

Construindo bases de dados via Webscraping

Benilton Carvalho & Guilherme Ludwig

Webscraping

- É possível "raspar" (*scrape*) informações de páginas da internet, e guardá-las em um banco de dados. A prática é chamada de *webscraping*.
- Nós iremos usar o pacote `rvest`, que simplifica algumas operações dos pacotes `xml2` e `httr`.

```
library(tidyverse)  
library(RSQLite)  
library(httr)  
library(rvest)
```

Idéias

Uma página da web é um documento que pode ser exibido por um navegador. Estes documentos normalmente exibem resultados de consultas à bancos de dados, que são nosso principal interesse nesta disciplina. Em geral:

- Páginas simples podem ser acessadas através do R com o pacote `rvest`.
- Páginas dinâmicas podem exigir alguma autenticação do usuário, na forma de *cookies*. Para acessar essas páginas, podemos precisar do pacote `httr`.
- Nosso objetivo é coletar dados com o `rvest` (e talvez `httr`) e armazená-los em um banco de dados.

Alguns recursos:

- <http://material.curso-r.com/scrape/>: Material organizado pelo pessoal do Curso-R sobre webscraping (com mais exemplos).
- <https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/>: Tutorial do `Rvest` (pode estar desatualizado).

HTML

Inevitavelmente, vocês deverão ter alguma idéia de HTML (pelo menos como funciona). Em geral, páginas html são texto estruturado, interpretado pelo navegador. Veja exemplos em:

https://www.w3schools.com/html/html_basic.asp

```
<!DOCTYPE html>  
<html>  
<body>  
  
<h1>My First Heading</h1>  
<p>My first paragraph.</p>  
  
</body>  
</html>
```

Note: "body" na linguagem do rvest é um node. Dentro desse node, há um node tipo h1 (cabecalho) e outro tipo p (parágrafo). Nodes interessantes incluem table, a (links) e img.

Exemplo: wikipedia

A *wikipedia* é particularmente interessante para scraping, pois ela possui muitas páginas com listas, de onde podemos começar nossas buscas. Por exemplo,

https://en.wikipedia.org/wiki/List_of_statisticians

Podemos estar interessados em compilar uma lista com nome, *alma mater*, data de nascimento (e local), e data de falecimento (caso já tenha falecido) de estatísticos famosos.

Lista de Estatísticos

The image shows a browser window displaying the Wikipedia page for 'List of statisticians'. The browser's address bar shows the URL 'https://en.wikipedia.org/wiki/List_of_statisticians'. The page title is 'List of statisticians'. Below the title, there is a sub-header 'From Wikipedia, the free encyclopedia'. The main content area contains a paragraph: 'This list of **statisticians** lists people who have made notable contributions to the theories or application of **statistics**, or to the related fields of probability or machine learning. Also included are **actuaries** and **demographers**.' Below this paragraph is a box containing the text: 'Contents: A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · Y · Z · See also · External links'. To the right of the main content is a sidebar with a 'Statistics' section, featuring a bar chart and a bell curve, and a list of links: 'Outline · Statisticians · Glossary · Notation · Journals · Lists of topics · Articles · Portal · Category'. The left sidebar of the Wikipedia page contains various navigation links such as 'Main page', 'Contents', 'Featured content', 'Current events', 'Random article', 'Donate to Wikipedia', 'Wikipedia store', 'Interaction', 'Help', 'About Wikipedia', 'Community portal', 'Recent changes', 'Contact page', 'Tools', 'What links here', 'Related changes', 'Upload file', 'Special pages', 'Permanent link', 'Page information', 'Wikidata item', 'Cite this page', 'Print/export', 'Create a book', 'Download as PDF', 'Printable version', and 'Languages'. The bottom of the image shows the Windows taskbar with the system tray displaying the time as 22:46 and the date as 21/10/2018.

Página: George Box

Wikipedia - George E. P. Box

https://en.wikipedia.org/wiki/George_E._P._Box

WIKIPEDIA The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects

George E. P. Box

From Wikipedia, the free encyclopedia

For the ice hockey player, see [George Box \(ice hockey\)](#).

