A UTILIZAÇÃO DE FERRAMENTAS DE *DATA SCIENCE* PARA A ANÁLISE DE DADOS ESPACIAIS: uma visão histórica do Brasil no contexto espacial

Cad Int Marcella Neves Boccardo Lanz[1]
Cap Eng Luisa Amaral de Almeida[2]
2º Ten QOCON MNS Evelyn Aparecida de Oliveira[3]

RESUMO

Este trabalho analisa a evolução histórica do setor espacial brasileiro, com ênfase na área de satélites. Tal abordagem se faz necessária para que seja possível traçar estratégias para o futuro a níveis estratégico, tático e operacional. Esse estudo, realizado por meio de ferramentas de *Data Science* e *Machine Learning*, como os modelos de análise de séries temporais multivariáveis VAR e VARMA, aplicadas sobre dados disponíveis na internet, em plataformas como *Our World in Data*, Dados.gov, CAPES, entre outros, juntamente com os conhecimentos obtidos através da metodologia bibliográfica foram utilizados na análise do setor espacial nacional de ontem, de hoje e de amanhã. Assim, foi possível concluir, por meio de parâmetros adequados dos dois modelos abordados, que o IDH, a indústria de inovação, a pesquisa científica, a capacitação de pesquisadores e o orçamento destinado à política espacial são de extrema importância para o crescimento do setor. Nesse sentido, também foi possível obter previsões futuras das séries temporais, e assim determinar uma equação da quantidade de satélites em um ano "x" para cada um dos modelos utilizados. No VARMA obteve-se um erro de previsão menor do que no VAR, entretanto a complexidade de manipulação do modelo também é superior. Por fim, vale ressaltar que a pesquisa evidenciou a necessidade do desenvolvimento e da consolidação do poder espacial nacional e do Programa Espacial Brasileiro como forma de manutenção da soberania nacional, por meio do direcionamento dos recursos orçamentários governamentais para fontes que de fato são aplicadas no setor.

Palavras-chave: Setor aeroespacial. Satélites. *Data-Science*. *Machine Learning.*

1 Curso de Formação do Oficiais Intendentes da Academia da Força Aérea. E-mail: marcella.lanz@gmail.com.
2 Mestre em Engenharia Eletrônica e Computação. Instituto Tecnológico de Aeronáutica, Academia da Força Aérea. E-mail: luamaralalmeida@gmail.com.
3 Mestre em Modelagem Computacional. Universidade Federal de Juiz de Fora, Academia da Força Aérea. E-mail: evelyn1.oliveira@gmail.com.

THE USE OF *DATA SCIENCE* TOOLS FOR THE ANALYSIS OF SPATIAL DATA: a historical view of Brazil in the spatial context

ABSTRACT

This work analyzes the historical evolution of the Brazilian space sector, with emphasis in the area of satellites. Such approach is necessary to create strategies for the future of the mission at strategic, tactical and operational levels. This study, carried out using Data Science and Machine Learning tools, such as the VAR and VARMA multivariate time series analysis models, applied to data available on the internet, in platforms such as Our World in Data, Data.gov, CAPES, among others, together with the knowledge obtained through bibliographic methodology, were used in the analysis of the national space sector of yesterday, today and tomorrow. Thus, it was possible to conclude, through appropriate parameters of the two models discussed, that the HDI, the innovation industry, scientific research, training of researchers and the budget allocated to space policy are extremely important for the sector's growth. In this sense, it was also possible to obtain future forecasts of the time series, and thus determine an equation for the number of satellites in a year "x" for each of the models used. The forecast error obtained in VARMA was smaller than in VAR, however the complexity of manipulating the first model is also higher. Finally, it is noteworthy that the research highlighted the need for the development and consolidation of national space power and the Brazilian Space Program as a way of maintaining national sovereignty, by directing government budget resources to sources that are actually applied in the sector.

