Webscraping em R: uma abordagem para investigação em ciências sociais*

Web Scraping in R: An Approach to Social Science Research

Web Scraping en R: una aproximación en la investigación en ciencias sociales

Recebido em 17-08-2021
Modificado em 19-10-2021
Aceito para publicação em 20-11-2021

https://doi.org/10.47456/simbitica.v8i4.37351

Quemuel Baruque de Freitas Rodrigues

Mestrando em Ciência Política pelo PPGCP da UFPE e graduado em Ciências Sociais pela UFAL. Tem experiência profissional com análise e avaliação de políticas públicas, criação e gerenciamento de banco de dados e desenvolvimento de aplicativos. E-mail: quemuelbaruque@gmail.com

Mayres Lane Pequeno dos Santos Silva

Graduanda em Ciências Sociais pela UFAL e pesquisadora nas temáticas de partidos políticos, reformas eleitorais, políticas públicas e metodologias quantitativas. E-mail: mayrespequeno@gmail.com

Marina Félix de Melo

Professora e vice-diretora do Instituto de Ciências Sociais da Universidade Federal de Alagoas. Doutora em Sociologia pelo PPGS da UFPE, coordena o Grupo de Pesquisa do CNPq "Produção Acadêmica, Científica e Tecnológica". E-mail: melomarina@msn.com

Amurabi Oliveira

Bolsista de Produtividade em Pesquisa do CNPq - Nível 2. Doutor em Sociologia pelo PPGS da UFPE com estágio pós-doutoral em Didática das Ciências Sociais pela Universidade Autônoma de Barcelona (2020). Professor da Universidade Federal de Santa Catarina. E-mail: amurabi_cs@hotmail.com

* Trabalho desenvolvido pela parceria de investigadores da Universidade Federal de Alagoas e da Universidade Federal de Santa Catarina no Grupo de Pesquisa do CNPq "Produção Acadêmica, Científica e Tecnológica".

Resumo

Este artigo apresenta uma breve introdução ao uso de algoritmos para coleta de dados em repositórios online em investigações no campo das ciências sociais a partir de um modelo empírico de pesquisa sobre produção acadêmica de bolsistas de produtividade do CNPq no Brasil. Metodologicamente, apresentamos uma pesquisa realizada com a aplicação da técnica computacional de elaboração de algoritmos e, em seguida, descrevemos o passo-a-passo do planejamento de algoritmo de *scrapping* a partir do *software* R. Busca-se não apenas tornar mais compreensível a técnica computacional para recolha de dados, como também fomentar sua utilização no campo das ciências sociais, tornando as coletas de dados em repositórios institucionais mais sistemáticas, transparentes, replicáveis e céleres.

Palavras-chave: ciências sociais; produção acadêmica; coleta de dados quantitativos; algoritmos.

Introdução

Imaginemos que uma organização internacional estivesse interessada em mapear todas as políticas públicas implementadas que promovessem a inserção da pessoa com deficiência no mercado de trabalho. Para tanto, esta empresa abriria edital para contratação de um grupo de pesquisa que pudesse produzir esse relatório que mapearia as políticas públicas voltadas para a inserção de pessoas com deficiência no mercado de trabalho. Sabendo que temos 5.568 municípios, 26 estados e 1 distrito federal, e também as políticas nacionais, qual a melhor forma de fazer essa pesquisa? Checar cada página de cada um desses entes elencados acima poderia dispender grande tempo da equipe, como incorrer em erros de recolha típicos aos limites das pesquisas que envolvem grande quantidade de dados. Uma alternativa a esta questão, e que propomos como técnica de recolha de informações, seria desenvolver um algoritmo que desempenhasse esse trabalho, envolvendo menos tempo dispendido e que permitisse replicabilidade e transparência sobre os coletados.

Tornou-se lugar-comum afirmar que o chamado novo movimento teórico (Alexander, 1987) representou não apenas uma guinada teórica, como também metodológica no campo das ciências sociais. Do mesmo modo que a dualidade micro/macro e agência/estrutura tem sido questionada, advogando-se uma perspectiva de síntese, como bem refletem a obra de autores contemporâneos como Pierre Bourdieu e Anthony Giddens, também a dualidade entre quanti/quali tem sido questionada, propondo-se em seu lugar abordagens metodológicas de cunho quali-quantitativas. Todavia, podemos observar que há ainda alguns obstáculos significativos para a consolidação dessa perspectiva.

No caso brasileiro existiu um entrincheiramento observado no campo das ciências sociais. Em outros termos, uma bipartição entre aqueles que se recusavam a qualquer

procedimento quantitativo de informações e os que tentavam enquadrar como pesquisa não científica toda e qualquer pesquisa que não fosse quantitativa (Cano, 2012; Soares, 2005). Isso foi reafirmado por Neiva (2015) quando avaliou aspectos metodológicos de 22 revistas científicas na área das ciências sociais. A ciência política mostrou que utiliza estatística avançada em cerca de 14% dos seus artigos, percentual mais elevado que nas demais áreas das ciências sociais. Quanto à utilização de metodologia quantitativa, a ciência política a apresenta em cerca de 43,8% dos seus artigos, a sociologia 20,7% e a antropologia 13,2%. Na perspectiva quantitativa, Figueiredo Filho *et al*. (2011) explica que a falta de domínio das técnicas de estatística descritiva e inferencial por grande parte dos pesquisadores é o que resulta em sua baixa aplicação nos trabalhos acadêmicos.

A produção científica está completamente ligada à sistematização do conhecimento. Kuhn (1962) propõe que o acúmulo de conhecimento, somado a novos problemas sem solução, produzem uma ciência dita melhor. De acordo com Wolf (1986), é imperativo que estudos tenham procedimentos nítidos e confiáveis, todavia o processo de coleta de informações em grande volume e com um alto rigor metodológico se propõe a ser um processo demorado, no que o longo tempo destinado à etapa de coleta parece justificável. Como Paranhos *et al*. (2013) afirma, a ausência de critérios sistemáticos gera efeitos perversos em investigações científicas. Minayo e Sanches (1993) já apontavam que um bom método sempre é aquele que, permitindo uma construção adequada dos dados, nos pode auxiliar no processo de inferência ou construção de uma explicação tangível sobre um fenômeno dentro da sociedade.

Como então produzir resultados robustos, com dados coletados com alto rigor metodológico, ao ponto de serem medidas confiáveis e replicáveis? Exemplificamos nesse artigo, seguindo tanto as orientações de Minayo e Sanches (1993), como as de Wolf (1986), para oferecer como alternativa o uso de ferramentas computacionais e como essas ferramentas podem auxiliar no processo de investigação e replicação em problemas típicos das ciências sociais. O avanço neste debate, bem como no refinamento das ferramentas metodológicas, nos possibilita avançarmos para além de uma dimensão meramente descritiva, como amiúde encontramos nos trabalhos que se propõem a analisar a produção acadêmica e o perfil dos pesquisadores em determinada área.

Durante o último meio século, para inferir algo sobre as populações humanas era necessário realizar pesquisas em eventos específicos. Hoje, na era do *big data*,[1] os mecanismos de coleta de dados existentes estão sendo usados e melhorados para responder a questões complexas dentro das relações humanas. Não é controlável que todos os que produzem dados os

[1] Grande volume de dados de informações armazenadas em diferentes mídias ou bases de dados, que podem ser de grande importância, principalmente, na tomada de decisões (Camargo-Vega *et al.*, 2015).

disponibilizem de maneira cessível e sistemática (King, 1995). Surge assim, a importância de um movimento dentro da academia científica que rearranja e utiliza métodos para coletar informações e tornar acessível o compartilhamento de dados. Existe um desafio atual para cientistas sociais, enquanto pesquisadores, de difundir novas técnicas para uma comunidade ampla que se prepara para a coleta e sistematização de novas fontes de dados.

Este artigo está subdivido em 2 partes. A primeira compreende a leitura de um modelo prático de uma pesquisa desenvolvida fazendo uso dessas estratégias computacionais. A segunda demonstra como entender os elementos de uma página da web e como aplicar um *webscraping* apresentando, de uma maneira não exaustiva, como operacionalizar um *webscraping* em diversos campos da *web*. Nesta segunda parte também demonstramos como construímos o algoritmo de coleta e sumarização da pesquisa apresentada na primeira parte deste artigo.