George Edward Pelham Box FRS^[1] (18 October 1919 – 28 March 2013) was a British statistician, who worked in the areas of [quality control](#), [time-series analysis](#), [design of experiments](#), and [Bayesian inference](#). He has been called "one of the great statistical minds of the 20th century".^{[3][4][5][6]}

Contents [hide]

- Education and early life
- Career and research
- Awards and honours
- Personal life
- References
- External links

Education and early life [edit]


He was born in [Gravesend, Kent](#), England. Upon entering university he began to study [chemistry](#), but was called up for service before finishing. During [World War II](#), he performed experiments for the [British Army](#) exposing small animals to poison gas. To analyze the results of his experiments, he taught himself statistics from available texts. After the war, he enrolled at [University College London](#) and obtained a bachelor's degree in mathematics and statistics. He received a PhD from the [University of London](#) in 1953, under the supervision of [Egon Pearson](#).^{[2][7]}

Career and research [edit]

From 1948 to 1956, Box worked as a statistician for [Imperial Chemical Industries](#) (ICI). While at ICI, he took a leave of absence for a year and served as a visiting professor at the North Carolina State University at Raleigh (now [North Carolina State University](#)). He later went to [Princeton University](#) where he served as Director of the Statistical Research Group.

In 1960, Box moved to the [University of Wisconsin–Madison](#) to create the Department of Statistics. He was appointed Vilas Research Professor of Statistics (the highest honor accorded to any faculty member at the University of Wisconsin–Madison) in 1980.^[8] Box and Bill Hunter co-founded

George Box



Born 18 October 1919
[Gravesend, Kent, England](#)

Died 28 March 2013 (aged 93)
[Madison, Wisconsin](#)

Residence United Kingdom, United States

Alma mater [University College London](#)

Known for "All models are wrong"
[Response-surface methodology](#)
[Box–Jenkins method](#)
[Box–Cox transformation](#)
[Box–Cox distribution](#)
[Box–Pierce test](#)
[Box–Behnken design](#)

22:48
21/10/2018

Tabela de Interesse

The screenshot shows a web browser window displaying the Wikipedia article for George E. P. Box. The browser's address bar shows the URL `https://en.wikipedia.org/wiki/George_E._P._Box`. The page title is "George E. P. Box". The article content includes a biographical paragraph and a table of contents. The browser's developer tools are open, showing the HTML structure of the infobox table and its CSS styles.

HTML Structure:

```
<a class="mw-jump-link" href="#p-search">Jump to search</a>
<div id="mw-content-text" class="mw-content-ltr" dir="ltr" lang="en">
  <div class="mw-parser-output">
    <div class="hatnote navigation-not-searchable" role="note"></div>
    <p class="mw-empty-elt"></p>
    <table class="infobox biography vcard" style="width:22em">
      <tbody>
        <tr></tr>
        <tr>
          <td colspan="2" style="text-align:center"></td>
        </tr>
      </tbody>
    </table>
  </div>
</div>
```

CSS Styles:

```
element {
  text-align: center;
}
.infobox td, .infobox th {
  vertical-align: top;
  text-align: left;
}
Inherited from table
.infobox {
  border-spacing: 3px;
  color: black;
}
```


SelectorGadget

Uma ferramenta recomendada pelo rvest é o chamado SelectorGadget (<https://selectorgadget.com/>), que mostra o nome de um "selector" em CSS. Há uma extensão para o navegador Chrome que permite que você use o SelectorGadget em qualquer página.

Com o selector correto, você pode acessá-lo usando `html_nodes()`. Selectors interessantes incluem `"table.<nome>"` e `"li"`. É preciso inspecionar as páginas de interesse caso a caso.

Using SelectorGadget (Chrome)

W George E. P. Box - Wikipedia x +

https://en.wikipedia.org/wiki/George_E._P._Box

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

George E. P. Box

From Wikipedia, the free encyclopedia

For the ice hockey player, see [George Box \(ice hockey\)](#).