Keywords: *Aerospace sector. Satellites. Data-Science. Machine Learning.*

1 INTRODUÇÃO

Tecnologias espaciais são reconhecidas como ferramentas importantes para o progresso científico e econômico das nações. Porém, o real crescimento e desenvolvimento das inovações nessa área tiveram o seu início com o final da Segunda Guerra Mundial e início da Guerra Fria. Até então, a conquista das novas fronteiras do espaço, além da atmosfera terrestre, não se encontrava dentro do escopo das principais pesquisas científicas, que se limitavam aos estudos teóricos da astronomia e da mecânica celeste.

Segundo Matos (2016), o lançamento do satélite Sputnik I ao espaço pela então União Soviética, em 1957, mostrou a possibilidade da realização da guerra no espaço por meio de satélites, tecnologias e veículos espaciais, fomentando a chamada corrida espacial, que teve o seu apogeu em 1969, com a chegada do homem à Lua. Tais eventos provocaram gigantescos impactos científicos, econômicos e sociais e induziram o dispêndio de grandes somas de recursos públicos das superpotências para o custeio da pesquisa e do desenvolvimento da tecnologia empregada.

Apesar da redução da força impulsora da atividade espacial no mundo findo o período da Guerra Fria, o conhecimento agregado e o *know-how* adquiridos possibilitaram a aplicação comercial das tecnologias desenvolvidas, como os satélites geoestacionários e de órbita baixa no ramo das telecomunicações, a observação da Terra através de satélites de sensoriamento remoto (SILVA FILHO, 1999) e a utilização do sistema GPS (*Global Positioning System*) na navegação aérea, marítima ou terrestre.

No Brasil, a Estratégia Nacional de Defesa (END) e a Política Espacial Brasileira, ou Política Nacional de Desenvolvimento Espacial (PNDE), preconizam o uso dual dos satélites, ou seja, militar e civil. Assim, de acordo com a END, cabe a Força Aérea Brasileira, juntamente com a Agência Espacial Brasileira (AEB), organização autárquica do Ministério da Ciência, Tecnologia e Inovação do Governo Federal, e o Instituto de Pesquisas Espaciais (INPE) o provimento da estrutura aeroespacial para as operações das Forças Armadas, gerando simultâneo benefício à sociedade civil, principalmente nas áreas de comunicação, meteorologia, observação da terra, navegação e monitoramento do espaço (BRASIL, 2008).

Deste modo, o presente trabalho aborda a utilização de ferramentas de Ciência de Dados (*Data Science*) para avaliar o avanço histórico do Brasil no setor espacial, de forma a detectar padrões úteis em dados públicos e tecer conclusões que possam auxiliar na tomada de decisões. A extração de conhecimento a partir de dados tem sido uma ferramenta adotada por muitas empresas e pelos setores públicos, visto que as previsões e simulações promovem uma nova visão singular da realidade (OZDEMIR, 2016). Esse tipo de análise é realizado sobre o chamado *Big Data*, que corresponde ao volume de dados não estruturados e de grande variedade circulantes nas redes de informação.

Tal abordagem se justifica pois o setor espacial e suas tecnologias, como satélites de geolocalização e de sensoriamento remoto, se fazem essenciais para a viabilização das inúmeras políticas sociais e de defesa no Brasil, como administração e controle das fronteiras, vigilância da Amazônia, defesa da costa e das reservas de petróleo, inclusão digital de populações geograficamente isoladas, etc. Nesse sentido, Rollemberg (2010) afirma que vislumbrar o Brasil em uma posição de destaque no importante setor espacial envolve extraordinário esforço e dedicação. Para tanto, a pesquisa científica nessa área é de extrema relevância.

É importante ressaltar também a contribuição do trabalho para Força Aérea Brasileira e para comunidade acadêmica. Para FAB, o estudo possibilitará a extração de conhecimento sobre o panorama atual do Brasil no setor espacial, de forma a auxiliar na tomada de decisão das autoridades, no que tange ao Programa Estratégico de Sistemas Espaciais (PESE). Para a comunidade acadêmica, a pesquisa contribuirá para que haja uma disseminação do interesse e do conhecimento relacionados a um setor tão relevante, além de mostrar como ferramentas de *Data Science* e *Machine Learning* podem contribuir nos mais diversos setores.