Métodos e técnicas de pesquisa em um estudo sobre a produção científica de bolsistas de produtividade do CNPq no Brasil

O estudo escolhido como demonstração do uso das tecnologias computacionais a problemáticas marcadas no campo das ciências sociais, que será apresentado nesta sessão, investigou diferentes formas de produções acadêmica e científica de bolsistas de produtividade em pesquisa do CNPq no Brasil. Analisamos 28 variáveis que respondiam a respeito da produção acadêmica dos bolsistas PQ no País nos últimos 5 anos, divididos em 3 grandes áreas do CNPq, a verificar possíveis diferenças e continuidades no modelo de produção dos pesquisadores de diversas áreas do conhecimento. Através de uma análise de regressão linear de mínimos quadrados ordinários realizada pelo *software Statistical Package for the Social Sciences* - SPSS, conseguimos identificar que variáveis tendem a influenciar o índice de produção acadêmica do grupo pesquisado. Utilizamos uma amostra probabilística, de tipo aleatória simples, para uma população de 13 mil PQs. Consideramos um intervalo de confiança de 3,91, com um nível de confiança de 95%. O cálculo amostral resultou em uma amostra de 601 bolsistas PQ. A coleta dos 601 casos foi selecionada pela função *sample()* do *software* R, de que falaremos em detalhamento no próximo tópico deste artigo. O desenho da pesquisa foi de tipo interseccional/corte transversal e a recolha dos dados ocorreu de janeiro a abril de 2020.

Os dados foram coletados a partir da plataforma *lattes*, que constitui uma principal base de dados acerca da formação e atuação dos pesquisadores brasileiros. Importante salientar que os dados desta plataforma são alimentados pelos próprios pesquisadores, de modo que eventuais

desatualizações, equívocos ou ausência de informações nos currículos relacionam-se com a forma como os próprios pesquisadores preenchem tais informações.

Posto que a produção acadêmica é um aspecto demasiadamente complexo, optamos pela criação de um "índice de produção geral", nossa primeira variável dependente. Este considera diversas variáveis que dizem respeito à produção acadêmica dos bolsistas PQ[2] através de uma média, *mean*, no comando do *software* SPSS. Para a composição desta variável dependente, realizamos o teste de Confiabilidade de *Crombach's Alpha* para verificar quais variáveis contribuíram para a formação do índice. Após vários testes com diversas variáveis, encontramos 0,75 como valor mais alto para o coeficiente de *Crombabh's Alpha* com 13 variáveis que diziam respeito à produção acadêmica.

O índice de produção acadêmica geral dos PQs brasileiros foi então formado com as seguintes variáveis: 1. Quantidade de apresentação de trabalhos em congressos, palestras, seminários etc; 2. Quantidade de textos apresentados em congressos, palestras, seminários etc; 3. Quantidade de livros publicados desde 2015; 4. Quantidade total de artigos publicados em Revistas desde 2015; 5. Quantidade total de artigos publicados avaliados pelo Webqualis da Capes; 6. Total de artigos publicados de Estrato Superior avaliados pelo Webqualis da Capes[3]; 7. Quantidade de publicações midiáticas não especializadas (TV, jornais etc); 8. Total de orientações concluídas desde 2015; 9. Quantidade de participações em bancas de doutorado desde 2015; 10. Quantidade de participações em bancas de mestrado desde 2015; 11. Quantidade de participações em bancas de graduação desde 2015; 12. Quantidade de participações em bancas de especialização desde 2015 e; 13. Quantidade de participação em eventos desde 2015[4]. As medidas de tendência centrais básicas do índice criado a partir das 13 variáveis e com 601 casos são: média de produção acadêmica geral/LG10 (0,919); desvio padrão (0,29) e; amplitude (de -0,20 a 1,92).

Como parte integrativa da análise utilizamos a comparação do índice geral com a variável que mensura apenas a quantidade de artigos publicados em estratos superiores do webqualis Capes[5], nossa segunda variável dependente. A produção de artigos avaliados como

[2] Os nomes dos sujeitos pesquisados, bem como qualquer forma de identificação dos bolsistas PQ, foi retirada da análise por razões éticas de pesquisa.

[3] São consideradas revistas de qualis superior pelo webqualis aquelas que recebem as seguintes notas: A1, A2 e B1. São consideradas revistas de qualis inferior pelo webqualis aquelas que recebem as notas: B2, B3, B4, B5 e C. Disponível em https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf - Acesso em 01-05-2021.

[4] Além destas 13 variáveis utilizadas para composição do índice, as demais variáveis analisadas neste estudo foram: sexo; tipo de bolsa; grande área do CNPq; sub área do CNPq; região; unidade federativa; universidade; principal país de internacionalização; área Capes do PPG (Programa de Pós-Graduação) ao qual está primordialmente vinculado; notas Capes do Programa ao qual está vinculado; quanto tempo demorou para fazer o doutorado; tipo de emprego; ano de informação mais antiga do lattes e; Observações Gerais.

[5] Consideramos as áreas Capes específicas às quais fazem parte cada um dos bolsistas para verificação das notas

superiores pela Capes (A1, A2 e B1) tem sido uma variável importante para o entendimento das dinâmicas de produção na academia brasileira atualmente por estar estritamente vinculada à avaliação dos Programas de Pós-Graduação aos quais se inserem os bolsistas. Importante salientar que a variável sobre a quantidade de publicações em artigos de estrato superior Capes foi utilizada a partir de uma correção logarítmica, feita através do comando LG10, devido à falta de simetria na distribuição original, que impediria a realização de testes paramétricos. Sobre os *outliers*[6] encontrados na distribuição, nomeadamente àqueles derivados do índice criado, optamos por não ocultá-los da base de dados justamente para compararmos os testes com e sem casos isolados, utilizando estes *outliers* para o entendimento de condições diferenciadas entre os escores da distribuição.

A partir das 28 variáveis do banco de dados realizamos algumas descrições iniciais e caras ao entendimento de nossa problemática, seguidas da aplicação de testes paramétricos e não paramétricos de análise, sendo a maioria destes testes de diferença entre médias e testes de correlação bivariada. Embora a variável dependente principal de investigação deste estudo seja o índice de produção geral, utilizamos, como já posto, a variável qualis superior também como variável dependente. Por toda a análise conseguinte, fizemos uma comparação entre os fatores determinantes para variação destas variáveis de forma paralela. Por fim, apresentamos um modelo de análise de regressão multivariada para cada qual destas duas variáveis de modo a explicarmos que aspectos influenciam a produção acadêmica deste público em dois modelos de produção: geral (índice) e a partir da publicação de qualis superior.

A amostra de nossa investigação, que contou com a análise de 601 currículos lattes recolhidos de forma aleatória - probabilística, em que todos os indivíduos do universo têm as mesmas chances de comporem a amostra, analisou 222 mulheres (37%) e 379 homens (63%)[7]. Através de um teste paramétrico de diferença entre duas médias, Teste *T de Student*, verificamos a diferença entre as médias de homens e mulheres no que se refere ao índice de produtividade criado para este trabalho. A média de produtividade de mulheres foi de 0,90, e a de homens, de 0,93. O valor do teste de significância correspondente foi de p=0,28[8], indicando que não existe diferença estatisticamente significativa entre as médias de produção de homens e mulheres. Pelo

qualis das revistas no webqualis da Capes. Isto é, um artigo publicado por um bolsista X pode considerar uma nota no webqualis diferente da atribuída a um bolsista Y que tenha publicação na mesma revista, caso estes pertençam a áreas distintas de conhecimento para a Capes. Atenta-se para o fato de que o qualis encontra-se atualmente em alteração, de modo que o qualis por área será substituído por um qualis unificado, entretanto, consideramos o qualis que estava em vigor no momento da pesquisa, que foi o mesmo utilizado para a avaliação da produção dos bolsistas no momento de solicitação de suas bolsas de produtividade.

