Introdução ao Mapeamento de Doenças

Renato Martins Assunção
Elias Teixeira Krainski

Outubro 2010

Neste capítulo nós introduzimos falando sobre a tipologia de dados espaciais. Nós apresentamos a classificação feita em [3] e outras extensões. Como o foco deste texto é o mapeamento de doenças, que é uma tópico dentro da análise espacial de dados de áreas, nós vamos também associar um mapa de áreas a um grafo e introduzir a leitura de mapas no **R**, [7].

1 Tipologia de Dados Espaciais

No livro *Statistics for Spatial Data* do Noel Cressie, [3], a Estatística Espacial foi dividida em três áreas: Processos Pontuais, Geoestatística e Dados de Áreas. Atualmente há sub-areas ou extensões dessas três áreas, ou ainda, novas áreas. A seguir, colocamos uma definição sucinta dessas três áreas, com exemplos de fenômenos cujos dados possuem referêcia geográfica, isto é, "dados espaciais".

1.1 Processos Pontuais

É a área que estuda a ocorrência de eventos georeferenciados. Os eventos ocorrem numa região $\mathbf{D} \in \Re^2$. Exemplos: localização de árvores numa floresta, localização de defeitos de pintura numa chapa de metal, localização das residências de pessoas com dengue num município. Neste caso, registra-se a localização dos eventos. O dado é apenas uma matriz $n \times 2$, onde as n linhas representam os n eventos e as duas colunas são as coordenadas geográficas.

1.2 Geoestatística

É a área da Estatística Espacial que estuda técnicas de análise estatística de fenômenos de variação espacial contínua. Uma variável aleatória pode assumir valores que dependem da localização de observação: $Y(s)$, onde s é um ponto na região $\mathbf{D} \in \Re^2$. Isto é, o fenômeno estudado pode ter valor diferente para qualquer um dos infinitos pontos nesse espaço.

Exemplos: temperatura máxima no estado do paraná hoje, nível de fertilidade do solo numa fazenda, profundidade até encontrar água. Neste caso, o fenômeno ocorre em qualquer lugar no espaço, mas é medido em apenas n locais distintos. O dado pode ser organizado numa matriz com três colunas: duas para as coordenadas espaciais e uma para o valor medido.

Os dados de fenômenos com variação espacial contínua também são chamados de dados pontualmente referenciados.

1.3 Dados de áreas

É a área que estuda fenômenos com variação espacial discreta. Neste caso, uma região $\mathbf{D}$ é dividida num conjunto de sub-regiões menores $D_1, D_2, \ldots, D_n$ que formam uma partição da região $\mathbf{D}$, com $D_i \cup D_j = \emptyset$ e $\cap_{i=1}^{n} D_i = \mathbf{D}$. O fenômeno estudado apresenta um valor para cada uma das sub-regiões.

Exemplos: PIB *per capita* em cada município do Paraná, taxa de mortalidade em cada município do Paraná, número médio de residentes por domicílio em cada setor censitário de uma cidade. Nesses exemplos, a variável em estudo foi agregada nas sub-areas da região. O dado é um conjunto de n polígonos e um vetor de tamanho n com o valor da variável observado em cada área. Vamos chamar o conjunto de polígonos de mapa.

1.4 Outros Tipos de Dados Espaciais

Na maioria das vezes, ao georeferenciar eventos, também referencia-se o tempo de ocorrência desses eventos, caracterizando um processo pontual espaço-temporal. Também, ao georeferenciar eventos, pode-se anotar uma característica de interesse desse evento, que pode variar de evento para evento, caracterizando um processo pontual marcado. Nos dados de áreas e nos dados pontualmente referenciados, podemos também coletar dados em mais de um tempo, caracterizando dados espaço-temporais.

Alguns fenômenos estudam por exemplo a interação entre dois processos pontuais. Um exemplo é a ocorrência de duas espécies de árvores distintas numa floresta. Outro tipo de interação de processos pontuais é quando dois processos pontuais são ligados. Um exemplo deste tipo é a ligação entre o local de roubo de um carro e o local de encontro do mesmo.

A interação espacial também é estudada quando a referência é área. Por exemplo, suponha uma matriz onde as linhas e colunas são os estados do Brasil e os valores nessa matriz são o número de pessoas que nasceram no referenciado na linha da matriz e vivem no estado referenciado na coluna da matriz.