O objetivo deste estudo é aplicar técnicas de ciência de dados, para avaliar o avanço histórico do Brasil no setor espacial, principalmente no que tange à área de satélites, a fim de obter conclusões e traçar estratégias para o futuro.

Este intento será conseguido mediante revisão bibliográfica acerca do setor espacial e das ferramentas de *Data Science* e *Machine Learning*, seguida de uma pesquisa experimental quantitativa. A pesquisa experimental terá como foco a obtenção e análise de dados, obtidos em plataformas públicas na internet.

2 REVISÃO BIBLIOGRÁFICA

A base bibliográfica deste trabalho será dividida em duas grandes áreas: o setor espacial e a ciência de dados. A primeira será composta por estudos, artigos, legislações e documentos que abordam o setor espacial nacional, como a Estratégia Nacional de Defesa e a Política Espacial Brasileira, com enfoque na área de satélites e uma análise da indústria espacial, tanto no Brasil como em países emergentes. Já a segunda abordará a metodologia que será utilizada para o tratamento dos dados e as ferramentas atuais disponíveis, como a linguagem de programação *Python,* e conceituará *Data Science, Machine Learning* e suas aplicações.

2.1 POLÍTICA ESPACIAL BRASILEIRA

A Estratégia Nacional de Defesa (END) é responsável por estabelecer diretrizes para a adequada preparação e capacitação das Forças Armadas, de modo que seja possível garantir a segurança nacional tanto em tempos de paz, quanto em situações de crise. A END abrange quatro eixos principais, sendo um deles a divisão das atribuições estratégicas entre as Forças Armadas, de modo que cada uma possa melhor desempenhá-las (BRASIL, 2013).

Assim, de acordo com Sá (2015, p.6), através da END, “o Brasil sinaliza sua posição no contexto internacional e indica a necessidade de uma nova e afirmativa postura no campo da soberania”. Uma das estratégias para se adquirir tal postura, prevista no respectivo documento, foi a implementação do Programa Estratégico de Sistemas Espaciais (PESE), cujo planejamento está essencialmente voltado à implantação da infraestrutura fundamental para o cumprimento da END. Pode-se afirmar que: “O PESE resulta das diretrizes estabelecidas na END que orientam as Forças Armadas a empregar o espaço para se tornar mais eficientes em suas operações e para contribuir com o desenvolvimento da indústria espacial brasileira.” (SÁ, 2015, p.8).

Já o Programa Nacional de Atividades Espaciais (PNAE), estruturado em etapas que serão implantadas até o final de 2021, pretende reordenar, de forma integrada, as iniciativas e os investimentos no setor (BRASIL, 2012). Segundo

Andrade (2015), a Agência Espacial Brasileira (AEB) formulou documentos que estabelecem metas desde a consolidação da indústria espacial brasileira até o desenvolvimento e absorção de tecnologias, capacitação de profissionais e parcerias internacionais e privadas.

Nesse sentido, ressalta-se que as atividades espaciais brasileiras são usualmente organizadas em programas, compostos por subprogramas e projetos (BRASIL, 1994). Dentre eles, destaca-se o Satélite Sino-Brasileiro de Recursos Terrestres (CBERS), de cooperação tecnológica entre China e Brasil para a produção de satélites de observação da Terra, e o Satélite Geoestacionário de Defesa e Comunicações Estratégicas (SGDC). Entretanto, apesar da existência de estruturação legal para o desenvolvimento desse tipo de tecnologia no país, o Brasil ainda possui um programa com baixo nível de competitividade internacional (MATOS, 2016).

