Extração de Dados via Web Scraping como Suporte em Análises Envolvendo a Geração Distribuída

Renan Moreira Soares. Guilherme Rezende Pereira Camargo.
Marcelo Escobar de Oliveira. Leonardo Garcia Marques.

Instituto Federal de Goiás, Itumbiara, Goiás. (e-mail: renanmoreirasoares@gmail.com)

Abstract: In recent years, relevant growth has been evidenced in the photovoltaic generation sector in Brazil. The analysis of this growth is fundamental for decision-making, both in the public and private sectors. This growth can cause major impacts, both on the electrical system and on the quality of electricity delivered to consumers. Therefore, investigating the factors that can influence this increase is of supreme importance so that actions and investments can be carried out towards the improvement of the electricity networks. To verify the influence of these socioeconomic factors on the growth, statistical studies have been developed. However, to carry them out, a huge amount of data needs to be collected to ensure robust analysis. These data can be collected manually, on the internet, on the websites where are available. However, this manual retrieval is slow, error-prone and can compromise data reliability. So, in this work, web scrapers tools are presented to collect data from two different sites, supporting different research that can be carried out on the growth of distributed generation. At the end, an analysis with the collected data is shown, demonstrating the usefulness of these tools.

Resumo: Nos últimos anos um crescimento relevante foi evidenciado no setor da geração fotovoltaica no Brasil. A análise deste crescimento é fundamental para a tomada de decisões, tanto do setor público quanto do setor privado. Este crescimento pode causar grandes impactos, tanto no sistema elétrico, quanto na qualidade da energia elétrica entregue aos consumidores. Logo, investigar os fatores que podem influenciar neste aumento é de suma importância para que se possa realizar ações e investimentos em prol da melhoria das redes de energia elétrica. Assim, para verificar a influência desses fatores socioeconômicos no crescimento da geração distribuída, estudos estatísticos vêm sendo realizados. No entanto, para realizá-los, uma enorme quantidade de dados precisa ser coletada, para garantir uma análise robusta. Estes dados podem ser coletados de forma manual, na internet, nos sites onde são disponibilizados. No entanto, esta obtenção manual é lenta, susceptível a erros e pode comprometer a confiabilidade dos dados. Assim, neste trabalho, são apresentadas ferramentas web scrapers para coletar dados de dois diferentes sites, oferecendo suporte a pesquisas que podem ser realizadas sobre o crescimento da geração distribuída. Ao fim, uma análise com os dados coletados é apresentada, demonstrando a utilidade dessas ferramentas.

Keywords: data extraction; distributed generation; socioeconomic factors; statistic; web scraping.

Palavras-chaves: estatística; extração de dados; fatores socioeconômicos; geração distribuída; web scraping.

1. INTRODUÇÃO

A quantidade de dados disponíveis na internet vem aumentando com o passar dos anos, fomentando novas áreas, principalmente quando se refere a ciência dos dados. Essa ciência tem como foco lidar com os dados, prepará-los para análises, utilizando de recursos de diversas áreas, desde ferramentas estatísticas até inteligência artificial. Portanto, a partir das informações extraídas, se faz possível tomar decisões e obter conhecimentos valiosos.

O processo de extrair estes dados varia, podendo ser de forma manual ou automatizada. A prática de extração de dados na internet existe desde seus primórdios, já tendo sido chamada de *screen scraping*, *data mining* e outras variações. Atualmente, o termo em consenso é *web scraping*.

Alguns autores consideram a programação como um tipo de mágica, assim como consideram o *web scraping* a prática dessa magia. Isto se dá pela capacidade desta ferramenta de "fazer proezas particularmente impressionantes e úteis" (MITCHELL, 2019).

O tradicional copiar e colar é uma prática de extração de dados, no entanto, sua utilização exige diversas repetições e esforço humano, além de despender tempo. Com base nisso, a citação anterior compara a pratica de *web scraping* à aplicação da magia, facilitando a coleta de dados a partir de sua automatização via programação, sendo um processo impressionante e útil. Além disso, a coleta realizada por um

humano está sujeita a erros, portanto, a coleta automatizada aumenta a confiabilidade dos dados.