6. Valor atípico, muito distante da média, sendo um escore (um dado) afastado dos demais na distribuição descritiva da série tratada em uma variável de análise.

7 Os resultados preliminares desta nossa pesquisa estão disponíveis em Autores (2020).

8 Neste estudo trabalhamos com um nível de significância de 95%, admitindo como significantes testes em que o p valor não ultrapassem 0,05.

mesmo teste, verificamos as médias de publicação em artigos de qualis superior de homens e mulheres. Elas possuem média de 0,87 e, eles, de 0,92. Embora a média do grupo masculino apresente-se numericamente como mais elevada, não existe diferença estatisticamente significativa (p=0,22) entre as médias de publicação em revistas de qualis superior no Webqualis Capes de homens e mulheres.

Ao que se refere às notas Capes de PPGs aos quais se vinculam homens e mulheres bolsistas PQ, de acordo com o teste diferença entre médias realizado, Teste *T de Student*, não existe diferença estatisticamente significativa entre as médias das notas Capes dos programas aos quais estão vinculados homens e mulheres (p=0,92). A média das notas Capes dos PPGs aos quais elas estão vinculadas é de 5,29, e aos quais eles estão vinculados, de 5,28. Investigamos na sequência a localização de homens e mulheres em diferentes áreas do conhecimento a partir de duas variáveis: 1. Subárea CNPq à qual vinculam suas bolsas PQ e 2. Área Capes à qual corresponde o PPG[9] ao que está primordialmente vinculado este bolsista PQ. Existe uma associação moderada entre o sexo dos bolsistas e a Subárea do CNPq em que estão inscritos. O teste de associação bivariada realizado foi o não-paramétrico *V de Crámer*, pelo qual encontramos um coeficiente de associação de 0,45, estatisticamente significativo (p=0,00).

A considerar o total de frequências da categoria do sexo feminino na amostra, as áreas que apresentam maiores percentuais de mulheres PQ são: Linguística 6,3%; Educação: 5,9%; História 5% e; Química 5%. As áreas sem nenhuma representação feminina são: Astronomia; Biofísica; Bioquímica; Engenharia Aeroespacial; Engenharia de Transportes; Probabilidade e Estatística e; Zootecnia. Já ao analisarmos o total de frequências da categoria do sexo masculino na amostra, as áreas que apresentam maiores percentuais de homens PQ são: Física 12,9 %; Agronomia 4,5%; Química 4,5%; Engenharia Mecânica 3,7% e; Medicina 3,7%. Áreas sem nenhuma representação masculina são: Artes; Ciência da Informação; Comunicação e; Fonoaudiologia. De forma semelhante, há uma associação moderada entre o sexo dos bolsistas e a Área Capes à qual corresponde o PPG ao qual este está vinculado. Mais uma vez, o teste realizado foi o *V de Crámer*, em que verificamos um coeficiente de correlação de 0,40, estatisticamente significativo (p=0,00).

As produções acadêmicas dos bolsistas PQ também se distribuem por diferentes regiões do Brasil e instituições. O entendimento sobre a localização desta produção é relevante na medida em que este acompanha a necessidade de compreensão de possíveis assimetrias internas ao País no que se refere à distribuição de produção acadêmica qualificada.

[9] Programa de Pós-Graduação da Capes.

Aplicamos um teste de análise de variância (*Anova*), teste de diferença entre médias estatísticas, para verificarmos o impacto da produção acadêmica, do índice de produção criado, de acordo com as regiões do País. Não existe diferença estatisticamente significativa entre as médias de produção das diferentes regiões, ou seja, embora estas médias sejam diferentes umas das outras, não implica que, na prática, tenhamos uma região que seja considerada como mais ou menos produtiva que as demais. Foram encontradas as seguintes médias de produtividade por região: Norte, com média de 1,06 (8 casos); Centro-Oeste 1,01 (24 casos); Sul 0,97 (127 casos); Nordeste 0,89 (95 casos) e; Sudeste 0,89 (346 casos). Nota-se a grande diferença de número de casos pra cada grupo, o que tende a indicar a falta de significância estatística para demonstrarmos reais diferenças entre as médias através de uma análise de variância. Em termos brutos, vemos como centro-oeste e norte têm médias mais altas em relação ao restante do País mesmo com números reduzidos de casos. Este dado salta à hipótese de que, para um pesquisador se tornar PQ nestas regiões, é preciso destacar-se acima de um quadro de produção nacional, especialmente por serem regiões com número menor de universidades às quais podem estar vinculadas. Seguimos a mesma lógica analítica com a variável que responde à quantidade de publicações em revistas de estrato superior no webqualis da Capes. Não foi encontrada diferença estatisticamente significativa (*Anova*, p=0,11) entre as médias de publicação de artigos com qualis superior no webqualis da Capes entre as diferentes regiões. As médias de publicação em qualis superior Capes, por regiões, são: Centro-Oeste 1,05; Norte 1,00; Sul 0,96; Sudeste 0,88 e; Nordeste 0,84. Após a realização de diversos testes de associação não-paramétricos, verificamos que não existe nenhuma associação entre as variáveis região ou Unidade Federativa com o nível da bolsa recebida pelo PQ, tampouco com a grande área do CNPq a que pertence, ou subárea, ou mesmo qualquer relação entre região e Unidade Federativa no que diz respeito à área Capes em que está inserido o PPG de atuação primordial de cada bolsista PQ analisado.

Depois de consideradas as análises bivariadas já descritas acima e verificadas as médias de nosso índice de produção acadêmico sob diversos aspectos da realidade, verificamos quais variáveis de nosso banco de dados melhor ajudam na compreensão da variação da variável dependente em tela, do índice de produção criado. Utilizamos a técnica de análise de regressão linear múltipla que, nada mais é, do que um conjunto de técnicas estatísticas para a construção de modelos que descrevem relações entre diversas variáveis previsoras, independentes, face à variação da variável dependente, no caso, o índice de produção.

Pelo método *stepwise*, em que são inseridas as variáveis “uma a uma” no modelo, inserimos 17 variáveis escalares na feitura do teste. Destas 17, 10 compuseram o modelo:

1. Artigos publicados em revistas desde 2015
2. Participações em eventos
3. Orientações concluídas desde 2015
4. Textos apresentados em congressos, seminários etc
5. Participações em bancas de mestrado
6. Publicações em revistas de qualis superior na área de atuação (LG10)
7. Livros publicados desde 2015
8. Publicações midiáticas não especializadas (TV, jornais etc)
9. Participações em bancas de doutorado
10. Participações em bancas de graduação

O Coeficiente de Determinação de R^2 Ajustado encontrado para este modelo foi de 0,89, o que significa dizer que estas 10 variáveis, na ordem apresentada, explicam a variância em 89% da produção acadêmica geral dos bolsistas de produtividade em sociologia do CNPq de acordo com o índice criado[10]. Importante reportarmos algumas informações técnicas utilizadas para esta análise de regressão: o teste de *Durbin Watson*, que verifica a homocedasticidade da análise foi de 1,95, o que nos leva a entender que o modelo não tem problemas de heterocedasticidade. O modelo todo tem significância estatística, aferido através de um teste ANOVA. Posto o resultado do teste de tolerância, com resultados todos inferiores a 0,1, diagnosticamos que não há colinearidade ou multicolinearidade no modelo, ou seja, estas variáveis não são "idênticas" e marcadas por nomenclaturas distintas. São genuinamente aspectos distintos da realidade analisada. O teste VIF corrobora na assertiva de que não encontramos casos de colinearidade, posto que os resultados alcançados foram todos inferiores a 10.

Encontramos poucos casos de *outliers*, apenas 9 escores. Estes não foram retirados da análise porque não se apresentaram muito distantes de 3 desvios padrões, conforme pré-estabelecido na solicitação do teste. Segue abaixo a representação pictórica da reta de regressão:

[10] A explicação para a obtenção de um valor de R^2 ajustado tão elevado está no princípio da endogeneidade, isto é, o efeito (a variável dependente) é medido de modo a praticamente coincidir com as suas causas. Porém, o recurso utilizado da análise de regressão nos permitiu ordenar o impacto de cada variável de causa na construção do próprio índice. Embora o modelo possa ser suposto como uma tautologia por esta análise, o que importa em termos explicativos é a capacidade de especificação de quais variáveis de fato importam e explicam a movimentação da variável dependente e não propriamente o valor de "prova" do R^2 ajustado.