> Segundo o PNAE, mais de 40 satélites geoestacionários de telecomunicações operam no Brasil, todos estrangeiros e usando artefatos fabricados no exterior. Empresas brasileiras fornecem apenas antenas para estações de controle e material de solo. Apesar disso, a disputa por um lugar no mercado não envolve conceito nem preconceito simplistas: além das questões de Defesa, o Brasil quer participar, com a sua indústria, seus centros tecnológicos e instituições, das novas demandas de serviços espaciais. A construção do Satélite Geoestacionário de Defesa e Comunicações Estratégicas, que terá uma vida útil de 15 anos, significa uma nova fase no contexto tecnológico e industrial do nosso setor espacial. (Andrade, 2015, p. 5).

Pela leitura do trecho acima, é possível aventar a necessidade do estabelecimento de diretrizes mais práticas e frequentemente atualizadas que possibilitem um melhor posicionamento nacional no setor espacial. Ressalta-se que as dificuldades enfrentadas devido à ausência dessas diretrizes não afetam apenas o Brasil, como também outros países em desenvolvimento, conforme será discutido no tópico seguinte.

2.2 INDÚSTRIA ESPACIAL EM PAÍSES EMERGENTES

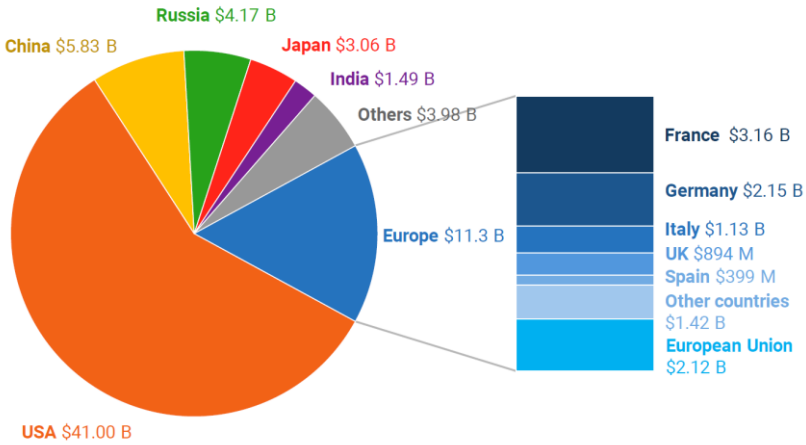
Tecnologias espaciais são importantes ferramentas para o progresso científico e econômico de nações. Isso ocorre porque a indústria espacial lidera um setor de

alta tecnologia que cria grande atividade econômica, fomentando inclusive outras áreas de pesquisa, principalmente em países emergentes. Desse modo, além de melhorar a qualidade de vida dos cidadãos direta e indiretamente, tecnologias espaciais também promovem a ciência (LELOGLU; KOCAOGLAN, 2008).

Por outro lado, nações desenvolvidas e poderosas econômica e cientificamente são capazes de desenvolver sistemas espaciais de ponta de maneira autônoma e sem grandes esforços. Assim, baseando-se no artigo *"Establishing space industry in developing countries: Opportunities and difficulties"* de Leloglu e Kocaoglan (2008), o mecanismo de *feedback* positivo dessas nações é responsável por aumentar significativamente as lacunas já existentes entre os países que possuem capacidade de lançar satélites e missões acima da estratosfera e aqueles que não, criando o que os autores definem como a "divisão estratosférica".

Ainda segundo Leloglu e Kocaoglan (2008), os motivos pelos quais países desenvolvidos apresentam esse *feedback* positivo são inúmeros e vão desde investimento em orçamento específico até capital humano. A exemplo disso, na Figura 01, pode-se verificar, no ano de 2018, a disparidade existente entre o orçamento espacial dos EUA com o dos demais países, em dólares. Ressalta-se que os recursos americanos somavam cerca de 55,89% do total global disponível, o que corrobora com a lacuna evidenciada anteriormente.

Figura 01 – Orçamento espacial institucional por país em 2018.



Fonte: European Space Policy Institute (ESPI) (2020, p.132).