Esse aumento de confiabilidade reflete na qualidade das análises realizadas. Diversas análises vêm sendo realizadas no setor elétrico, auxiliando na tomada de decisões. No trabalho de Borges (2020), uma análise foi realizada coletando dados de potência instalada e indicadores socioeconômicos de diversas cidades para analisar a correlação entre o crescimento da geração distribuída e os indicadores.

No entanto, esta coleta foi realizada de forma manual, sendo exaustiva e susceptível a erros. Além disso, para realizar esta análise considerando o país inteiro, o autor despenderia bastante tempo, devido à imensa quantidade de dados que devem ser coletados, podendo também comprometer as informações obtidas.

Portanto, com base nessa dificuldade, este trabalho apresenta a construção, em linguagem *Python*, de *ferramentas web scrapers* de forma a coletar estes dados, visando dar suporte a análises do crescimento da geração distribuída. Além disso, apresenta a robustez dessas ferramentas aplicadas em uma análise estatística similar a realizada por Borges (2020), no entanto, considerando todo o território brasileiro.

2. WEB SCRAPING

Ferramentas que são utilizadas para automatizar a coleta de dados na internet são conhecidas como web scrapers, também chamadas de *web* robôs, devido ao seu comportamento automatizado. Conforme já citado, diminuem o esforço humano, realizando as tarefas repetitivas de forma automatizada. O ato de acessar repetidamente uma página da internet, requisitar os dados e então analisá-los para extrair as informações desejadas pode ser automatizado por meio da técnica conhecida como *web scraping*.

Estas ferramentas são utilizadas, principalmente, quando um serviço ou site não disponibiliza, ou limita, uma interface de programação de aplicações – API, recurso que oferece acesso a dados e serviços. Mesmo com API disponível, vários sites e ferramentas não cobrem todas as demandas que o usuário possa necessitar (GLEZ-PEÑA et al. 2014). Nestes cenários, as ferramentas automatizadas se tornam valiosas, permitindo o acesso às informações não disponibilizadas, mas que estão visíveis nos sites.

Existem diferentes técnicas para realizar *web scraping*, desde o tradicional copiar e colar, assim como técnicas que levam em consideração a posição do texto contendo os dados, assim como diversas outras, segundo Saurkar, Pathare e Gode (2018). Comparando as técnicas, com base no estudo anterior e nos estudos de Sirisuriya (2015) e Gunawan et al. (2019), verifica-se que o formato do site é um fator determinante na escolha da técnica que mais se adequa para a coleta de dados.

A escolha da técnica influi no tempo de execução, na utilização de memória (GUNAWAN et al. 2019), e depende dos dados e do local em que estão disponibilizados. Assim, a forma que o site dispõe as informações, a linguagem de marcação utilizada em sua construção são fatores que devem ser analisados.

2.1 Aplicações no setor elétrico

As ferramentas *web scrapers* são utilizadas em diversas análises, conforme já explanado. Considerando o setor elétrico, estas ferramentas foram utilizadas em estudos de previsão de demanda elétrica (ÇAMURDAN e GANIZ, 2017), extraindo dados públicos como medidas e observações relacionadas ao mercado elétrico na Turquia. O modelo construído determina a quantidade de eletricidade, em diferentes condições, que deverá ser produzida de acordo com o tempo, como anos, meses e dias.

Na Itália (NOUSSAN, ROBERTO e NASTASI, 2018), estas ferramentas foram utilizadas para coletar dados de indicadores de performance baseados no consumo primário de energia e no compartilhamento de fontes renováveis, além de emissões de CO2. De tal forma, uma análise de dados da produção da energia pode ser realizada.

A utilização de *web scrapers* para coletar dados para diferentes estudos é uma prática comum e facilitadora. Para automatizar a coleta de dados para estudos que analisam o crescimento da geração distribuída e fatores socioeconômicos (BORGES, 2020), se faz necessário coletar dados de dois domínios.