George Edward Pelham Box FRS^[1] (18 October 1919 – 28 March 2013) was a British statistician, who worked in the areas of [quality control](#), [time-series analysis](#), [design of experiments](#), and [Bayesian inference](#). He has been called "one of the great statistical minds of the 20th century".^{[3][4][5][6]}

Contents [hide]

- [Education and early life](#)
- [Career and research](#)
- [Awards and honours](#)
- [Personal life](#)
- [References](#)
- [External links](#)

Education and early life [edit]

He was born in [Gravesend, Kent](#), England. Upon entering university he began to study [chemistry](#), but was called up for service before finishing. During [World War II](#), he performed experiments for the [British Army](#) exposing small animals to poison gas. To analyze the results of his experiments, he taught himself statistics from available texts. After the war, he enrolled at [University College London](#) and obtained a bachelor's degree in mathematics and statistics. He received a [PhD](#) from the [University of London](#) in 1953, under the supervision of [Egon Pearson](#).^{[2][7]}

Career and research [edit]

From 1948 to 1956, Box worked as a statistician for [Imperial Chemical Industries](#) (ICI). While at ICI, he took a leave of absence for a year and served as a visiting professor at the [North Carolina State University](#) at Raleigh (now [North Carolina State University](#)). He later went to [Princeton University](#) where he served as Director of the [Statistical Research Group](#).

In 1960, Box moved to the [University of Wisconsin–Madison](#) to create the [Center for Quality and Productivity Improvement of the University of Wisconsin–Madison](#) in 1984. Box officially retired in 1999, becoming an

George Box



Born 18 October 1919
[Gravesend, Kent, England](#)

Died 28 March 2013 (aged 93)
[Madison, Wisconsin](#)

Residence [United Kingdom](#), [United States](#)

Alma mater [University College London](#)

Known for "All models are wrong"
[Response-surface methodology](#)
[Box–Jenkins method](#)
[Box–Cox transformation](#)

Tabela de interesse

```
url = "https://en.wikipedia.org/wiki/George_E._P._Box"
webpage <- read_html(url)

table <- webpage %>%
  html_nodes("table.vcard") %>% # SG
  html_table(header = FALSE)

# Retorna uma lista de tabelas... eu só quero a primeira
table <- table[[1]]
```

Conteúdo

```
table %>% as.tibble
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new
## This warning is displayed once per session.
```

```
## # A tibble: 17 x 2
```

```
##   X1           X2
```

```
##   <chr>       <chr>
```

```
## 1 George Box  George Box
```

```
## 2 ""         ""
```

```
## 3 Born      (1919-10-18)18 October 1919Gravesend, Kent, England
```

```
## 4 Died      28 March 2013(2013-03-28) (aged&nbsp;93)Madison, Wisconsin
```

```
## 5 Residence United Kingdom, United States
```

```
## 6 Alma&nbsp;mater University College London
```

```
## 7 Known&nbsp;for "All models are wrong"\nResponse-surface methodol
```

```
## 8 Awards    "Shewhart Medal (1968)\nWilks Memorial Award (1972)\nR.
```

```
## 9 Scientific car... Scientific career
```

```
## 10 Fields   "Statistics\nDesign of experiments\nBayesian statistics
```

```
## 11 Institutions "ICI\nPrinceton University\nUniversity of Wisconsin-Madison
```

```
## 12 Thesis    Departures from Independence and Homoskedasticity in th
```

```
## 13 Doctoral advis... "Egon Pearson\nH.&nbsp;O. Hartley[2]"
```

```
## 14 Doctoral stude... John F. MacGregor[2]Greta M. Ljung
```

