Data Scientist - Visualização espacial

Soraia Pereira[a], Tiago Marques[a,b]

[a] CEAUL e FCUL, Universidade de Lisboa
[b] CREEM, University of St Andrews, e Dept de Biologia Animal, FCUL

FCUL, 18 de fevereiro de 2020

House Keeping

Recursos do curso disponíveis na pasta

https://tinyurl.com/CEAULGADESCursoRM4

O curso decorre entre as 18:30 e as 22:30

Entre as 20:30 e as 20:45 faremos uma pausa para café.

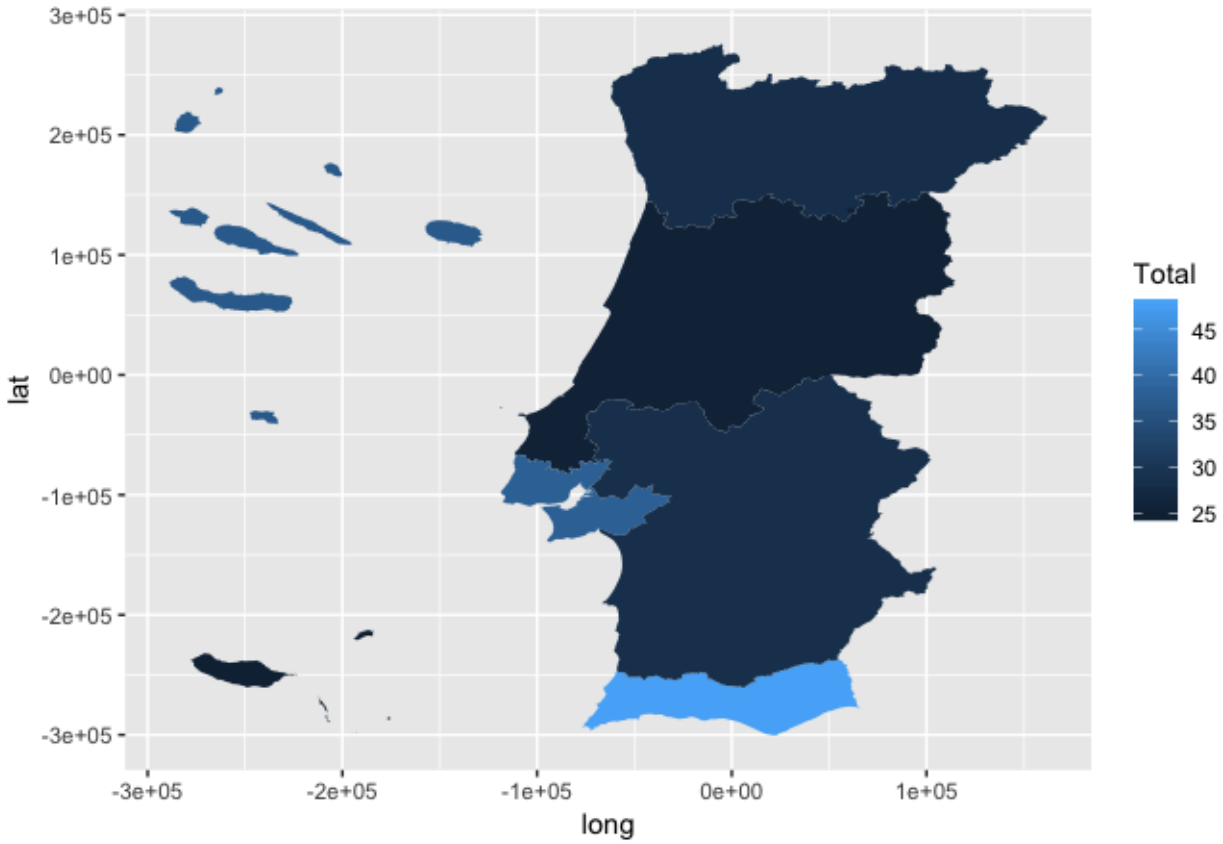
Introdução

A estatística espacial tem ganho interesse nas mais diversas áreas, tais como epidemiologia, ambiente, engenharia, saúde, biologia, ecologia, economia, entre outras.
Dados com uma natureza espacial fornecem não só informação sobre os atributos de interesse mas também sobre a localização espacial desses atributos.
A literatura existente de estatística espacial faz uma distinção clara entre três tipos de dados espaciais:

- Dados de área
- Padrões pontuais espaciais
- Dados referenciados por pontos

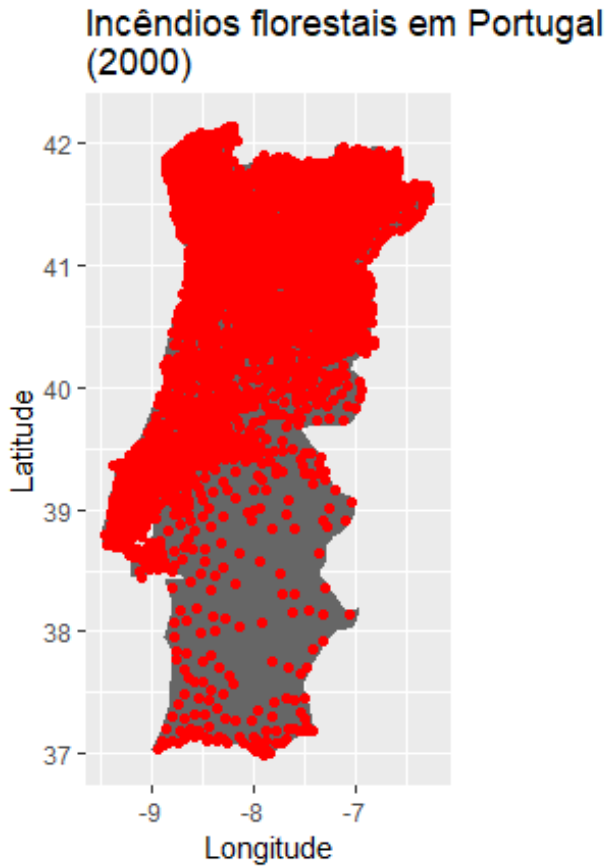
Dados de área

Exemplo ilustrativo: Taxa de criminalidade (permilagem) por regiões NUTS II de Portugal no ano de 2018.