Primeiramente, a coleta de dados socioeconômicos é realizada no site do Instituto Brasileiro de Geografia e Estatística – IBGE. Assim, são coletadas informações que refletem as características da sociedade.

Além disso, é necessário coletar dados de geração distribuída, que são disponibilizados pela Agência Nacional de Energia Elétrica – ANEEL. Os dados se encontram abertamente no banco de dados chamado Dados Compilados de Geração Distribuída, podendo ser consultados por qualquer pessoa com acesso à internet.

Apesar de disponíveis publicamente, estes dados não podem ser exportados facilmente, exigindo que a coleta dos mesmos seja feita de forma manual. Assim, a utilização das ferramentas automatizadas se mostra eficaz para coletar grandes quantidades de dados.

A seguir, estão descritos os procedimentos utilizados na construção das ferramentas para cada um dos sites, considerando suas peculiaridades.

2.2 Web Scraper IBGE

O formato do site influencia nas técnicas utilizadas para a coleta de dados, portanto, deve ser analisado a fim de verificar a que melhor se encaixa. No caso do IBGE, as informações estão separadas em diversas páginas, de acordo com as esferas. São apresentados dados municipais, estaduais e do país inteiro. Em cada uma dessas esferas, existem diversas outras páginas que apresentam dados gerais ou específicos.

As páginas são construídas em linguagem HTML (*HyperText Markup Language*), uma linguagem de marcação de texto. Apesar de construídas na mesma linguagem, algumas páginas apresentam disposição diferente dos dados. Portanto, a ferramenta *web scraper* deve se adaptar à esta peculiaridade.

Para iniciar a extração, precisa-se primeiro conhecer a esfera que o usuário deseja coletar os dados: estadual ou municipal. Uma lista com as localidades é pré-carregada, no caso dos municípios, estas foram obtidas no próprio site do IBGE por outra ferramenta *web scraper*. Assim, a ferramenta irá percorrer essa lista, estabelecendo a conexão com a página que se encontra os dados para cada localidade.

Além disso, o usuário deve inserir qual tipo de dados ele quer coletar, baseado na lista de pesquisas disponíveis do IBGE. A depender da pesquisa selecionada, o *web scraper* se adapta a localização dos dados.

A pesquisa IBGE chamada Panorama, apresenta dados gerais da localidade como número de habitantes, rendimento, entre outros. A Tabela 1 apresenta as *tags* e atributos que indicam a localização destes dados no código fonte da página.

As outras pesquisas apresentam dados específicos da localidade, a depender de qual seja selecionada, podendo apresentar dados apenas sobre educação, rendimento, agricultura, dentre diversas outras. A Tabela 2 apresenta a localização dos dados para essas pesquisas.

Tabela 1. Localização dos dados - Panorama

Dados	*Tag*	Atributo	Valor do atributo
Nomes	tr	class	lista__indicador
	td	class	lista__nome
Valores	tr	class	lista__indicador
	div		
Unidades	tr	class	lista__indicador
	span		

Tabela 2. Localização dos dados – Pesquisas tipo 1 e 2

Dados	*Tag*	Atributo	Valor do atributo (1)	Valor do atributo (2)
Valores dos indicadores	td	class	valor s	valor s sem-valor
		class	“nível-”	
Nomes dos indicadores	div	class	label--com-acao	label--com-acao
	td	class	valor s	valor s sem-valor

Para os dados localizados de acordo com a Tabela 1, utiliza-se a técnica de HTML *Parsing* para encontrar os dados. Essa técnica consiste em analisar o código HTML do site e extrair as informações relevantes. Assim, utilizando as bibliotecas *Requests*, *Selenium* e *BeautifulSoup*, o código fonte da página é coletado, e é feita uma busca nas estruturas de marcação HTML do documento. Obtendo os dados desejados, que são organizados em forma de planilha.