Conteúdo (limpeza com regex)

```
## # A tibble: 17 x 2
##   X1                X2
##   <chr>            <chr>
## 1 George Box       George Box
## 2 ""              ""
## 3 Born             (1919-10-18)18 October 1919Gravesend, Kent, England
## 4 Died            28 March 2013(2013-03-28) (aged 93)Madison, Wisconsin
## 5 Residence       United Kingdom, United States
## 6 Alma mater      University College London
## 7 Known for       "All models are wrong" Response-surface methodology EVO
## 8 Awards          Shewhart Medal (1968) Wilks Memorial Award (1972) R. A.
## 9 Scientific car... Scientific career
## 10 Fields          Statistics Design of experiments Bayesian statistics Ti
## 11 Institutions    ICI Princeton University University of Wisconsin-Madiso
## 12 Thesis         Departures from Independence and Homoskedasticity in th
## 13 Doctoral advis... "Egon Pearson H. O. Hartley "
## 14 Doctoral stude... John F. MacGregor Greta M. Ljung
## 15 Influences      Ronald Fisher
## 16 Influenced      Norman Draper George C. Tiao
## 17 ""            ""
```

Procurando Links

Inspecionando a página no navegador, eu descobri que dentro de body #content (o conteúdo da página) os links estão guardados no node "li".

```
url = "https://en.wikipedia.org/wiki/List_of_statisticians"  
listPages <- read_html(url)  
links <- listPages %>%  
  html_nodes("body #content") %>% # Inspect Object...  
  html_nodes("li") # All links
```

Procurando Links

links

```
## {xml_nodelist (691)}
## [1] <li><a href="/wiki/Outline_of_statistics" title="Outline of statist .
## [2] <li><a class="mw-selflink selflink">Statisticians</a></li>
## [3] <li><a href="/wiki/Glossary_of_probability_and_statistics" title="G .
## [4] <li><a href="/wiki/Notation_in_probability_and_statistics" title="N .
## [5] <li><a href="/wiki/List_of_statistics_journals" title="List of stat .
## [6] <li><a href="/wiki/Lists_of_statistics_topics" title="Lists of stat .
## [7] <li><a href="/wiki/List_of_statistics_articles" title="List of stat .
## [8] <li>\n<a href="/wiki/File:Nuvola_apps_edu_mathematics_blue-p.svg" c .
## [9] <li><a href="/wiki/Category:Statistics" title="Category:Statistics" .
## [10] <li class="nv-view"><a href="/wiki/Template:Statistics_topics_sideb .
## [11] <li class="nv-talk"><a href="/wiki/Template_talk:Statistics_topics_ .
## [12] <li class="nv-edit"><a class="external text" href="https://en.wikip .
## [13] <li><a href="#A">A</a></li>
## [14] <li><a href="#B">B</a></li>
## [15] <li><a href="#C">C</a></li>
## [16] <li><a href="#D">D</a></li>
## [17] <li><a href="#E">E</a></li>
## [18] <li><a href="#F">F</a></li>
## [19] <li><a href="#G">G</a></li>
```

"Sajid Ali Khan, Rawalakot" até "Zipf, George Kingsley"

```
estat1 = links %>%  
  as.character %>%  
  grep("Sajid Ali Khan, Rawalakot", .)  
estatN = links %>%  
  as.character %>%  
  grep("Zipf, George Kingsley", .)  
estat1
```

```
## [1] 40
```

```
estatN
```

```
## [1] 679
```

```
links <- links[estat1:estatN]
```


Páginas individuais

links

```
## {xml_nodeset (640)}
## [1] <li>\n<a href="/w/index.php?title=Sajid_Ali_Khan&action=edit&am .
## [2] <li>\n<a href="/wiki/Odd_Aalen" title="Odd Aalen">Aalen, Odd Olai</ .
## [3] <li>\n<a href="/wiki/Edith_Abbott" title="Edith Abbott">Abbott, Edi .
## [4] <li>\n<a href="/wiki/Robert_P._Abelson" class="mw-redirect" title=" .
## [5] <li>\n<a href="/wiki/Moses_Abramovitz" title="Moses Abramovitz">Abr .
## [6] <li>\n<a href="/wiki/Gottfried_Achenwall" title="Gottfried Achenwal .
## [7] <li>\n<a href="/wiki/Abraham_Manie_Adelstein" title="Abraham Manie .
## [8] <li>\n<a href="/wiki/Riaz_Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a> .
## [9] <li>\n<a href="/wiki/Beatrice_Aitchison" title="Beatrice Aitchison" .
## [10] <li>\n<a href="/wiki/John_Aitchison" title="John Aitchison">Aitchis .
## [11] <li>\n<a href="/wiki/Alexander_Aitken" title="Alexander Aitken">Ait .
## [12] <li>\n<a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="H .
## [13] <li>\n<a href="/wiki/Mir_Masoom_Ali" title="Mir Masoom Ali">Ali, Mi .
## [14] <li>\n<a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, .
## [15] <li><a href="/wiki/David_B._Allison" title="David B. Allison">Allis .
## [16] <li>\n<a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug< .
## [17] <li>\n<a href="/wiki/Takeshi_Amemiya" title="Takeshi Amemiya">Amemi .
## [18] <li>\n<a href="/wiki/Oskar_Anderson" title="Oskar Anderson">Anderso .
## [19] <li><a href="/wiki/Theodore_Wilbur_Anderson" title="Theodore Wilbur
```