A consolidação de atividades e programas espaciais em um país depende, dentre muitos fatores, de uma indústria espacial forte e capaz de atender às demandas advindas desse tipo de setor. Portanto, com o intuito de determinar esses fatores, bem como as particularidades do desenvolvimento de uma indústria espacial brasileira significativa, serão analisadas a seguir as oportunidades e dificuldades encontradas.

2.2.1 Oportunidades e Dificuldades

O desenvolvimento de um setor espacial próprio em países emergentes como o Brasil apresenta inúmeros benefícios, que foram retratados no estudo de Leloglu e Kocaoglan (2008, p. 3 e 4), sendo os mais relevantes elencados a seguir:

- Com o fomento da ciência e da pesquisa, o poder humano capacitado, utilizado na inovação do setor espacial, conta com uma redução no que se conhece por “fuga de cérebros”. Isso ocorre porque pesquisadores recebem maior incentivo em suas pesquisas e trabalhos, culminando na valorização de profissionais que anteriormente buscavam melhores condições de produção científica em outros países, principalmente aqueles que já estão em um patamar elevado na conquista espacial.
- O derrame tecnológico decorrente do desenvolvimento de novas tecnologias espaciais promove benefícios para a sociedade no geral, como avanços na saúde, educação e economia, principalmente em países emergentes.
- Novas pesquisas e o fomento da área de satélites culminam em melhores serviços em terra, ou *downstream*, como estações de telecomunicações, de sensoriamento remoto e de controle aeroespacial, que são de grande importância para o contexto tanto científico quanto militar;
- A cooperação internacional e regional, essencial para o desenvolvimento das capacidades espaciais, contribui para a paz e a estabilidade globais.

Em contraposição com as oportunidades apresentadas acima, segundo Lall (1992), embora muitos países emergentes estejam determinados em desenvolver sua capacidade espacial, o orçamento alocado por eles ainda é modesto quando

comparado aos orçamentos gastos pelos países desenvolvidos. Desse modo, os recursos disponíveis podem e devem ser gastos, mas de maneira eficiente. Ressalta-se que essa diretriz vai de encontro à ideia globalmente difundida de que países emergentes apresentarão um desempenho inferior, por competirem apenas no mercado doméstico, quando comparado ao de países que já possuem histórico de projetos no setor, justificando assim a sua atuação na área espacial.

Além disso, Leloglu e Kocaoglan (2008, p. 4 e 5) também apontam outras dificuldades enfrentadas pelos países emergentes. Dentre elas podemos citar a falta de mão de obra qualificada e a dificuldade das empresas nacionais em competirem com empresas internacionais sem o incentivo governamental.

Desse modo, uma das maneiras de se garantir a construção e a manutenção de um setor espacial significativo pode ser através do incentivo e fomento aos programas regionais e multilaterais e à indústria espacial nacional, que beneficiam e desenvolvem conhecimentos específicos e as atividades na área. No tópico a seguir será discutida a importância da indústria espacial no contexto apresentado.

2.2.2 Relevância

Segundo Leloglu e Kocaoglan (2008 apud DAHLMAN et al, 1987), combinar tecnologia de fontes externas com componentes domésticos é a melhor forma de se pavimentar o caminho para o crescimento industrial. Isso pode ser afirmado porque importar o *know-how* inicial, já desenvolvido e amadurecido em mercados internacionais, permite o direcionamento dos recursos e insumos nacionais para o objeto final do processo.

Nesse sentido, no contexto da indústria espacial, observa-se que a demanda é puxada pela procura por satélites, que por sua vez se expande ou se contrai de acordo com o comportamento da procura por serviços orbitais, como as telecomunicações. Assim, é importante salientar um conceito definido pela *Satellite Industry Association* (SIA) (2010), que estabelece o termo “indústria de satélites” como um subconjunto de empresas que operam na interseção entre a indústria espacial e a indústria de telecomunicações em geral. Ela é formada por quatro segmentos que são os componentes principais da economia espacial: fabricação de satélites, lançamento,

equipamento de solo e serviços satelitais, conforme apresentado por Schmidt (2012, p. 20). Dessa forma, pode-se aventar que um indicativo eficiente para a avaliação quantitativa da pesquisa e desenvolvimento no setor é a produção e o lançamento de satélites nacionais.