Um processo parecido é realizado para as demais pesquisas, separadas em dois tipos, a depender da localização dos dados. Além disso, estas são organizadas em tabelas, possuindo níveis hierárquicos. Portanto, foi necessária a inclusão, em algumas páginas, de expressões regulares, técnica onde realiza-se busca de textos que obedecem a um padrão. Neste caso, o padrão era a presença do valor do atributo “nível-”, onde na sequência apresenta-se um número que indica o nível hierárquico da tabela. Assim, os dados são coletados, a depender do tipo da pesquisa, e salvos em uma planilha. A Figura 1 apresenta, de forma resumida, o fluxograma do funcionamento do *web scraper* IBGE.

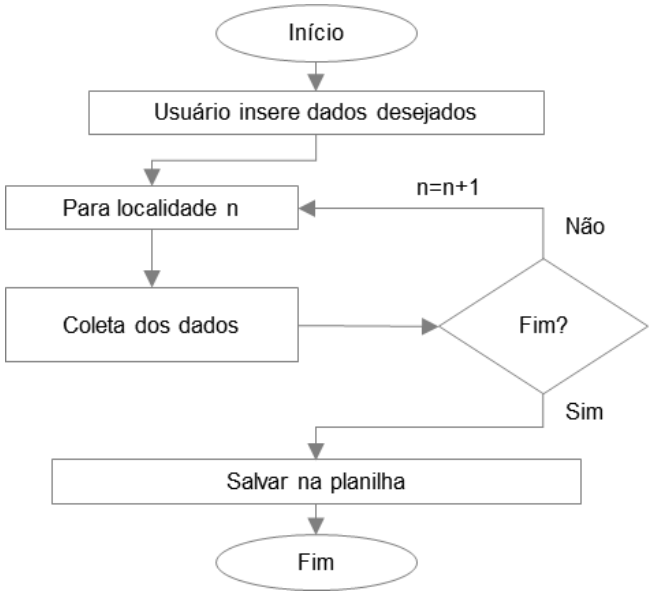

Fig. 1 Fluxograma do *web scraper* IBGE.

É importante ressaltar que esta ferramenta percorre diversas páginas, realizando diversas conexões ao site, podendo ser visto como um ataque. Portanto, cuidados éticos e legais devem ser tomados, levantados por Mitchell (2019) e Krotov e Silva (2018).

O primeiro cuidado é verificar se o site permite ou não a prática de *web scraping* por meio do seu arquivo “robots.txt”. No caso do IBGE, este não deixa claro se permite ou não, porém, apresenta bloqueios depois de conexões sucessivas, que são vistas como um ataque.

Para resolver este impasse, o *web scraper* deve ter seu comportamento otimizado, visando se passar por um humano acessando o site. Assim, a cada conexão, a ferramenta foi programada para esperar tempos aleatórios, se aproximando do comportamento de uma pessoa.

Além disso, outras informações foram enviadas ao tentar acesso, como dados de diferentes navegadores e sistemas operacionais reais, impedindo que o site identifique a ação de um robô. Também foi enviado o último site acessado como sendo o último site coletado, evitando mais uma vez a detecção da ação automatizada ao simular a navegação humana.

Estes procedimentos não garantem que a ferramenta não seja bloqueada, mas evita. Além disso, garantem o pleno funcionamento do site, já que os acessos são realizados com espaçamento de tempo, não sobrecarregando o servidor.

A Figura 2 apresenta um formulário criado para facilitar a inserção dos dados pelo usuário, selecionando o tipo de análise e pesquisa, além do ano e estado, caso seja possível.

Fig. 2 Entrada de dados *web scraper* IBGE

A ferramenta é capaz de coletar dados de quase todas as pesquisas disponíveis no site do IBGE, expandindo as possibilidades.