1.5 Mudança de Suporte Espacial

Há situações quando temos dados de diferentes suportes (ou resoluções) espaciais. A resolução espacial é geralmente dependente do nível de precisão das informações. Os dados de áreas por exemplo, podem ser resultados de agregações de que poderiam ser medidos pontualmente.

Um exemplo é se temos a localização pontual dos pacientes com Dengue numa cidade e queremos analisar o efeito de alguma variável medida a nível de setor censitário. Neste caso a resposta é pontualmente referenciada mas a variável auxiliar é referenciada por áreas. A situação inversa também pode ocorrer. Por exemplo, quando temos dados climáticos medidos em algumas estações metereológicas de um estado e queremos usar um índice climático para explicar a produtividade média de municípios.

A interação espacial pode servir também para mapear ligações telefônicas. Por exemplo, mapear a interação entre áreas, anotando o código de área de destino e origem das chamadas. Neste caso, uma empresa telefônica poderia ter o dado pontualmente referenciado, tendo o ponto exato de origem e destino das chamadas.

2 Mapas como Grafos

Neste texto vamos chamar de mapas um conjunto de polígonos. Podemos considerar que cada um desses polígonos pode ter fronteira com algum(ns) outro(s) polígono(s) do mapa. Se não tiver, será uma ilha no mapa. Essa estrutura de vizinhança pode ser representada por uma matriz, a matriz de adjacência **A**. Vamos considerar o mapa da Figura 2.

No mapa da esquerda da Figura 2 vemos a divisão municipal da microregião de Londrina, composta por seis municípios. No mapa da direita dessa Figura, vemos o grafo associado à vizinhança por fronteira comum entre esses municípios. Na Tabela **??** temos a matriz de adjacência associada ao grafo da Figura 2.

	Pitangueiras	Rolândia	Cambé	Londrina	Tamarana	Ibiporã
Pitangueiras	0	1	0	0	0	0

```
Rolândia                1       0     1       0       0       0
Cambé                   0       1     0       1       0       0
Londrina                0       0     1       0       1       1
Tamarana                0       0     0       1       0       0
Ibiporã                 0       0     0       1       0       0
attr(,"call")
nb2mat(neighbours = nb1, style = "B")
```

Em muitas análises, é bastante comum fazer uma padronização na matriz de adjacência, de forma que as linhas somem 1 (um). A matriz de vizinhança ponderada dessa forma é como a Tabela **??**.

```
             Pitangueiras Rolândia     Cambé Londrina  Tamarana   Ibiporã
Pitangueiras          0.0      1.0 0.0000000      0.0 0.0000000 0.0000000
Rolândia              0.5      0.0 0.5000000      0.0 0.0000000 0.0000000
Cambé                 0.0      0.5 0.0000000      0.5 0.0000000 0.0000000
Londrina              0.0      0.0 0.3333333      0.0 0.3333333 0.3333333
Tamarana              0.0      0.0 0.0000000      1.0 0.0000000 0.0000000
Ibiporã               0.0      0.0 0.0000000      1.0 0.0000000 0.0000000
attr(,"call")
nb2mat(neighbours = nb1, style = "B")
```

Além da matriz de adjacência, onde as arestas tem peso igual, podemos definir matrizes que representem vizinhança com algum tipo de ponderação. Por exemplo, poderíamos usar pesos proporcionais ao inverso do tamanho da fronteira entre os municípios. Qualquer outro critério que seja julgado adequado pode ser usado para representar a estrutura de vizinhança entre as áres.

As técnicas estatisticas aplicadas a dados de áreas geralmente vão estar baseadas na matriz de ajacência ou matriz de adjacência ponderada. Ou seja, dada essa matriz, não precisamos mais do mapa. Vamos usar o mapa apenas para visualizar resultados. Por esse motivo, as técnicas de análise de dados de áreas são extensivas a qualquer estrutura de dados que possa ser representado por um grafo. Um exemplo, seria analisar dados de uma rede de relacionamentos do Orkut, considerando um grafo para representar a estrutura de amizades.

3 Mapas em R

Um mapa de dados de áreas é, basicamente, um conjunto de polígonos. Porém, geralmente temos atributos (variáveis) associados aos polígonos. Assim, um formato padrão de mapa é o *shapefile*, que é um conjunto de pelo menos três arquivos: um arquivo com os polígonos (com extensão .shp), um arquivo com os atributos (com extensão .dbf), e um arquivo com índices (com extensão .shx). Além desse formato há outros que também podem ser lidos em **R**.