Baseando-se nos conhecimentos acerca das aplicações espaciais abordados nas seções 2.1 e 2.2, aliados aos conceitos de *Data Science* que serão apresentados a seguir, foram realizadas análises da atuação do Brasil nesse setor por meio de programação em *Python* aplicada em dados públicos obtidos na internet.

2.3 *DATA SCIENCE*

Na segunda parte da base bibliográfica do presente estudo, será abordado o conceito de ciência de dados, do inglês*, Data Science*. Segundo Moreira, Carvalho e Horváth (2018) a ciência de dados é responsável pela criação de modelos capazes de extrair padrões de dados complexos, aplicáveis em problemas da vida real. Essa técnica infere conhecimento útil e significativo dos dados com o suporte de tecnologias adequadas.

Nesse sentido, um conceito importante é o de *Machine Learning* ou aprendizado de máquina. O SAS Institute (2021) o define como um método de análise de dados que automatiza a construção de modelos analíticos. Assim, pode ser entendido como um ramo da inteligência artificial, no qual os sistemas aprendem com dados, identificam padrões e tomam decisões com baixa ou nenhuma intervenção humana.

O aprendizado de máquina pode ser dividido em dois tipos: o aprendizado supervisionado e o não supervisionado. No primeiro, utilizado na presente pesquisa, o modelo aprende a partir de resultados pré-definidos, já no segundo, não existem resultados pré-definidos para o modelo utilizar como referência para aprender (MONARD; BARANAUSKAS, 2003).

Existem várias formas de obtenção de análises e de *insights* de dados, como estatísticas descritivas, análises descritivas de múltiplas variáveis, *data quality*, *clustering* ou análise de agrupamento de dados, auto regressão vetorial (VAR), mineração de padrões e *cheat sheet*, como também ferramentas de simulação e

previsão futuras (MOREIRA; CARVALHO; HORVÁTH, 2018). O método a ser utilizado na análise e tratamento dos dados depende dos *datasets* obtidos. Entende-se por *datasets* os conjuntos de dados tabulares que são disponibilizados no domínio público.

Por fim, a linguagem de programação mais utilizada em *Data Science* é o *Python*. Ela é uma linguagem acadêmica, bastante empregada em cursos de matemática e estatística, de tipagem dinâmica, funcional e que tem como base a orientação a objetos. Atualmente, é a mais utilizada em ciência de dados no Brasil, possuindo fácil integração com outras linguagens (VITÓRIA, 2019).

Nos tópicos a seguir, serão abordados alguns dos métodos que foram utilizados na produção das análises deste trabalho e suas especificidades.

2.3.1 Cinco etapas de um projeto de *Data Science*

De acordo com Ozdemir (2016, p. 47), uma das características mais particulares, que caracterizam *Data Science*, quando comparada com *Data Analytics,* por exemplo, é que a ciência de dados segue um processo estruturado passo a passo, que quando aplicado, preserva a integridade e a confiabilidade dos resultados. Assim, cumprir esse rigoroso método permite que até mesmo cientistas de dados amadores obtenham resultados com mais rapidez do que explorando os dados sem uma visão mais clara. É importante ressaltar que apesar de esses passos serem uma lição orientada para pesquisadores novatos, eles também estabelecem as bases para todos os cientistas de dados, até mesmo aqueles nos mais elevados patamares acadêmico e de negócios.

Os cinco passos ou etapas, definidos por Ozdemir (2016, p. 47), para a realização de um projeto de *Data Science* são:

1. Fazer uma pergunta interessante: Nessa fase inicial, deve-se começar como em uma seção de *brainstorming* - escrevendo perguntas independentemente de suas repostas serem factíveis ou não, sem restrições. Esse passo serve para que o pesquisador não siga a tendência de descartar possibilidades importantes para o desenvolvimento da pesquisa, que inicialmente podem não ser claras ou significativas.