2.3 Web Scraper ANEEL

De maneira diferente dos dados do IBGE, os dados de geração distribuída disponibilizados pela ANEEL se apresentam em apenas uma página, no entanto, possuindo conteúdo dinâmico. Os dados são apresentados através da ferramenta *Power BI*, da *Microsoft*, cujo objetivo é transformar uma fonte de dados em informações coerentes e visíveis.

Devido a necessidade de interação do usuário com a página para selecionar os dados que serão exibidos, se fez necessário a utilização da biblioteca *Selenium* na extração. Essa biblioteca foi criada para automatizar testes em sites, abrindo uma página no navegador onde os comandos são realizados. Assim, neste caso, a página é aberta e o usuário interage, configurando a exibição de dados, antes de solicitar a coleta.

O site apresenta diversas tabelas, cada uma apresentando as informações com base em algum critério, como dado por região, por municípios, fonte de geração, ano de conexão, dentre outros. Assim, o *scraper* deve ser capaz de identificar de qual tabela o usuário deseja coletar os dados.

Para isso, utilizou-se a técnica *XPath*, que consegue localizar pontos específicos do documento HTML com endereços, navegando pelas *tags*. Verificou-se que o endereço a ser localizado sempre continha o nome da tabela. Portanto, para coletar os dados, o usuário deve inserir o nome da tabela.

Assim, ao executar o *scraper*, primeiramente, a página é aberta, permitindo que o usuário realize todos os ajustes para a exibição dos dados possíveis do site. Em seguida, o usuário retorna o comando para iniciar a coleta da tabela desejada. Ao receber este comando, o *scraper* irá expor os dados da tabela, e iniciar a coleta dos dados, conforme apresentado pela Figura 3.

Conforme explanado, a apresentação dos dados no site é dinâmica. Inicialmente, ao expor os dados, apenas vinte linhas de dados são exibidas. Portanto, de forma automatizada, utilizando a ferramenta chamada *ActionChains*, cliques para baixo são realizados. Essa ferramenta, disponível na biblioteca *Selenium*, permite automatizar por linha de código ações de usuário como pressionar alguma tecla.

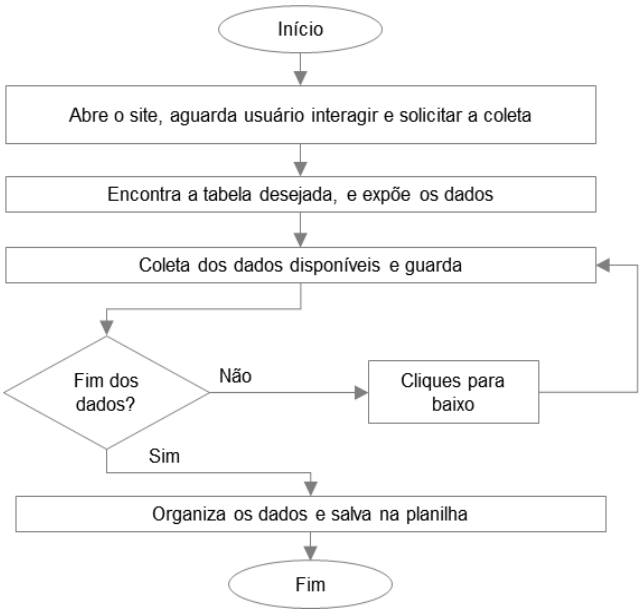

Fig. 3 Fluxograma *web scraper ANEEL*.

Assim, automatizando os cliques para baixo, mais dados são apresentados, até que todos os dados sejam coletados. A coleta de dados só é finalizada quando a soma dos dados numéricos seja igual ao total, que é o primeiro dado a ser coletado.

Em cada repetição de coleta, os dados devem ser localizados e coletados. A posição destes segue um padrão, no entanto, devido a características construtivas do site, o tamanho da tela do usuário, o tamanho da janela e outros fatores acabam influenciando. Assim, as *tags* possuem atributos de estilos mutáveis, impossibilitando a utilização de marcadores fixos, como no caso do scraper IBGE.