3.1 Fonte de Mapas Territoriais no Brasil

O Instituto Brasileiro de Geografia e Estatística (IBGE) disponibiliza um conjunto de milhares de mapas. Há mapas do Brasil dividido por municípios e mapas de cada município dividido por setores censitários. São disponibilizados mapas em três diferentes resoluções gráficas. Os polígonos de cada área (município/setor censitário), podem ser representados por um conjunto de pontos. Quanto mais pontos utilizados na representação de um polígono, mais apropriado será o mapa para a produção de mapas em alta resolução gráfica. O IBGE disponibiliza mapas para três diferentes resoluções.

Um polígono pode estar representados em diferentes projeções ou cartográficas, [6]. O IBGE disponibiliza mapas em duas projeções: Projeção Geográfica e Projeção Policonica. Além disso,

os mapas são disponibilizados em dois formatos de arquivo diferentes: *shapefiles* e *Mge_Dgn*. O mapa de municípios está disponível em arquivo para o Brasil todo, por Região e por Unidade da Federação.

3.2 Lendo *Shapefiles*

Há mais de um pacote para leitura de mapas em **R** [8]. Nós vamos ler *shapefiles* usando o pacote **maptools** [4]. Este pacote usa as classes de dados espaciais definidas no pacote **sp** [1].

Vamos considerar o mapa do Estado do Paraná dividido em municípios. No exemplo, nós usamos o mapa disponível no *site* do IBGE, seguindo os *links* <Geociências> + <Mapeamento das unidades territoriais> + <Produtos> + <Malha municipal digital 2007>. Na nova janela, escolhemos E2500 (entre as opções E500, E100 e E2500), escolhemos a projeção geográfica <Proj_Geográfica>, escolhemos o formato *shapefiles* <ArcView_shp>, escolhemos o nível territorial Unidade da Federação <UF> e escolhemos o estado do Paraná <PR>.

Inicialmente carregamos o pacote **maptools**.

```
> require(maptools)
```

Vamos considerar que o conjunto de arquivos do *shapefiles* está do diretório "mapas" do diretório corrente. Vamos usar a função `readShapePoly()` para ler o mapa.

```
> pr <- readShapePoly("mapas/41mu2500g")
```

O objeto `pr` é contém o conjunto de polígonos dos 399 municípios do paraná e uma tabela de atributos, um `data.frame` com 399 linhas. Podemos visualizar esse mapa simplesmente fazendo

```
> par(mar = c(0, 0, 0, 0))
> plot(pr)
```

onde `par(mar=c(0,0,0,0))` foi usado para tirar as margens da janela gráfica, aumentando sua área útil. O mapa produzido por este comando está na Figura 3.2.

3.3 Obtendo Matriz de Adjacência

A matriz de adjacência é uma matriz que na maioria das vezes é esparsa, isto é, contém muitos zeros. Por este motivo, podemos economizar memória do computador representado-a por uma lista de adjacência. Neste caso, cada elemento da lista representa uma área e seu conteúdo é um vetor indicando seus vizinhos.

A lista de adjacência do mapa lido pode ser obtida usando a função `poly2nb()`, assim:

```
> nbpr <- poly2nb(pr)
> length(nbpr)

[1] 399

> nbpr[[1]]

[1]  29  85 177 302 317 332
```

Neste exemplo, as áreas 29, 85, 177, 302, 317 e 332 são vizinhas da área 1. Obtento a matriz de adjacência neste caso, dos 399 elementos da primeira linha, teríamos apenas seis elementos não nulos! Podemos ver também um sumário da lista de vizinhança

```
> summary(nbpr)
```

```
Neighbour list object:
Number of regions: 399
Number of nonzero links: 2240
Percentage nonzero weights: 1.407026
Average number of links: 5.614035
Link number distribution:

 2  3  4  5  6  7  8  9 10 11 12
 7 39 72 80 88 56 29 15 10  2  1
7 least connected regions:
30 49 172 253 273 330 356 with 2 links
1 most connected region:
68 with 12 links
```

Se realmente houver necessidade em obter a matriz de adjacência ou a matriz de adjacência ponderada, podemos usar a função `nb2mat()`. A matriz de adjacência pode ser obtida por

```
> A.pr <- nb2mat(nbpr, style = "B")
```

e matriz de adjacência ponderada pode ser obtida por

```
> W.pr <- nb2mat(nbpr)
```

4 Mapa Temático

Um mapa temático é um mapa onde as sub-áreas do mapa são coloridas de acordo com um atributo (variável). Por exemplo, podemos colorir o mapa do Paraná dividido em municípios, colorindo cada município com as cores definidas em função da Taxa de Mortalidade infantil.