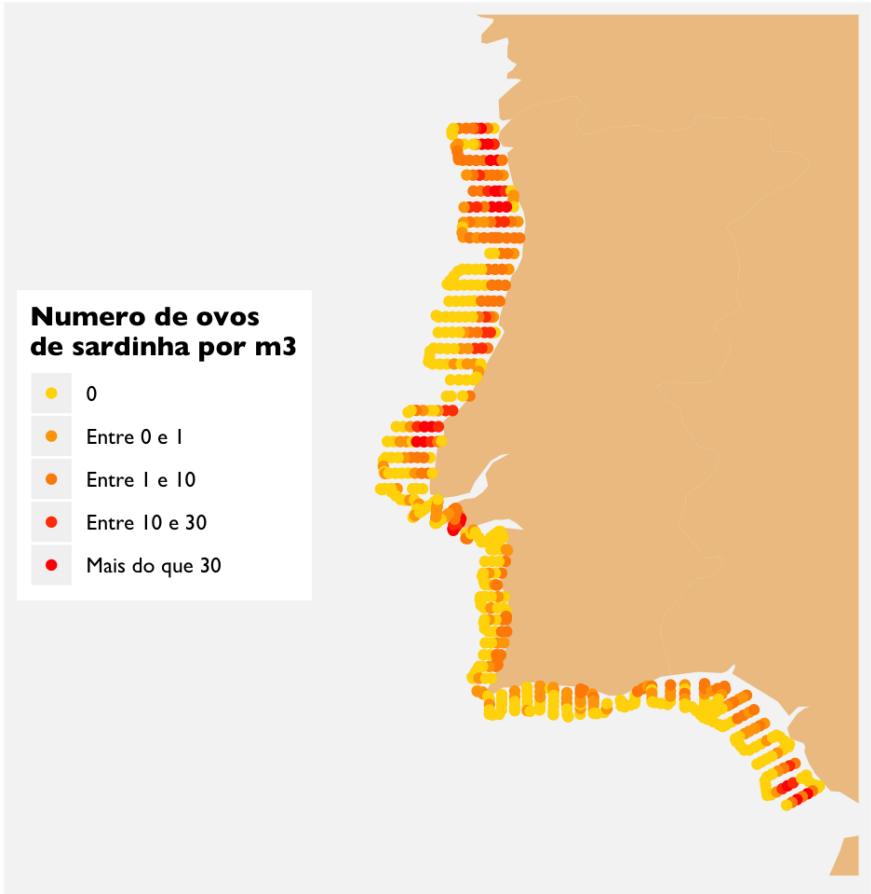
Padrões pontuais espaciais

Exemplo ilustrativo: Localização dos fogos florestais registados em Portugal Continental no ano de 2000.



Dados referenciados por pontos

Exemplo ilustrativo: Número de ovos de sardinha por m^3 observadas em localizações igualmente espaçadas, em transectos perpendiculares à linha da costa, numa campanha do IPMA.



Caso prático - criminalidade 2018

Suponhamos que pretendemos obter a representação da taxa de criminalidade em 2018 por região NUTS II de Portugal, conforme o slide 3. Este conjunto de dados foi extraído do site do INE.

```
> crim2018<-read.table("criminalidade2018.csv",header=TRUE,sep=";")
> head(crim2018)
                      Regiao Codigo Total Crimes.integridade.fisica Furto.via.publica
1                   Portugal     PT  32.4                       5.1                0.8
2                 Continente   CONT  32.1                       5.0                0.8
3                      Norte     NT  28.2                       4.9                0.5
4                     Centro     CT  25.4                       4.4                0.3
5 Area Metropolitana de Lisboa    AML  38.0                       5.3                1.7
6                   Alentejo     AL  28.4                       5.0                0.3
  Furto.veiculo Conducao.alcool Conducao.sem.habilitacao Crimes.patrimonio
1           3.3             1.8                      0.9               16.6
2           3.4             1.7                      0.9               16.6
3           4.2             1.4                      0.7               14.6
4           1.9             1.8                      0.7               12.4
5           3.9             1.5                      1.1               21.6
6           1.7             1.7                      0.9               13.1
```

Importação de ficheiro com formato shapefile

- Uma das funções mais utilizadas para importação de shapefiles é a `readOGR` do package `rgdal`.
- A importação do ficheiro com informação espacial das regiões NUTS II ("NutsII2002.shp") e a respetiva visualização pode ser feita da seguinte forma:

```
nuts2<-readOGR("NutsII2002.shp")
plot(nuts2).
```