Para contornar isso, sabendo que os dados estão disponíveis numa *tag div* com atributo *class* com valor "b*odyCells*", coletou-se os estilos de cada coluna da tabela. Assim, informações únicas são coletadas, como largura e distância referente à margem esquerda.

Sabendo disso, construiu-se um caminho *XPath* para cada coluna, com base nos atributos de estilo coletados. Assim, os dados disponíveis são coletados, cliques para baixo são realizados, disponibilizando mais dados que também são coletados.

Os dados são organizados e salvos em uma planilha, assim como os dados obtidos do IBGE. De tal forma, os dados estão prontos para serem utilizados.

3. ANÁLISE ESTATÍSTICA

Conforme proposto, para demonstrar a relevância destas ferramentas em estudos do setor elétrico, uma análise estatística semelhante a realizada por Borges (2020) foi feita, expandindo a quantidade de dados. No trabalho em questão, o autor realizou análises em apenas um estado brasileiro, verificando a correlação entre diversos fatores socioeconômicos e geração distribuída.

Assim, para expandir a análise, dados dos mais de 5 mil municípios brasileiros foram coletados, separados por estados. Para esta análise, decidiu-se comparar a correlação entre a geração distribuída e os fatores socioeconômicos PIB per capita e Índice de Desenvolvimento Humano – IDH para cada estado.

Os indicadores socioeconômicos para cada município foram coletados utilizando a ferramenta *web scraper* IBGE. Além disso, coletou-se também a população total de cada munícipio. Devido a atrasos na publicação de indicadores recentes por parte do órgão, considerou-se que o IDH (dados de 2010) e PIB (dados de 2018) não tenham sofrido grandes alterações para o ano de 2021.

Já os dados de potência instalada fotovoltaica foram coletados pelo *web scraper* ANEEL, considerando o mês de novembro de 2021. Interações foram feitas com o site para exibir apenas os dados originados da geração fotovoltaica antes da coleta ser realizada.

Assim como na análise em que se baseia (BORGES, 2020), para garantir uma comparação justa entre grandes e pequenas cidades, construiu-se o Índice de Geração. Este índice calcula a potência instalada por habitante. A Equação (1) demonstra como este cálculo foi realizado, para cada cidade.

$$Índice\ de\ Geração = \frac{Potência\ Instalada\ [kW]}{\frac{Número\ de\ habitantes}{1000}} \quad (1)$$

Assim, a análise pode ser realizada, reduzindo erros. Para esta análise, decidiu-se por utilizar o coeficiente de correlação de Pearson, que mede a relação entre duas variáveis. O coeficiente retorna valores num intervalo de -1 a +1, onde valores positivos indicam uma correlação positiva e negativos indicam uma correlação negativa.

Uma correlação positiva significa que à medida que uma variável cresce, a outra também cresce. Já a correlação negativa, indica que a medida que uma cresce, outra decresce.

Quanto mais próximo de 1 este coeficiente se encontra, mais forte é esta correlação, independente do sinal. Quanto mais próximo de 0, menor a correlação.

A correlação de Pearson é calculada, matematicamente, de acordo com a Equação 2.

$$r = \frac{1}{n-1}\sum_{n=i}^{n}\left(\frac{x_i-\bar{X}}{S_x}\right)\left(\frac{y_i-\bar{Y}}{S_y}\right) \quad (2)$$

Onde:

- n - Quantidade de dados
- $\bar{X}, \bar{Y}$ - Média das variáveis x e y;
- S_x, S_y - Desvio padrão das variáveis x e y.

Assim, a correlação para cada estado é calculada. A Tabela 3 apresenta os resultados das correlações, organizado por estado. A correlação 1 se refere a correlação entre geração distribuída e PIB per capita, ao passo que a correlação 2 apresenta resultados entre geração distribuída e IDH.

Comparando com os resultados obtidos para o estado de Goiás em Borges (2020), verifica-se que a correlação entre geração distribuída e PIB per capita divergiu bastante (0,535). No entanto, deve-se levar em consideração que o autor realizou manipulações estatísticas adicionais que não foram realizadas aqui, além de coletar dados de geração distribuída de um período diferente.