Como exemplo, vamos produzir um mapa temático usando os dados de Mortalidade Infantil da Carolina do Norte, disponíveis no pacote **spdep**, [2].

4.1 Um Pouco Mais Sobre Mapas

Inicialmente vamos buscar o nome do *shapefile* a ser lido:

```
> file <- system.file("etc/shapes/sids.shp", package = "spdep")[1]
```

Agora vamos carregar pacote **maptools** para carregar importar o *shapefile*:

```
> require(maptools)
```

e ler o *shapefile*:

```
> nc <- readShapePoly(file, ID = "FIPSNO", proj4string = CRS("+proj=longlat +ellps=clrk66"))
```

Após isso, vamos inspecionar o objeto que guarda o mapa.

```
> class(nc)
```

```
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
```

```
> names(nc)
```

```
 [1] "SP_ID"     "CNTY_ID"   "east"      "north"     "L_id"      "M_id"
 [7] "names"     "AREA"      "PERIMETER" "CNTY_"     "NAME"      "FIPS"
[13] "FIPSNO"    "CRESS_ID"  "BIR74"     "SID74"     "NWBIR74"   "BIR79"
[19] "SID79"     "NWBIR79"
```

Esse objeto é da classe `SpatialPolygonsDataFrame`, que é um objeto que contém basicamente uma lista de polígonos e um `data.frame` com atributos. O `names(nc)` retorna os nomes das variáveis contidas no `data.frame`.

O objeto `nc` possui atributos, que são os elementos contidos nele:

```
> names(attributes(nc))

[1] "bbox"        "proj4string" "polygons"    "plotOrder"   "data"
[6] "class"
```

A lista de polígonos é o `attribute polygons` do objeto e o `data.frame` é o `attribute data`. Podemos inspecionar esses atributos colocando após o nome do objeto:

```
> nc@bbox

        min       max
x -84.32385 -75.45698
y  33.88199  36.58965

> nc@proj4string

CRS arguments: +proj=longlat +ellps=clrk66

> length(nc@polygons)

[1] 100

> nc@plotOrder

  [1]  82  24  78   9  92  71  10  51  31   7  76  97  67  42  63  19   8  25
 [19]  26  41  11  74  48  90  49  43  79  66  81  29  44  60  39  86  96  64
 [37]  87   4  80  12  33  57  62  35  50  16  85  14  52  77  20  93  59  58
 [55]  23  56   5  17   1  73  34  18  68  54  84  94  89  47  88  46  45  98
 [73]  28  13  36  37  99  69  95 100  83  55  38  32  91  40  27  30   2  53
 [91]   6  72  15   3  75  61  70  22  21  65

> dim(nc@data)

[1] 100  20

> nc@class

[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
```

Podemos também inspecionar cada polínogo colocando duplo colchetes, assim:

```
> str(nc@polygons[[1]])
```

```
Formal class 'Polygons' [package "sp"] with 5 slots
  ..@ Polygons :List of 1
  .. ..$ :Formal class 'Polygon' [package "sp"] with 5 slots
  .. .. .. ..@ labpt  : num [1:2] -79.4 36
  .. .. .. ..@ area   : num 0.111
  .. .. .. ..@ hole   : logi FALSE
  .. .. .. ..@ ringDir: int 1
  .. .. .. ..@ coords : num [1:10, 1:2] -79.2 -79.2 -79.5 -79.5 -79.5 ...
  ..@ plotOrder: int 1
  ..@ labpt    : num [1:2] -79.4 36
  ..@ ID       : chr "37001"
  ..@ area     : num 0.111
```