Representação espacial usando o ggplot

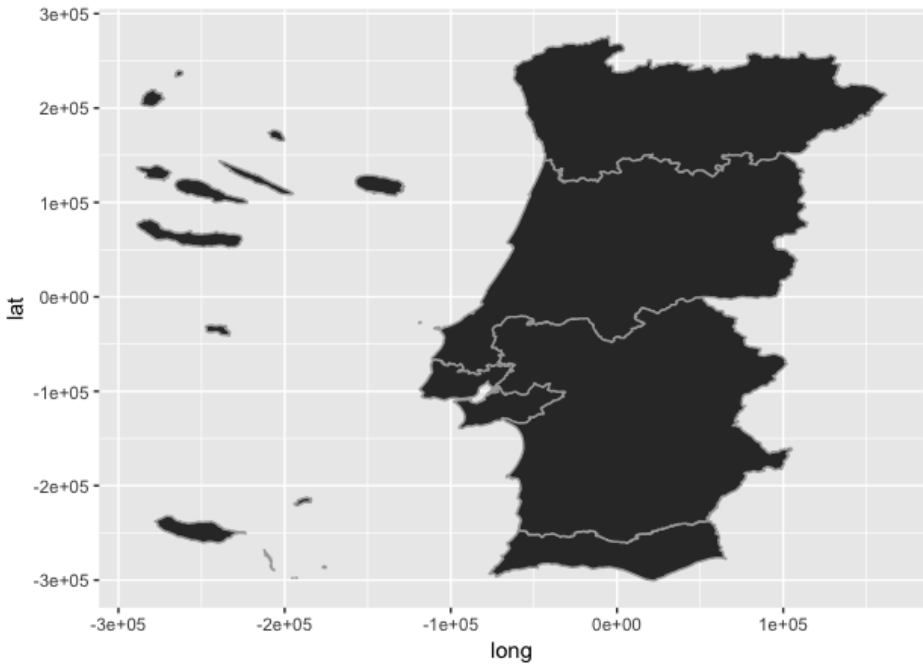
- Alternativamente, podemos recorrer à função `ggplot` do package `tidyverse` para a representação espacial. Esta função traz muitas vantagens quando pretendemos associar atributos às regiões.
- As funções `gSimplify` do package `rgeos` e `tidy` do package `broom` são muito úteis quando pretendemos utilizar a função `ggplot`. A função `gSimplify` permite simplificar a fronteira do domínio (tornando o objecto menos pesado) e a função `tidy` permite converter um objecto espacial com formato .shp numa tabela data.frame.

```
library(rgeos)
nuts2_s<-gSimplify(nuts2,tol=100)

library(broom)
nuts2_dt<-tidy(nuts2_s)

library(tidyverse)
ggplot(nuts2_dt, aes(x = long, y = lat)) +
  geom_polygon(aes( group = group),colour="darkgray")
```

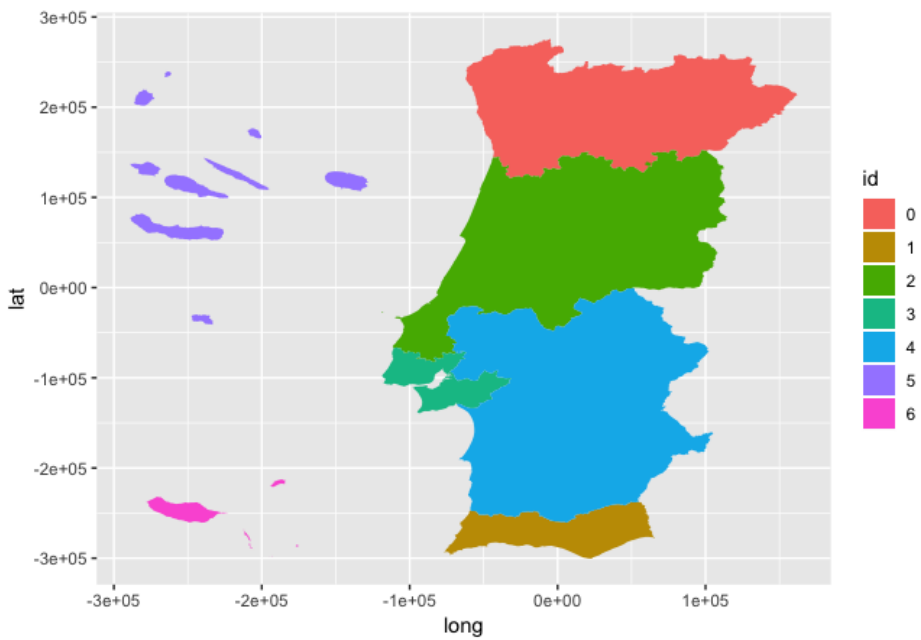
Representação espacial usando o ggplot



Representação espacial com atributos

- Para a representação da variável de interesse por região, devemos "ligar" as duas fontes de informação (informação da variável e informação espacial).
- Neste caso, o objecto espacial não inclui os nomes das regiões, pelo que vamos cruzar os dois ficheiros pelo "id".
- Correspondência entre o "id" e a região:

```
ggplot(nuts2_dt, aes(x = long, y = lat)) +
geom_polygon(aes( group = group,fill=id))
```