Gráfico 1 - Reta de Regressão para variável "Índice de Produção (LG10)"

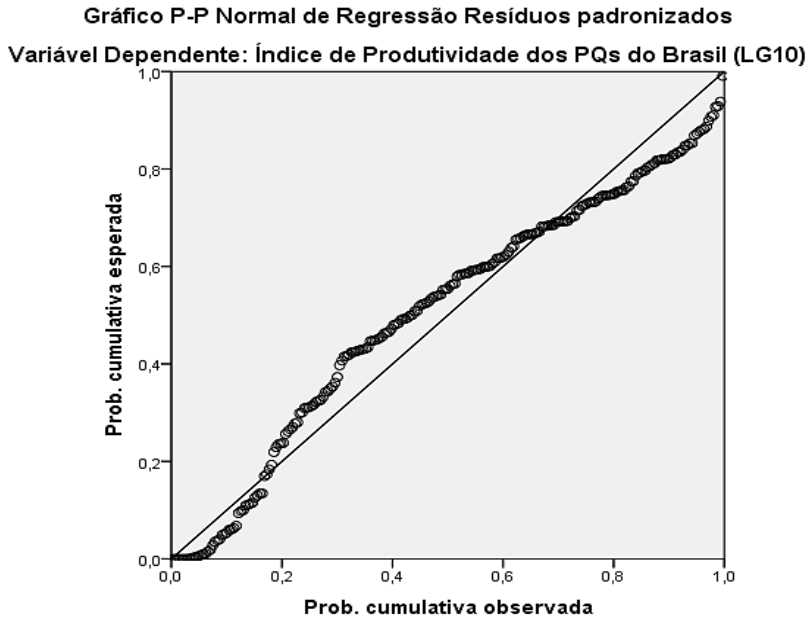
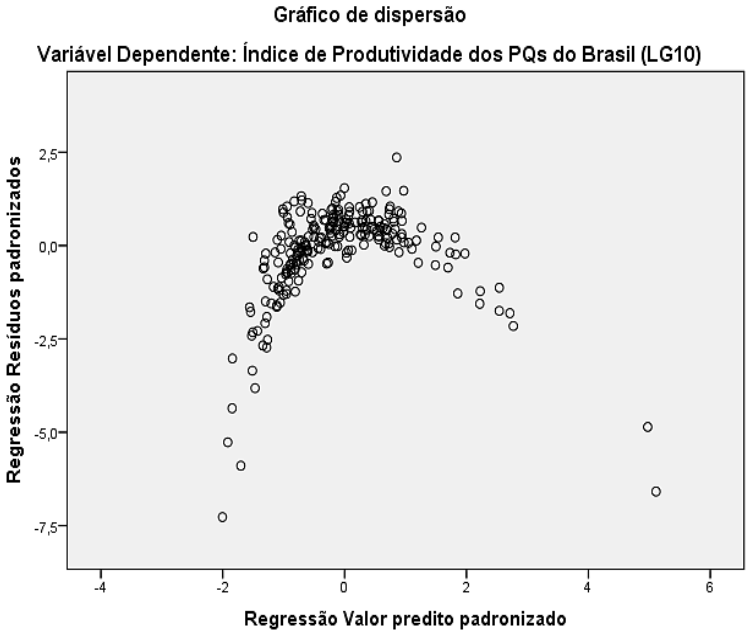


Gráfico 2 - Dispersão para variável "Índice de Produção (LG10)"



Procedemos na sequência com a aplicação da análise de regressão linear múltipla para a variável que mensura a quantidade de artigos publicados pelos bolsistas de produtividade em revistas de estrato superior da Capes. Novamente por *stepwise*, inserimos as mesmas 17 variáveis, invertendo apenas as posições das variáveis índice de produção acadêmica geral e

qualis superior, no que a primeira passou a figurar como variável previsora e não mais como variável de saída.

Ao verificarmos o teste de tolerância da análise, diagnosticamos que havia colinearidade no modelo entre as variáveis "Artigos avaliados pelo Qualis Capes" e "Artigos publicados em revistas desde 2015", posto que os valores de tolerância encontrados foram superiores a 0,1. A colinearidade indica que duas variáveis estão tão associadas que se reduzem à mesma informação prática. O teste VIF corroborou o problema da colinearidade, apresentando resultados superiores a 10 para estas duas variáveis. Optamos por refazer o teste sem a variável "Artigos publicados em revistas desde 2015", presumindo que os PQs da amostra, logo, tendem fortemente a publicações em revistas que são avaliadas pelo qualis da Capes. Assim, resolvemos o problema da colinearidade inicialmente verificado.

Refizemos a análise de regressão linear múltipla após esta correção, com as 16 variáveis previsoras e sem colinearidade. Destas 16, 6 compuseram o modelo:

1. Artigos avaliados pelo Qualis Capes
2. Índice de Produtividades do PQs do Brasil (LG10)
3. Apresentação de trabalhos em congressos, palestras etc
4. Participações em bancas de doutorado
5. Livros publicados desde 2015
6. Textos apresentados em congressos, seminários etc

O Coeficiente de Determinação de R^2 Ajustado encontrado para este modelo foi de 0,65, o que significa dizer que estas 4 variáveis, na ordem apresentada, explicam 65% da variância de publicação de artigos de qualis superior da Capes entre os bolsistas de produtividade. Evidente salientar que este resultado é restrito às 16 variáveis propostas ao modelo e que este resultado supostamente se alteraria com a introdução de outras variáveis.

Assim como no modelo anterior, o teste de *Durbin Watson* (1,83) conferiu que não há problemas de heterocedasticidade na análise. A análise de regressão foi realizada com significância estatística conferida através do teste ANOVA. Posto o resultado do teste de tolerância, novamente diagnosticamos que não há colinearidade ou multicolinearidade no modelo face os valores encontrados nestes resultados, inferiores a 10. Encontramos poucos casos de *outliers* e estes não foram retirados da análise porque não estavam muito distantes de 3 desvios padrões, conforme pré-estabelecido desde a solicitação do teste no *software* SPSS.

Gráfico 3 - Reta de Regressão para variável "Artigos Qualis Superior Capes (LG10)"