Páginas individuais

```
links %>%  
  html_nodes("a")
```

```
## {xml_nodeset (640)}  
## [1] <a href="/w/index.php?title=Sajid_Ali_Khan&action=edit&redl .  
## [2] <a href="/wiki/Odd_Aalen" title="Odd Aalen">Aalen, Odd Olai</a>  
## [3] <a href="/wiki/Edith_Abbott" title="Edith Abbott">Abbott, Edith</a>  
## [4] <a href="/wiki/Robert_P._Abelson" class="mw-redirect" title="Robert .  
## [5] <a href="/wiki/Moses_Abramovitz" title="Moses Abramovitz">Abramovit .  
## [6] <a href="/wiki/Gottfried_Achenwall" title="Gottfried Achenwall">Ach .  
## [7] <a href="/wiki/Abraham_Manie_Adelstein" title="Abraham Manie Adelst .  
## [8] <a href="/wiki/Riaz_Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a>  
## [9] <a href="/wiki/Beatrice_Aitchison" title="Beatrice Aitchison">Aitch .  
## [10] <a href="/wiki/John_Aitchison" title="John Aitchison">Aitchison, Jo .  
## [11] <a href="/wiki/Alexander_Aitken" title="Alexander Aitken">Aitken, A .  
## [12] <a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="Hirotsu .  
## [13] <a href="/wiki/Mir_Masoom_Ali" title="Mir Masoom Ali">Ali, Mir Maso .  
## [14] <a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, R. G. .  
## [15] <a href="/wiki/David_B._Allison" title="David B. Allison">Allison, .  
## [16] <a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug</a>  
## [17] <a href="/wiki/Takeshi_Amemiya" title="Takeshi Amemiya">Amemiya, Ta .  
## [18] <a href="/wiki/Oskar_Anderson" title="Oskar Anderson">Anderson, Osk .
```

Páginas Individuais

```
links %>%  
  html_nodes("a") %>%  
  html_attr("href") # Salvar title também!
```

```
## [1] "/w/index.php?title=Sajid_Ali_Khan&action=edit&redlink=1"  
## [2] "/wiki/Odd_Aalen"  
## [3] "/wiki/Edith_Abbott"  
## [4] "/wiki/Robert_P._Abelson"  
## [5] "/wiki/Moses_Abramovitz"  
## [6] "/wiki/Gottfried_Achenwall"  
## [7] "/wiki/Abraham_Manie_Adelstein"  
## [8] "/wiki/Riaz_Ahsan"  
## [9] "/wiki/Beatrice_Aitchison"  
## [10] "/wiki/John_Aitchison"  
## [11] "/wiki/Alexander_Aitken"  
## [12] "/wiki/Hirotsugu_Akaike"  
## [13] "/wiki/Mir_Masoom_Ali"  
## [14] "/wiki/R._G._D._Allen"  
## [15] "/wiki/David_B._Allison"  
## [16] "/wiki/Doug_Altman"  
## [17] "/wiki/Takeshi_Amemiya"  
## [18] "/wiki/Oskar_Anderson"
```