Ou inspecionar o `data.frame`:

```
> str(nc@data)
```

```
'data.frame':        100 obs. of  20 variables:
 $ SP_ID    : Factor w/ 100 levels "37001","37003",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ CNTY_ID  : num  1904 1950 1827 2096 1825 ...
 $ east     : num  278 179 183 240 164 138 406 411 321 353 ...
 $ north    : num  151 142 182 75 176 154 118 148 53 6 ...
 $ L_id     : num  1 2 1 3 1 1 2 1 4 4 ...
 $ M_id     : num  3 2 2 2 2 2 4 4 3 3 ...
 $ names    : Factor w/ 100 levels "Alamance","Alexander",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ AREA     : num  0.111 0.066 0.061 0.138 0.114 0.064 0.203 0.18 0.225 0.212 ...
 $ PERIMETER: num  1.39 1.07 1.23 1.62 1.44 ...
 $ CNTY_    : num  1904 1950 1827 2096 1825 ...
 $ NAME     : Factor w/ 100 levels "Alamance","Alexander",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ FIPS     : Factor w/ 100 levels "37001","37003",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ FIPSNO   : num  37001 37003 37005 37007 37009 ...
 $ CRESS_ID : num  1 2 3 4 5 6 7 8 9 10 ...
 $ BIR74    : num  4672 1333 487 1570 1091 ...
 $ SID74    : num  13 0 0 15 1 0 7 6 8 5 ...
 $ NWBIR74  : num  1243 128 10 952 10 ...
 $ BIR79    : num  5767 1683 542 1875 1364 ...
 $ SID79    : num  11 2 3 4 0 0 4 5 5 6 ...
 $ NWBIR79  : num  1397 150 12 1161 19 ...
 - attr(*, "data_types")= chr  "C" "N" "N" "N" ...
```

4.1.1 Mapa Temático

Vamos usar `SID79` e `BIR79` para produzir um mapa temático da Mortalidade Infantil na Carolina do Norte em 1979. Inicialmente vamos carregar o pacote **spdep** para usar função que calcula a taxa de mortalidade em cada área do mapa, o número esperado de óbitos se a taxa em cada área fosse a mesma na região toda, o risco relativo (razão entre o número de óbitos observado e esperado) e a probabilidade de se observar um número observado de óbitos maior que o observado em cada área se a taxa de mortalidade fosse constante na região. Carregando o pacote **spdep** e calculando essas estatísticas:

```
> require(spdep)
> summary(pm79 <- probmap(nc$SID79, nc$BIR79))
```

```
      raw              expCount          relRisk            pmap
 Min.   :0.000000   Min.   : 0.6314   Min.   :  0.00   Min.   :0.0002270
```

```
 1st Qu.:0.001249   1st Qu.: 2.6447   1st Qu.: 63.13   1st Qu.:0.3241328
 Median :0.002075   Median : 5.2172   Median :104.86   Median :0.6515481
 Mean   :0.002039   Mean   : 8.3600   Mean   :103.02   Mean   :0.5794970
 3rd Qu.:0.002539   3rd Qu.: 9.6763   3rd Qu.:128.27   3rd Qu.:0.8117612
 Max.   :0.006114   Max.   :60.8744   Max.   :308.91   Max.   :0.9999693
```

No sumário observamos que houve taxa número de óbitos observado igual a zero.

Agora, vamos classificar a taxa em 7 categorias. Para isso, vamos criar sete pontos de corte para taxa:

```
> (q7 <- quantile(pm79$raw, 0:7/7))

         0%   14.28571%   28.57143%   42.85714%   57.14286%   71.42857%
0.000000000 0.000930612 0.001356446 0.001869994 0.002160810 0.002428930
  85.71429%        100%
0.003226070 0.006113871
```

e classificar a taxa em sete classes:

```
> table(cl7 <- findInterval(pm79$raw, q7, TRUE))

 1  2  3  4  5  6  7
15 14 14 14 14 14 15
```

O número de sete cores foi recomendado por Linda Picle do *Center of Disease Control* (CDC) através de experimentação. Além disso, ela sugere também as cores. Estas abrange tons de Azul e Vermelho, organizados no sentido de que áreas com menor taxa de mortalidade sejam pintadas em Azul mais forte e áreas com maior taxa sejam pintadas em Vermelho mais forte. Os tons sugeridos pode ser criado usando a função `brewer.pal()` do pacote **RColorBrewer** [5].

```
> require(RColorBrewer)
> c7 <- c(rev(brewer.pal(4, "Blues")), brewer.pal(3, "Reds"))
```

A partir da classificação de cada área numa das sete categorias e da definição das cores, podemos visualizar o mapa (Figura 4.1.1):

```
> par(mar = c(0, 0, 0, 0))
> plot(nc, col = c7[cl7])
> legend("bottomleft", leglabs(format(1000 * q7, dig = 2), "<",
+     ">", "a"), bty = "n", ncol = 2, fill = c7, title = "Taxa por mil")
```

Referências

[1] 2005. R package version 2.1-4.