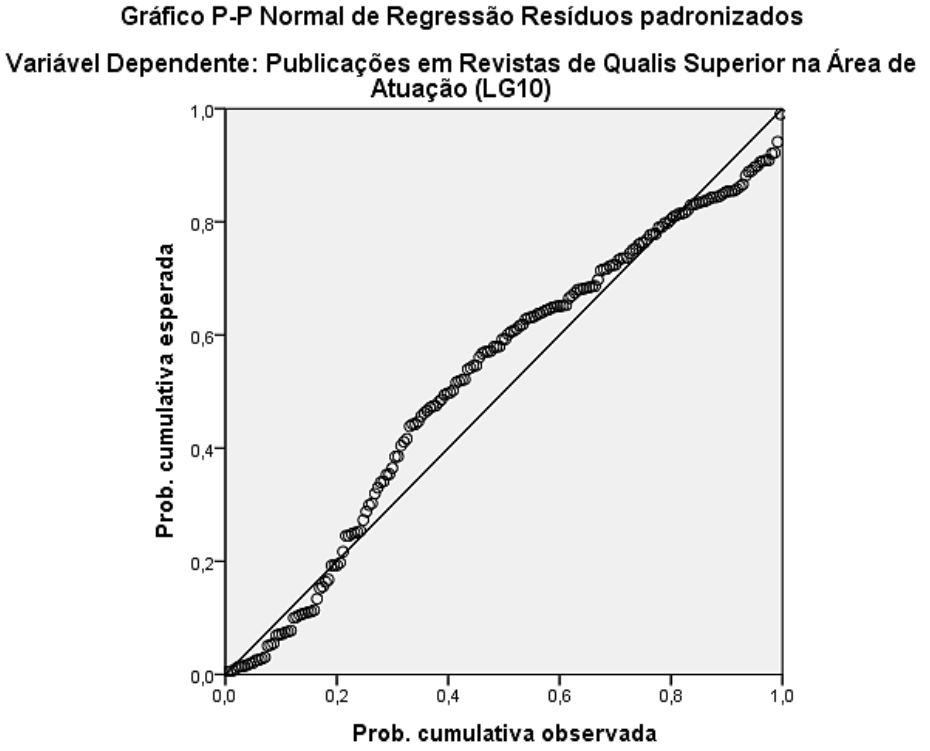
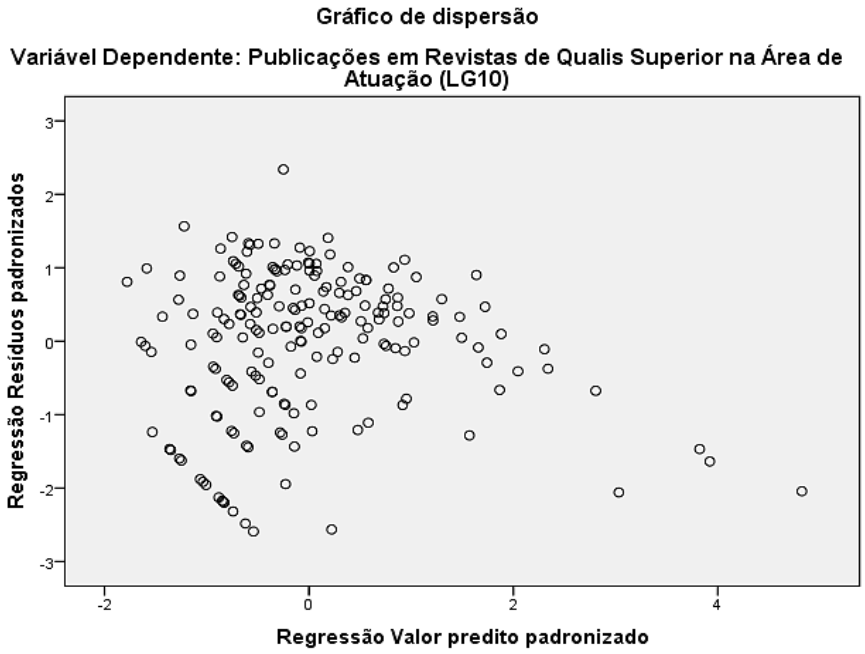


Gráfico 4 - Dispersão para variável "Artigos Qualis Superior Capes (LG10)"



Os resultados das duas análises de regressão corroboram com os nossos testes anteriores de diferença entre médias. Tratamos de um público que tende a produzir de maneira simular, sobretudo pelo fato destes sujeitos terem feito parte de um processo seletivo que considera diversos aspectos da produção acadêmica levantados nestas páginas. Estudamos um grupo que produz, notadamente, em 89%, a partir das mesmas variáveis, haja vista a pluralidade de 10 variáveis independentes como explicativas do primeiro modelo de regressão, sem colinearidade, que explicam a movimentação da variável dependente (índice de produção).

O R2 ajustado da segunda análise de regressão (0,65) se mostrou bem abaixo do R2 ajustado da primeira regressão (0,89). A considerar que estes valores permitem predizer a respeito da movimentação de variáveis importantes ao trabalho, convém atentarmos para o fato de que o índice, por ser uma variável composta, tem mais chances de se associar com indicadores que muitas vezes fizeram parte da composição do próprio índice, por endogenia, sem que isso seja considerado como colinearidade[11]. No caso da análise de regressão da variável sobre a publicação de artigos de estrato superior, não entraram no modelo outras variáveis que certamente explicariam melhor o modelo de regressão resultante, que aumentariam o valor do R2 ajustado. Que variáveis são estas, não alcançadas por este estudo? Apenas um estudo qualitativo poderia subsidiar esta resposta para a composição de um novo banco de variáveis para avaliação de nossa problemática. Entretanto, algumas poderiam ser ensaiadas: IDH entre diferentes regiões, número de vagas por universidades, economias de transferência, financiamento de projetos, financiamento de programas de pós-graduação, dentre outras.

Modelo de construção técnica de algoritmo para recolha de dados de currículos da plataforma lattes do CNPQ

Elaboração de um Webscrapping

Afinal, o que é *webscrapping*? É o processo de minerar ou coletar informações e/ou dados, que são estruturados ou não estruturados. Esse processo atende à finalidade de capturar dados ou informações em um tempo menor que um humano faria, maximizando a assertividade, ou seja, minimizando a chance de capturar uma informação errada.

[11] A explicação para a obtenção de um valor de R^2 ajustado tão elevado está no princípio da endogeneidade, isto é, o efeito (a variável dependente) é medido de modo a praticamente coincidir com as suas causas. Porém, o recurso utilizado da análise de regressão nos permitiu ordenar o impacto de cada variável de causa na construção do próprio índice. Embora o modelo possa ser suposto como uma tautologia por esta análise, o que importa em termos explicativos é a capacidade de especificação de quais variáveis de fato importam e explicam a movimentação da variável dependente e não propriamente o valor de "prova" do R^2 ajustado.

Antes de entrarmos na operacionalização de como fazer um *webscraping* evidenciamos que existem *websites* que fornecem sistemas de raspagens de dados[12], cujo objetivo é, através de uma interface *point and click*, tornar a extração de dados em *websites* mais simples possível, sendo desnecessária uma linha de código.

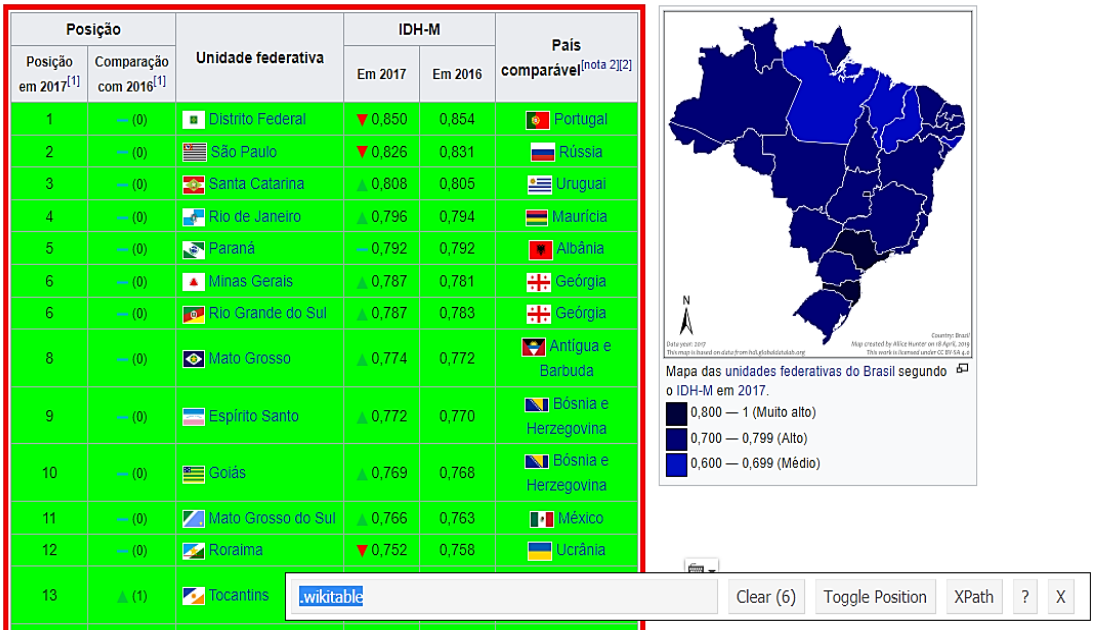
Operacionalizar um *webscraping* é passar para a sua máquina em qual elemento da página da web a informação alvo está localizada. Descreveremos o processo na linguagem R, amplamente usada na comunidade científica. A lógica do *scraping* é universal, todavia precisamos deixar claro que cada linguagem de programação tem suas especificidades, vantagens e desvantagens, tornando o processo mais demorado a depender da linguagem utilizada.