Tabela 3. Correlação obtida por estado

Estado	Correlação 1 (PIB)	Correlação 2 (IDH)
Acre	0,2229	0,2475
Alagoas	0,2352	0,3791
Amazonas	0,5738	0,4919
Amapá	0,0867	0,5643
Bahia	0,1279	0,2875
Ceará	0,3579	0,3442
Espírito Santo	-0,0631	-0,0427
Goiás	0,2231	0,3224
Maranhão	0,4652	0,5349
Minas Gerais	0,0226	0,3952
Mato Grosso	0,1631	0,1033
Mato Grosso do Sul	0,0107	0,2263
Pará	0,2544	0,6132
Paraíba	0,0814	0,1943
Paraná	0,1909	0,4101
Pernambuco	0,1313	0,2248
Piauí	0,0319	0,1916
Rio de Janeiro	0,0583	-0,0776
Rio Grande do Norte	-0,0245	0,0786
Rio Grande do Sul	0,0947	0,3021
Rondônia	0,0728	0,3315
Roraima	0,5150	0,4317
Santa Catarina	0,0123	0,0958
Sergipe	-0,0018	-0,0032
São Paulo	-0,0443	0,0744
Tocantins	0,0108	0,0838

Quanto a correlação entre geração distribuída e IDH, os valores se mostraram próximos, sendo que o autor obteve uma correlação de 0,323. Com base nos outros resultados, pode-se verificar que Amazonas é o estado que possui maior correlação entre geração distribuída e PIB per capita, ao passo que o Espírito Santo possui a menor, sendo negativa.

Quanto ao IDH, o estado com maior correlação é o Pará, ao passo o com menor correlação é o Rio de Janeiro, também apresentando correlação negativa. Além disso, essa análise permite verificar que o IDH possui maior correlação com a geração distribuída em grande parte dos estados brasileiros, quando comparado ao PIB per capita.

Assim, investir em geração distribuída pode possuir maior impacto positivo nos indicadores onde a correlação é forte. No entanto, conforme ressaltado por Borges (2020), a correlação é apenas um indício da existência de casualidade entre as variáveis, sendo que esta deve ser verificada por outros métodos estatísticos.

4. RESULTADOS

Este trabalho teve como objetivo a criação de ferramentas que facilitem análises realizadas envolvendo a geração distribuída. Essa facilidade se refere a coleta automatizada de dados, reduzindo o esforço humano, aumentando a velocidade de coleta e a confiabilidade dos dados.

Analisando as ferramentas criadas individualmente, pode-se verificar os resultados obtidos por cada uma. A principal característica que pode ser analisada é a velocidade extração, conforme exposto na Tabela 4.

Tabela 4. Comparação entre os scrapers

	IBGE	ANEEL
Característica	Várias páginas estáticas	Uma única página dinâmica
Técnica	HTML *Parsing* e expressões regulares	*Xpath*, principalmente
Velocidade	Baixa	Alta

Quanto *ao web scraper* IBGE, tem-se como resultado uma ferramenta capaz de coletar quase todos os dados disponibilizados pelo IBGE. No entanto, devido a característica de precisar percorrer diversas páginas para coletar estes dados, as medidas utilizadas para garantir seu funcionamento reduziram sua velocidade. Quando coletando dados de estados com elevado número de cidades, como Minas Gerais, a ferramenta pode chegar a despender horas.

De maneira análoga, o *web scraper* ANEEL não precisa simular o comportamento humano para evitar bloqueios, já que apenas um acesso ao site é realizado, portanto, sua coleta é bastante veloz. Como maneira de exemplificar, para o pior caso, coletando os dados de todos os munícipios brasileiros ao mesmo tempo, a ferramenta leva menos de uma hora.