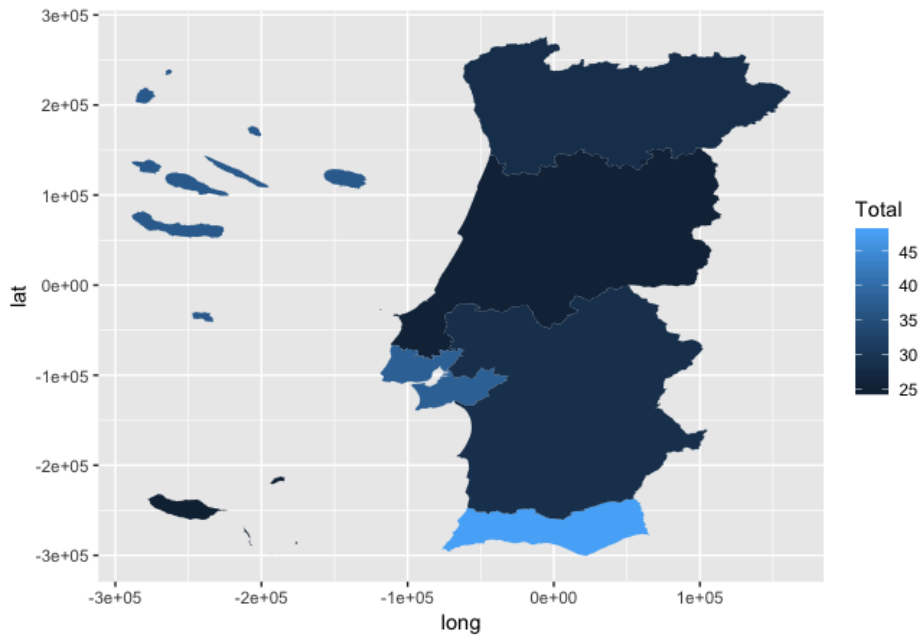
Representação espacial com atributos

- Note-se que as duas primeiras linhas da data.frame crim2018 não são regiões NUTS II, pelo que devem ser eliminadas.
- Por fim, devemos acrescentar o "id" correspondente a cada região e juntar os dois data.frames por "id".

```
> crim2018<-crim2018[3:9,]
> crim2018$id<-c(0,2,3,4,1,5,6)
> crim<-merge(nuts2_dt,crim2018,by="id")
> head(crim)
  id      long      lat order  hole piece group Regiao Codigo Total Crimes.integridade.fisica
1  0 -59383.48 223843.4     1 FALSE     1   0.1  Norte     NT  28.2                       4.9
2  0 -59590.31 225055.6     2 FALSE     1   0.1  Norte     NT  28.2                       4.9
3  0 -59749.39 225137.4     3 FALSE     1   0.1  Norte     NT  28.2                       4.9
4  0 -59903.22 225768.2     4 FALSE     1   0.1  Norte     NT  28.2                       4.9
5  0 -60244.29 225876.4     5 FALSE     1   0.1  Norte     NT  28.2                       4.9
6  0 -61071.63 228387.8     6 FALSE     1   0.1  Norte     NT  28.2                       4.9
  Furto.via.publica Furto.veiculo Conducao.alcool Conducao.sem.habilitacao Crimes.patrimonio
1               0.5           4.2             1.4                     0.7              14.6
2               0.5           4.2             1.4                     0.7              14.6
3               0.5           4.2             1.4                     0.7              14.6
4               0.5           4.2             1.4                     0.7              14.6
5               0.5           4.2             1.4                     0.7              14.6
6               0.5           4.2             1.4                     0.7              14.6
```