Em nosso exemplo faremos uso da extensão *selector gadget* para acessar as informações alvo e 3 *librarys* do R: *dplyr*, apropriada para a manipulação de dados, xml2, apropriada para a leitura e manipulação de arquivos em formato xml e html e rvest, que lida com a raspagem propriamente dita. Em nosso primeiro exemplo, vamos descrever como um processo de raspagem genérico é feito. Para tal, escolhemos o site do Wikipedia, posto este ser de amplo acesso a buscas pelo site google.

Acessamos a lista de Índice de Desenvolvimento Humano das Unidades Federativas do Brasil[13]. Para fazer a coleta dos dados no R é muito simples e rápido. Primeiro, identificamos no site o item que desejamos raspar. Em seguida, clicamos na extensão *selector gadget* e, depois, clicamos no *link*.

[12] Disponível em https://webscraper.io/. Acesso em 01-06- 2021
[13] Disponível em https://pt.wikipedia.org/wiki/Lista_de_unidades_federativas_do_Brasil_por_IDH. Acesso em 01-06-2021.

Imagem 1 – Identificação do Caminho do Objeto Alvo



Posição		Unidade federativa	IDH-M		País comparável[nota 2][2]
Posição em 2017[1]	Comparação com 2016[1]		Em 2017	Em 2016	
1	— (0)	Distrito Federal	▼0,850	0,854	Portugal
2	— (0)	São Paulo	▼0,826	0,831	Rússia
3	— (0)	Santa Catarina	▲0,808	0,805	Uruguai
4	— (0)	Rio de Janeiro	▲0,796	0,794	Maurícia
5	— (0)	Paraná	—0,792	0,792	Albânia
6	— (0)	Minas Gerais	▲0,787	0,781	Geórgia
6	— (0)	Rio Grande do Sul	▲0,787	0,783	Geórgia
8	— (0)	Mato Grosso	▲0,774	0,772	Antígua e Barbuda
9	— (0)	Espírito Santo	▲0,772	0,770	Bósnia e Herzegovina
10	— (0)	Goiás	▲0,769	0,768	Bósnia e Herzegovina
11	— (0)	Mato Grosso do Sul	▲0,766	0,763	México
12	— (0)	Roraima	▼0,752	0,758	Ucrânia
13	▲ (1)	Tocantins			

Vemos que a extensão no retornou um item chamado “.wikitable” que é o nome do node onde estão armazenadas as informações tipo tabela do Wikipedia. Agora, só precisamos passar essas informações para o R para processamento.

Código 1 – Transferência das Informações para o R

```
25#passando o link alvo para um objeto

26 url <-
"https://pt.wikipedia.org/wiki/Lista_de_unidades_federativas_do_Brasil_ por_IDH”

27#acessando o objeto desejado

28 urlTables<-url%>%read html(#função que lê a página em um formatohtml

          ) %>% html_nodes  ( #funçao que acessa o node indicado

".wikitable"#Vode desejadoo

                        )

32  #transformando o objeto acessado em um banco de dados

33   df<-as.data.frame( # função que transforma o objeto em um banco de
dados

          html_table( #função que passa a pagina html para uma tabela

35 urlTables[1] # localização do objeto
```

Dentro do ambiente do R, primeiro, passamos a página alvo para um objeto, cujo nome fica a critério do operador. Fizemos uso do nome url para facilitar a compreensão e o salvamos na interface do R. Na linha 28, transformamos a nossa página da web em um html, e indicamos para o R qual o nodo, ou endereço da informação alvo. Em seguida, transformamos o objeto que estava em formato html em uma tabela html, e passamos o endereço da tabela desejada, o endereço [1] indica que é a primeira tabela a aparecer no site. Por fim, transformamos o objeto em banco de dados. Essa é a rotina padrão de um *webscraping* para tabelas, todavia, também podemos coletar dados que não estão em formatos tabulares, e sim desagregados em uma página.

Imagem 2 – Página Web da Confederação Brasileira de Futebol

Neste exemplo faremos uso do site da Confederação Brasileira de Futebol/CBF - para coletar os resultados das partidas do Campeonato Brasileiro de Futebol/Brasileirão. Seguindo o mesmo procedimento, fazemos uso do *selector gadget*, ou também podemos clicar com o botão direito do mouse, e clicamos no elemento alvo, onde será redirecionado ao html da página que também nos retorna com a localização do elemento, em seguida passamos as informações alvo para o R.

Código 2 – Modelo de *Script*

```
47 #coletando as informações da rodada

48   rodada<- resultadosBrasileirão >> html_nodes (# declarando a
localização".aside-header .text-center")%>%

50     html_text ()#passando as informações para um formato de texto

51   mandante <- resultadosBrasileirão > html_nodes declarando a
localização

".pu11-left .time-sigla")%>%

53   html_text ()#passando as informações para um formato de texto

54   visitante <- resultadosBrasileirão > html_nodes ( #declarando a
localização

".pu11-right .time-sigla")%>%

56   html_text ()#passando as informações para um formato de texto

57

58 tabelaBrasileirao <- data.frame (rodada = rodada,

mandante = mandante,

visitante = visitante


61      )
```


Já no *script* abaixo seguimos os mesmos procedimentos. Primeiro, passamos ao site alvo um objeto, o qual foi declarado nesse exemplo como resultados do Brasileirão e o salvamos no ambiente do R. Em seguida, transformamos a página um formato html. Realizamos o mesmo processo de identificar o elemento que continha as informações sobre rodadas através do *selector gadget* e passamos para o *script*. Transformamos o objeto em html text. Fizemos o mesmo processo para a variável mandante e visitante.

Os exemplos apresentados tratam das formas de *scrap* mais comuns que um cientista social lida no dia a dia de pesquisa. A primeira forma, por coleta de dados tabulares, já contém as informações sistematizadas. A segunda, por coleta de dados não sistematizados, que algumas vezes necessita de mais tempo explorando a página, seja usando extensões, como o selector gadget, ou explorando o código fonte da página, via inspecionar o elemento, ou acessando o código fonte da página.

A busca por dados contidos em Currículos Lattes da plataforma do CNPq

O presente subtópico tomará como exemplo a investigação realizada sobre a produtividade acadêmica de bolsistas PQs do CNPq, metodologicamente criamos uma função dentro do ambiente R que raspava as informações alvo em cada um dos currículos lattes, seja do universo dos 200 bolsistas PQ de sociologia ou da amostra dos bolsistas PQ do Brasil. A função foi criada por necessidade da pesquisa, que era mensurar as dimensões de produtividade científica dos pesquisadores. O primeiro passo de recolha técnica foi a seleção de quais bibliotecas deveriam ser usadas na recolha de dados. Em nosso trabalho, utilizamos as bibliotecas do pacote *tidyverse*, que trazem funcionalidades desde a manipulação de banco de dados até a visualização dos dados em si.