DB

A melhor maneira de armazenar o conteúdo das páginas é através de um banco de dados.

```
li <- links %>% html_nodes("a") %>% html_attr("href")
li <- paste0("https://en.wikipedia.org", li)
names <- links %>% html_nodes("a") %>% html_attr("title")
db = dbConnect(SQLite(), "estatisticos.db")
bad = c("page does not exist", "Florence Nightingale", "Harold Wilson")
bad1 = unlist(sapply(bad, grep, names))
bad2 = unlist(sapply(c("mshkhan", "redlink", "orghhttp"), grep, li))
names = names[-c(bad1, bad2)]
li = li[-c(bad1, bad2)]
dbWriteTable(db, "person", data.frame(id = seq_along(names),
                                       names = names,
                                       links = li))

dbExecute(db, "CREATE TABLE info
              (id INTEGER, Born TEXT, Died TEXT,
               AlmaMater TEXT)")
```

```
## [1] 0
```

Conferindo...

```
dbGetQuery(db, "SELECT * FROM person LIMIT 4")
```

##	id	names	links
## 1	1	Odd Aalen	https://en.wikipedia.org/wiki/Odd_Aalen
## 2	2	Edith Abbott	https://en.wikipedia.org/wiki/Edith_Abbott
## 3	3	Robert P. Abelson	https://en.wikipedia.org/wiki/Robert_P._Abelson
## 4	4	Moses Abramovitz	https://en.wikipedia.org/wiki/Moses_Abramovitz

Extraindo tabelas (demora alguns minutos...)

```
library(doMC) ## se windows library(doParallel)

## Loading required package: foreach

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##   accumulate, when

## Loading required package: iterators

## Loading required package: parallel

registerDoMC(4) ## se windows registerDoParallel(nproc)
f <- function(x) if(length(x) == 0) return(NA_character_) else return
out = foreach(i=seq_along(li), .combine=rbind) %dopar% {
  webpage <- read_html(li[i])

  table <- webpage %>% html_nodes("table.vcard") %>% html_table(heac

  if(length(table) == 0) return(NULL)
```

Consultando nossa tabela

O código quebrou com *Florence Nightingale* e *Harold Wilson*, porque tinham muitos campos diferentes do esperado, ficando inconsistente com as tabelas anteriores. Estes podem ser trabalhados manualmente. Fora isso, o código de scraping funciona bem:

```
dbGetQuery(db, "SELECT names, Born, Died, AlmaMater FROM info
              INNER JOIN person ON info.id = person.id
              LIMIT 8") %>% as.tibble
```

```
## # A tibble: 8 x 4
##   names      Born      Died      AlmaMater
##   <chr>      <chr>      <chr>      <chr>
## 1 Odd Aalen (1947-05-06) May ... <NA>      University of Oslo
## 2 Edith Ab... (1876-09-26) Septe... July 28, 1957(195... <NA>
## 3 Moses Ab... (1912-01-01) Janua... December 1, 2000(... Harvard University and C
## 4 Gottfrie... (1719-10-20) 20 Oc... 1 May 1772(1772-0... <NA>
## 5 Riaz Ahs... 1951, December 25... November 8, 2008(... University of KarachiAda
## 6 Beatrice... (1908-07-18) 18 Ju... 22 September 1997... Goucher College (BA) Johr
## 7 John Ait... 22 July 1926 East ... 23 December 2016 ... University of Edinburgh
## 8 Alexande... (1895-04-01) 1 Apr... 3 November 1967(1... University of EdinburghU
```

Crawlers + Cron

- *Crawlers* são programas de scraping que interpretam conexões nas páginas e continuam buscando novas páginas. Estão bem além do escopo do nosso curso.
- `cron` é um programa de Linux que permite executar outros programas em intervalos fixos de tempo. Existe um pacote, `cronR` (<https://cran.r-project.org/web/packages/cronR/README.html>), que permite executar scripts de R em períodos fixos de tempo. Em conjunto com uma base de dados, é possível coletar dados periodicamente (por exemplo, com o `tweetR` ou em páginas de notícias).

Fechando base de dados...

```
dbDisconnect(db)
```