[2] Roger Bivand, with contributions by Micah Altman, Luc Anselin, Renato Assunço, Olaf Berke, Andrew Bernat, Eric Blankmeyer, Marilia Carvalho, Yongwan Chun, Bjarke Christensen, Carsten Dormann, Stéhane Dray, Rein Halbersma, Elias Krainski, Nicholas Lewin-Koh, Hongfei Li, Jielai Ma, Giovanni Millo, Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Gianfranco Piras, Markus Reder, Michael Tiefelsdorf, , and Danlin Yu. *spdep: Spatial dependence: weighting schemes, statistics and models*, 2010. R package version 0.5-10.

[3] Noel A. C. Cressie. *Statistics For Spatial Data*. Wiley, 1995. Revised Edition.

[4] Nicholas J. Lewin-Koh, Roger Bivand, contributions by Edzer J. Pebesma, Eric Archer, Adrian Baddeley, Hans-Jörg Bibiko, Stéphane Dray, David Forrest, Michael Friendly, Patrick Giraudoux, Duncan Golicher, Virgilio Gómez Rubio, Patrick Hausmann, Thomas Jagger, Sebastian P. Luque, Don MacQueen, Andrew Niccolai, Tom Short, and Ben Stabler. *maptools: Tools for reading and handling spatial objects*, 2010. R package version 0.7-34.

[5] Erich Neuwirth. *RColorBrewer: ColorBrewer palletes*, 2007. An R package version 1.0-2.

[6] C. Oliveira. *Curso de Cartografia Moderna*. Ed. IBGE, 1988.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[8] Virgilio Gomez-Rubio Roger S. Bivand, Edzer J. Pebesma. *Applied spatial data analysis with R*. Springer, New York, 2008. http://www.asdar-book.org/.

Figura 1: Microregião de Londrina

Figura 2: Mapa do estado do Paraná dividido em municípios.

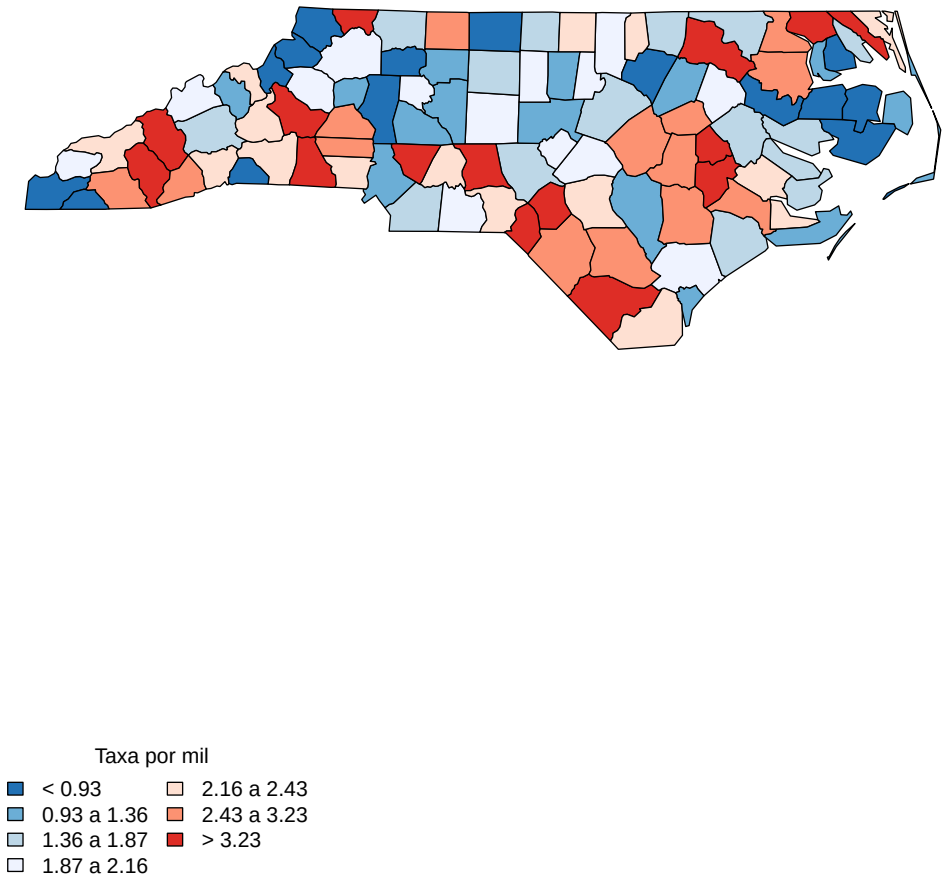

Figura 3: Mapa de mortalidade Infantil na Carolina do Norte em 1979