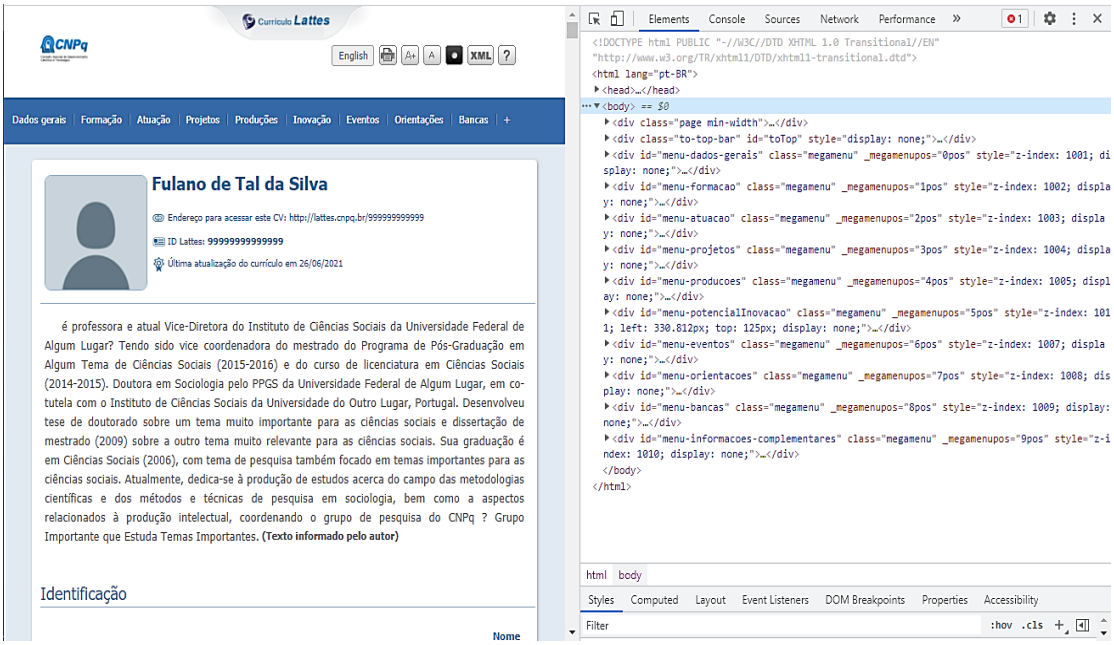
O segundo passo foi declarar qual página seria alvo da raspagem. Como no nosso caso era a página de currículos lattes da Plataforma do CNPq[14], tínhamos a limitação de que o CNPq faz uso do reCAPTCHA2, uma API do google que tem como objetivo limitar algoritmos de raspagem de coletarem um grande volume de informações sobre os pesquisadores brasileiros. Entretanto, toda página que é acessada pelo nosso navegador pode ser baixada localmente. Foi a estratégia usada na investigação, onde acessaríamos manualmente as páginas do lattes e baixaríamos os currículos para rodarmos o algoritmo localmente nos computadores.

O terceiro passo foi definir quais informações seriam coletadas. Utilizamos as variáveis que já tínhamos previstas desde uma pesquisa anterior, que tratavam acerca de informações sobre a formação do pesquisador, premiações, produções bibliográficas, orientações e organizações de eventos. A partir do desenho já estruturado, pudemos avançar para a parte mais simples da pesquisa, que era a execução do algoritmo.

O quarto passo consistiu em encontrar onde estavam localizadas as informações alvo dentro dos elementos da página do currículo lattes. Apertando a tecla F11 acessamos os elementos da página *web*. Como podemos ver abaixo os elementos que compõem a página, cada informação está estruturada em uma *div*. Só precisamos acessá-la dentro do ambiente do R para coletar essas informações.

[14] Disponível em http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar. Acesso em 01-06-2021.

Imagem 3 – Elementos da Página Web de Currículos Lattes CNPq



Fonte: CNPQ/Lattes (2021).

Código 3 – Descrição da Passagem da Página Web para o Ambiente R


```
#onde my.l é o objeto que vai conter o currículo
    #XML :: é a biblioteca utilizada para manipular páginas da web
    my.l <- XML::xmlToList (#função que transforma o formato XML em objeto do tipo lista
      XML :: xmlParse (#funcão para gerar uma estrutura que OR pode manipular e ler
      file.path ( my. Tempdir, #função do que define o diretório que contem o curriculo
            "curriculo.xml"#nome do arquivo com o formato.xml
      ) ,  encoding = my.encoding) #codificação usada no sistema operacional
          ) #fim da função
```

A passagem para o ambiente R é muito simples. Agora que já temos acesso ao currículo no nosso ambiente de trabalho, precisamos apenas definir onde estão localizadas as informações e coletar cada uma delas.

Código 4 – Declaração da Localização das Informações a Serem Raspadas

```
10  #para informações da graduação GRAD é o produto final
11   GRAD <- do.call ( #funcão para concatenar mutiplas funções
12              c, #função para combinar elementos
13           list (#função para listar os elementos
14           #localização das informações sobre a graduação do
pesquisador
15    my.l $'DADOS-GERAIS' $'FORMACAO-ACADEMICA-
TITULACAO'$GRADUACAO$.attrs))
16   #para informações de mestrado e doutorado o processo é o mesmo
17    MESTRADO <- do.call (c, list (my.l$'DADOS-GERAIS'$ 'FORMACAO-
ACADEMICA-TITULACAO $MESTRADO$.attrs))
18   DOUTORADO <- do.call (c, list (my.l$'DADOS-GERAIS'$ 'FORMACAO-
ACADEMICA-TITULACAO $DOUTORADO$.attrs))
19   DADOS. GERAIS <- do.call(c. list (my.l$'DADOS-GERAIS'$.attrs))
20   AREAS <- do.call(c, list(my.l$'DADOS-GERAIS'$'AREAS-DE-ATUACAO'))
21   #Também podemos acessar definindo onde a informação está localizada
22   #fazendo uso de colchetes
23   GArea <- my.1'DADOS-GERAIS'$'AREAS-DE-ATUACAO' [[1]] [2]
24   AArea <- my.l'DADOS-GERAIS'$'AREAS-DE-ATUACAO' [[1]][3]
```


Após os elementos já coletados, iniciamos o processo de sistematização em um banco de dados. Fazemos uso das funções *bind_cols*, as *tibble* e *t* para transformar os dados desagregados em um banco de dados de formato tabular.

Para itens que contêm uma série de elementos, precisamos construir uma estrutura que testa se o valor do item é nulo ou não. Isto pode ser resolvido a partir do uso da função *if*, que faz um teste lógico, seguido da função *for* que é uma função utilizada para estruturas de repetição. Nestes casos, ela percorre cada unidade do objeto e aplica as funções que inserimos na estrutura de repetição, tal como podemos ver na imagem abaixo.

Código 5 – Declaração da Localização das Informações a Serem Raspadas

```
36    #Declaração a localização da informação desejada
37   ORIENTACOES <- my.l$ 'OUTRA-PRODUCAO'$ 'ORIENTACOES-CONCLUIDAS'
38
39   if ( #função para fazer um teste lógico
40        !is.null (ORIENTACOES # se diferente de nulo para orientações
então
41                              )) {
42   for (#função for prepara uma estrutura de repetição
43    i.orient inORIENTACOES #para cada 7 em
orientações 44  )) {
45   i.orient[[1]]#declara o caminho do objeto desejado i.orient
46   i.orient[[2]]
47   course <-i.orient[[1]]["NATUREZA"]
48   type.course <-i.orient[[1]]["TIPO"]
49   std.name <-i.orient[[2]]["NOME-DO-ORIENTADO"]
50   year. supervision <-as.numeric #tratada o vetor como numerico
51                         i.orient[[1]]["ANO"])
52   temp.df <- dplyr::tibble( #a função as tibble transforma a lista em
um banco de dados
54                        id.file =basename(zip.in), #identifica com o
nome do pesquisador
56                        name = data. Tpesq$name, situation =
"CONCLUIDA",
58                        type.course, course, std. name, year.
supervision)
60                         rownames(temp.df) <- NULL#remove a coluna
rowrames
62                         data. supervisions <-rbind( #função para
empilhar as linhas do dataframe
64                       data.supervisions, temp.df)
```


Depois das informações já sistematizadas, podemos partir para a etapa de análise. A principal vantagem de termos construído um *webscraping* foi a certeza que nossos dados

mediam exatamente aquilo que queríamos que fosse medido. A probabilidade de erro de coleta era muito próxima de 0, permitindo assim achados mais robustos, transparentes e replicáveis.

O *webscraping* como técnica de coleta de dados se mostra como uma ferramenta elementar para a pesquisa em ciências sociais, já que hoje documentações, relatórios, indicadores, estatísticas e parâmetros estão disponibilizados em repositórios digitais. Seu uso tem muitas vantagens que vão desde o tempo destinado à coleta das informações (e ao tratamento dos dados que é inúmeras vezes menor quando comparada à coleta manual) até a qualidade do dado produzido pela pesquisa, que depende diretamente do repositório onde os dados da pesquisa são coletados, evitando assim erros de coleta.

