array` ou `geodata`, podem ser salvos em arquivo texto. Ambos os formatos podem ser salvos usando a função `write.citrus()`, porém o arquivos de dados terão formatos diferentes.

Para o objeto da classe `geodata`, a planilha terá duas colunas indicando as coordenadas das plantas e colunas adicionais dos atributos em cada avaliação.

```
> write.citrus(o303.geo, "og303.txt")
```

Para a classe `array`, o arquivo terá os atributos de cada avaliação abaixo uma das outras, como no arquivo `original303.csv`.

```
> write.citrus(o303.array, "oar303.txt")
```

Esses arquivos podem ser lidos novamente. Usando a função `read.citrus()` para ler o arquivo dos dados da classe `array`:

```
> oar303 <- read.citrus("oar303.txt", nrow = 20, row.id = 1,
+     n.att = 13)
> oar303
```

```
Disease plant data in 25 evaluations of
20 rows of plants and 48 plants in each row.

> class(oar303)

[1] "citrus" "array"
```

A função `read.citrus.geo()` pode ser usada para ler o arquivo dos dados da classe `geodata`. Observando os argumentos dessa função:

```
> args(read.citrus.geo)

function (file, n.att = NULL, header = TRUE, coords.col = 1:2,
    data.col = NULL, sep = "", dec = ".", na.strings = "NA",
    col.nam.att = 1, col.val.att = 2)
NULL
```

Lendo o arquivo:

```
> ogr303 <- read.citrus.geo("og303.txt", n.att = 13)
> ogr303

Disease plant data in 25 evaluations of
20 rows of plants and 48 plants in each row.

> class(ogr303)

[1] "citrus"  "geodata"

> names(ogr303)

$coords
[1] "nam.at" "val.at"

$data
 [1] "X1"  "X2"  "X3"  "X4"  "X5"  "X6"  "X7"  "X8"  "X9"  "X10" "X11"
[12] "X12" "X13" "X14" "X15" "X16" "X17" "X18" "X19" "X20" "X21" "X22"
[23] "X23" "X24" "X25"

$other
[1] "dates"
```

Também é fácil salvar arquivos em formato de texto ".csv" e abrir em planilha OpenOffice ou Excel. No Excel há um detalhe: para abrir arquivos com colunas separadas por ",", deve ser dados um "duplo-clique" no ícone do arquivo, enquanto que arquivos separados por ";", deve ser pelo *menu File Open* ou *Arquivo Abrir*.

```
> write.citrus(ogr303, "og303.csv", sep = ",")
> write.citrus(oar303, "oar303.csv", sep = ",")
```

Agradecimentos

Este trabalho foi desenvolvido como parte das atividades do convênio firmado entre o Fundo de Defesa da Citricultura (FUNDECITRUS) e o Departamento de Estatística da Universidade Federal do Paraná e financiado pelo FUNDECITRUS.