Representação espacial com atributos

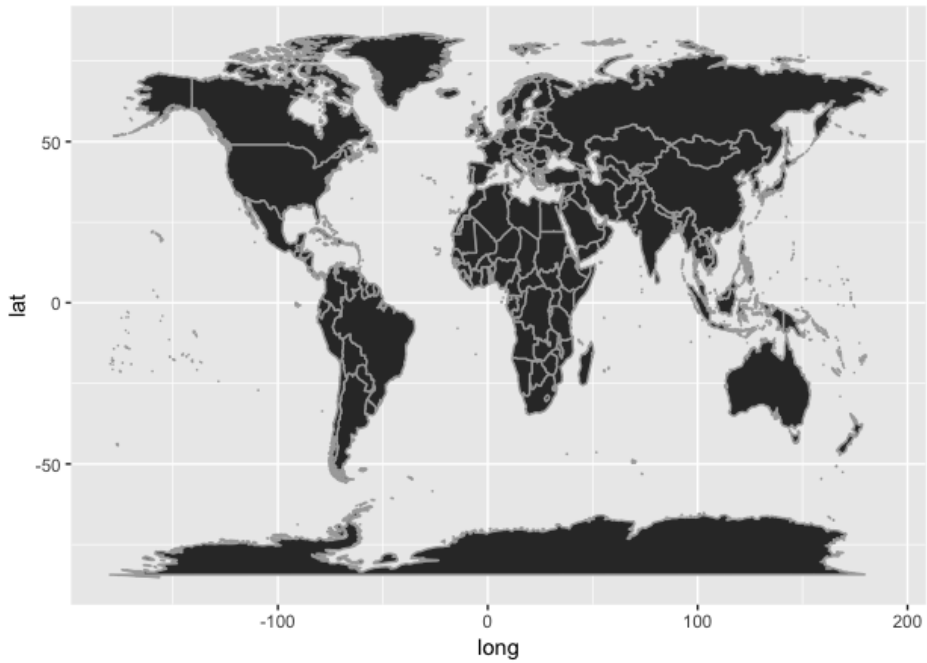
- A representação espacial do atributo "Total" (taxa de criminalidade medida em permilagem) pode ser obtida por:
  `ggplot(crim, aes(long, lat))+ geom_polygon(aes(group = group,fill = Total ))`



Package maps

Quando se pretende fazer uma representação espacial de dados de área por país, o package `maps` pode ser muito útil.
Este package contém informação geográfica de todos os países do mundo e é possível selecionar um conjunto de países específico.

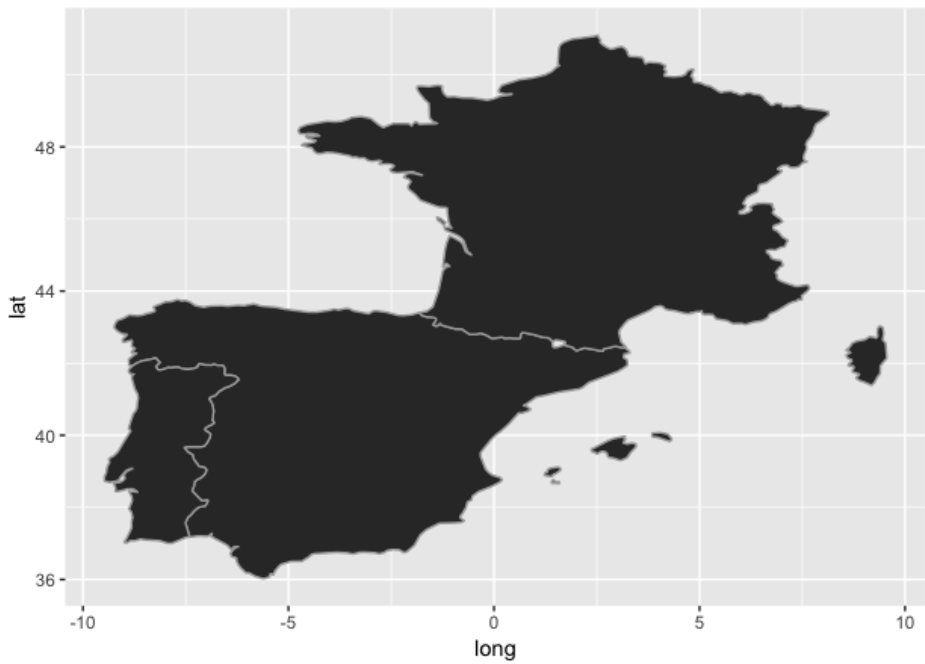
```
library(maps)
world_map <- map_data("world")
ggplot(world_map, aes(x = long, y = lat)) +
geom_polygon(aes( group = group),colour="darkgray")
```