No quadro abaixo comparamos a duração da pesquisa em que usamos coleta manual das informações de currículos lattes, e da pesquisa que fazemos uso do algoritmo de raspagem.

Quadro 1 – Comparação entre Técnicas de Coleta

Etapa da Pesquisa	Coleta Manual	Coleta Automatizada
Validação da coleta / pré teste	15 dias	3 segundos
Tempo para coleta dos dados	15 minutos por currículo	3 segundos por currículo
Tempo para tabulação dos dados	8 minutos por currículo	
Tempo de desenho do método de coleta	1 dia	10 dias
Tamanho da amostra	200 casos	601 casos
Período total de coleta	120 dias	12 dias

Considerações finais

Tendo em vista que a relevância dos resultados de uma investigação depende especialmente da qualidade dos dados utilizados, nos dedicamos a tratar metodologicamente sobre a importância da utilização de dados primários precisos, replicáveis e com processo de coleta de dados transparente. Fazer uso da técnica de *webscraping* é uma alternativa bastante interessante ao universo da investigação social, posto que esta técnica permite a coleta de grande volume de dados, de forma mais rápida quando comparada à recolha de informações de forma manual. Ademais, o método possibilita que outros pesquisadores repliquem a pesquisa a partir de diferentes situações nas quais a replicabilidade venha ser importante para validação de resultados.

Especificamente no caso da pesquisa sobre os bolsistas PQ isso nos possibilitou avançar em relação à literatura existente, que de forma geral tende a se concentrar em elementos mais descritivos. O acesso a uma ampla base de dados e a formulação de indicadores constitui parte

indispensável para a construção de uma análise quali-quantitativa, e como indicado no início deste artigo, este ainda é um desafio para o campo das ciências sociais brasileiras.

Para pesquisadores que estudam temas como administração pública ou judiciário, o uso da técnica *webscraping* parece ainda mais imprescindível atualmente, posto que grande parte dos documentos da administração pública se encontra em formato digital. Outro ponto importante a ser salientado é que o *webscraping* permite a realização de pesquisas que outrora seriam impossíveis sem um *budget* de pesquisa muito alto, já que permite que grande volume de dados seja coletado sem que pesquisadores tenham de realizar a coleta de forma manual, de caso a caso.

Por fim, buscamos por este breve artigo aproximar a técnica aos pesquisadores brasileiros, bem como trazer como alternativa investigativa uma abordagem mais computacional às ciências sociais brasileiras. Entendemos que o amplo uso de técnicas *webscraping* poderá promover novas possibilidades à produção científica das ciências sociais, certa vez que permite que pesquisadores dediquem mais tempo à construção robusta de argumentos teóricos e analíticos a seus problemas de pesquisa do que, necessariamente, à fase de coleta e sistematização de dados.

Referências

ALEXANDER, Jeffrey (1987), "O Novo movimento teórico". *Revista Brasileira de Ciências Sociais*, v. 2, n. 4, pp. 5-28 [Consult. 01-07-2021]. Disponível em https://www.anpocs.com/index.php/publicacoes-sp-2056165036/rbcs/233-rbcs-04

CANO, Ignacio (2012), "Nas trincheiras do método: o ensino da metodologia das ciências sociais no Brasil". *Sociologias*, Ano 14, n. 31, pp. 94-119 [Consult. 01-07-2021]. Disponível em https://www.scielo.br/j/soc/a/QC6rphm93gZgXmt6FSqWJys/abstract/?lang=pt

CAMARGO-VEGA, Juan José; CAMARGO-ORTEGA, Jonathan Felipe; JOYANES-AGUILAR, Luis. (2015), *Conociendo big data*. Facultad de Ingeniería, v. 24, n. 38, pp. 63-77 [Consult. 01-11-2021]. Disponível em http://www.scielo.org.co/pdf/rfing/v24n38/v24n38a06.pdf

FIGUEIREDO FILHO, Dalson *et al*. (2011), "O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO)". *Revista Política Hoje*, v. 20, n. 1 [Consult. 01-11-2021]. Disponível emhttps://periodicos.ufpe.br/revistas/politicahoje/article/view/3808

IBM SPSS Statistics for Windows (2019), Versão 26.0. Armonk, NY, IBM Corp.

KING, Gary (1995), "Replication". *Political Science and Politics*, v. 28, pp. 443-499 [Consult. 01-07-2021]. Disponível em https://doi.org/10.2307/420301

MINAYO, Maria Cecilia de S.; SANCHES, Odécio. (1993), "Quantitativo-qualitativo: oposição ou complementaridade". *Cadernos de Saúde Pública*, v. 9, pp. 237-248 [Consult.01-09-2021]. Disponível em https://doi.org/10.1590/S0102-311X1993000300002

KUHN, Thomas. (1978), *Revoluções Científicas*. São Paulo, Perspectiva.

MINAYO, Maria C.S. (1994). *O desafio do conhecimento científico: Pesquisa Qualitativa em Saúde* (2a edição). SP-RJ, Hucitec-Abrasco.

NEIVA, Pedro (2015), "Revisitando o calcanhar de aquiles metodológico das ciências sociais no Brasil". *Sociologia, Problemas e Práticas*, v. 79, pp. 65-83 [Consult. 01-11-2021]. Disponível em https://journals.openedition.org/spp/2232

PARANHOS, Ranulfo *et al*. (2013), "Corra que o survey vem aí: Noções básicas para cientistas sociais". *Revista Latinoamericana de Metodología de la Investigación Social*, n. 6, pp. 7-24 [Consult. 01-11-2021]. Disponível em https://dialnet.unirioja.es/servlet/articulo?codigo=5275921

R Core TEAM (2021), "R: A language and environment for statistical computing". *R Foundation for Statistical Computing*. Vienna [Consult. 01-07-2021]. Disponível em https://www.R-project.org/

SILVA, Rosalina C. (1998), "A falsa dicotomia qualitativo-quantitativo: paradigmas que informam nossas práticas de pesquisas", in Romanelli G., Biasoli-Alves Z. M. M. *Diálogos metodológicos sobre prática de pesquisa. Programa de Pós-Graduação em Psicologia da FFCLRP USP/Capes*. Ribeirão Preto, Legis-Summa. pp. 159-174.

SOARES, Gláucio Ary Dillon. (2005), "O calcanhar metodológico da ciência política no Brasil". *Sociologia, Problemas e Práticas*, Oeiras, n. 48 [Consult 01-07-2021]. Disponível em http://www.scielo.pt/scielo.php?script=sci_arttext&pid=S0873-65292005000200004&lang=pt

WOLF, Fredric M. Wolf, Fredric M., *Meta-Analysis*: Quantitative Methods for Research Synthesis. Beverly Hills, *CA: Sage*, 1986.

Abstract

This article presents a brief introduction to the algorithms for data collection used in online repositories in social sciences investigations based on empirical research on academic production of CNPq scientific productivity grants in Brazil. Methodologically, we present a study carried out to apply the computational technique of elaborating algorithms. Then we describe the step-by-step planning of the scrapping algorithm using the R software. The computational approach for data collection is understandable and promotes its use in the field of social sciences, making data collection in institutional repositories more systematic, transparent, replicable and quick.

Keywords: Social sciences; academic production; quantitative data collection; algorithms.

Resumen

Este artículo presenta una breve introducción al uso de algoritmos para la recolección de datos en repositorios en línea en investigaciones en el campo de las ciencias sociales a partir de un modelo de investigación empírica sobre la producción académica de becarios de productividad del CNPq en Brasil. Metodológicamente, presentamos una investigación realizada con la aplicación de la técnica computacional de elaboración de algoritmos y luego describimos la planificación paso a paso del algoritmo de *scrapping* utilizando el software R. La técnica computacional para la recolección de datos intenta promover su uso en el campo de las ciencias sociales, haciendo que la recolección de datos en los repositorios institucionales sea más sistemática, transparente, replicable y rápida.

Palavras clave: ciencias sociales; producción académica; recopilación de datos cuantitativos; algoritmos.