Quanto aos dados que podem ser coletados, incialmente, o foco seria apenas os dados de geração distribuída fotovoltaica. No entanto, a ferramenta foi construída sendo capaz de coletar dados de geração distribuída, independentemente do tipo, classificados por localidade, fonte de geração, tipo e modalidade de geração, além de ano da conexão e outros. Além disso, as diversas opções do site permitem que dados específicos sejam coletados.

Quanto a análise estatística realizada na seção anterior, esta demonstra o poder das ferramentas ao coletar enormes quantidades de dados. Estes dados demorariam semanas para serem coletados por humanos, além desse processo acabar comprometendo a confiabilidade dos mesmos, já que a coleta humana pode conter erros.

5. CONCLUSÕES

Neste trabalho, foi apresentada a construção de *web scrapers* para realizar coleta de dados a fim de dar suporte em pesquisas relacionadas a geração distribuída. Estas ferramentas já são consagradas e aplicadas em diversos estudos, facilitando-os.

Assim, construiu-se uma ferramenta automatizada a fim de coletar dados socioeconômicos disponibilizados pelo IBGE, resultando numa ferramenta capaz de coletar a maioria dos dados disponibilizados pelo órgão. Isto permite que estudos além do que este trabalho espera dar suporte possam ser realizados com maior confiança, robustez e velocidade.

Além disso, uma ferramenta automatizada para coletar os dados de geração distribuída da ANEEL foi produzida. Estes dados combinados podem ser utilizados em estudos da influência dos fatores socioeconômicos no crescimento da geração distribuída.

Estas ferramentas tiveram que ser construídas de formas distintas, devido a características intrínsecas dos sites onde seus dados estão alocados. Isto implicou na velocidade de extração, que apesar de longa no caso da extração de dados do IBGE, é mais rápida e confiável do que se fosse realizada manualmente.

A fim de demonstrar o poder destas ferramentas, uma análise foi realizada, com uma grande base de dados coletados sem esforço. Com isso, verificou-se a existência de correlação entre geração distribuída e dois indicadores socioeconômicos para cada estado brasileiro, verificando um destaque para o IDH.

Com estas ferramentas, a coleta de dados para pesquisas envolvendo geração distribuída brasileira foi aprimorada. Diminuindo esforços, aumentando a confiabilidade e permitindo que diversas análises sejam realizadas, além da aqui exemplificada. Com estas ferramentas, análises do crescimento e tendência no cenário da geração distribuída podem ser realizadas, em um leque de possibilidades infinitas. Assim, os pesquisadores podem se concentrar no desenvolvimento das análises, aproveitando o tempo e energia que seriam gastos coletando os dados para realizar proezas fascinantes.

AGRADECIMENTOS

Este foi um projeto de iniciação científica financiado pelo CNPq.

REFERÊNCIAS

Borges, L. M. S. (2020). Análise de fatores socioeconômicos em relação ao crescimento da geração distribuída no estado de goiás.

Çamurdan, Z., and Ganiz, M. C. (2017, October). Machine learning based electricity demand forecasting. In *2017 International Conference on Computer Science and Engineering (UBMK)*, p. 412-417.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., and Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in bioinformatics*, 15(5), p. 788-797.

Gunawan, R., Rahmatulloh, A., Darmawan, I., and Firdaus, F. (2019, March). Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. In

International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) Comparison, Vol. 2, pp. 283-287.

Krotov, V., and Silva, L. (2018). Legality and ethics of web scraping. *Twenty-fourth Americas Conference on Information Systems.*

Mitchell, R. (2019). Web Scraping com Python: Coletando mais dados da web moderna. Novatec Editora.

Noussan, M., Roberto, R., and Nastasi, B. (2018). Performance indicators of electricity generation at country level—The case of Italy. *Energies*, 11(3), p. 650.

Saurkar, A. V., Pathare, K. G., & Gode, S. A. (2018). An overview on web scraping techniques and tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4), p. 363-367.

Sirisuriya, D. S. (2015). A comparative study on web scraping. Proceedings of *8th International Research Conference, KDU*, p. 135-140.