Package maps

Suponhamos que estamos interessados apenas nos mapas de Portugal, Espanha e França.

```
countries <- c("Portugal", "Spain", "France")
countries.maps <- map_data("world", region = countries)
ggplot(countries.maps, aes(x = long, y = lat)) +
geom_polygon(aes( group = group),colour="darkgray")
```



Caso prático - fogos florestais 2000

- Suponhamos agora que pretendemos representar as localizações dos fogos florestais registados em 2000 em Portugal. Este conjunto de dados foi extraído do seguinte link: http://www2.icnf.pt/portal/florestas/dfci/inc/cartografia/base-dados-1980a2000

```
> fogos<-read.table("FOGOS_1980_2000.csv",header=TRUE,sep=";",dec=",",na.strings="")
> fogos<-filter(fogos,Ano=="2000")
> head(fogos)
     ID   INE DISTRITO CONCELHO   FREGUESIA         x        y          lat          lon  Ano
1 151436 11107   Aveiro Mealhada Vacari\x8da -8.404719 40.36434 40 21 51.616 -8 24 16.987 2000
2 151579 11102   Aveiro Mealhada Barcou\x8do -8.484393 40.30259 40 18 09.323 -8 29 03.813 2000
3 151580 11106   Aveiro Mealhada  Pampilhosa -8.429501 40.33872 40 20 19.392 -8 25 46.205 2000
4 151605 11106   Aveiro Mealhada  Pampilhosa -8.429501 40.33872 40 20 19.392 -8 25 46.205 2000
5 151644 11102   Aveiro Mealhada Barcou\x8do -8.484393 40.30259 40 18 09.323 -8 29 03.813 2000
6 151645 11102   Aveiro Mealhada Barcou\x8do -8.484393 40.30259 40 18 09.323 -8 29 03.813 2000
          Data_inicio             Data_fim CAUSA A_POV A_MATOS Area_Total_Florestal A_PUB A_PRIVADA REAC
1 2000-10-28 14:28:00.000 2000-10-28 16:25:00.000  <NA>  0.00    0.15                0.15     0      0.15    0
2 2000-06-26 21:29:00.000 2000-06-26 23:05:00.000  <NA>  0.00    0.12                0.12     0      0.12    0
3 2000-06-26 20:29:00.000 2000-06-26 21:05:00.000  <NA>  0.00    0.01                0.01     0      0.01    0
4 2000-06-30 15:03:00.000 2000-06-30 18:10:00.000  <NA>  0.20    0.00                0.20     0      0.20    0
5 2000-05-19 21:24:00.000 2000-05-19 22:30:00.000  <NA>  0.02    0.00                0.02     0      0.02    0
6 2000-05-31 18:23:00.000 2000-05-31 19:35:00.000  <NA>  0.40    0.00                0.40     0      0.40    0
```

Caso prático - fogos florestais 2000

Para a representação dos pontos, basta acrescentar `geom_point` com a informação georreferenciada pretendida:

```
portugal <- map_data("world", region = "Portugal")
ggplot() + geom_polygon(data=portugal, aes(x=long, y=lat,
group=group))+ geom_point(data=fogos, aes(x=x, y=y),
color="red",size=1.3)+ coord_fixed(1.2) +
xlab("Longitude") + ylab("Latitude") + ggtitle("Incêndios
florestais em Portugal (2000)")
```