2. Obter os dados: Uma vez escolhida(s) a(s) pergunta(s) a ser(em) respondida(s) na pesquisa, nessa fase deve-se fazer uma varredura em busca de dados públicos ou não, que serão utilizados nas análises. Esses dados podem ser obtidos em diversas fontes, e dependem exclusivamente da criatividade do pesquisador.
3. Explorar os dados: Nessa etapa, o cientista de dados já é capaz de conhecer e entender os dados com os quais está lidando, e utiliza programação para manipulá-los e explorá-los, fazendo com que estejam mais adequados para a pesquisa realizada. Assim, já é possível vislumbrar as informações que os dados manipulados são capazes de fornecer.
4. Modelar os dados: Essa etapa envolve a utilização de modelos estatísticos e de *Machine Learning*. Nela, são implantadas as métricas de validação matemática que são responsáveis pela quantificação dos modelos e a avaliação de sua eficácia.
5. Comunicar e visualizar os resultados: Essa é, indiscutivelmente, a etapa mais importante de todo o processo. Apesar de parecer óbvia e simples, a habilidade de concluir os resultados obtidos em um formato de fácil compreensão é mais difícil do que aparenta.

Ao longo da pesquisa realizada neste trabalho, foram aplicados os cinco passos apresentados anteriormente e, além disso, nas etapas de exploração e modelagem dos *datasets*, realizou-se a análise de séries temporais, do inglês *Time Series*, que será abordada no tópico seguinte.

2.3.2 *Time Series*

Time Serie, ou série temporal, é um compilado de observações feitas sequencialmente ao longo do tempo (EHLERS, 2009, p. 1). Uma característica importante desse tipo de dado é que os seus vizinhos apresentam uma dependência lógica entre si, e cabe ao cientista de dados analisá-la e modelá-la.

A maioria dos procedimentos estatísticos utilizados atualmente foram desenvolvidos para analisar *datasets* independentemente de sua evolução no tempo. Assim, é importante ressaltar que a manipulação e exploração de dados variáveis em

séries temporais requer a utilização de técnicas específicas, que serão abordadas no tópico seguinte.

Apesar da complexidade existente em sua utilização e manipulação, elas podem ser muito úteis em diversas áreas do conhecimento, como medicina, meteorologia, ciências naturais, economia, etc. e são empregadas com objetivos descritivos, explicativos, preditivos ou de controle.

Segundo Singh (2018), as séries temporais podem ser classificadas em univariáveis ou multivariáveis. As univariáveis, como o próprio nome sugere, possuem apenas uma única variável dependente do tempo, já as multivariáveis, possuem duas ou mais. Nesse caso, cada variável depende não somente de seus valores anteriores, como também de outras variáveis. Assim, essa dependência é utilizada para prever modelos futuros, exemplificando a aplicação preditiva citada anteriormente.

2.3.3 Modelos de previsão para *Time Series*

Um dos métodos mais comuns e utilizados para previsão em *Time Series* multivariáveis é a Auto-Regressão Vetorial, do inglês*, Vector Auto Regression* (VAR). Em um modelo VAR, cada variável é uma função linear de valores anteriores dela mesma, e os valores anteriores, de todas as outras variáveis (SINGH, 2018). De forma mais profunda, este modelo é composto por *n* equações (representando *n* variáveis endógenas) e inclui *p "lags"* das variáveis (CLOWER, 2021). Ressalta-se que variáveis endógenas são determinadas dentro do modelo.

Como uma extensão do VAR, existe também o VARMA, do inglês*, Vector Autoregression Moving-Average*, ou Auto-Regressão Vetorial de Médias Móveis. Esse método possui maior precisão de previsão quando comparado ao VAR porque ele inclui um ruído branco de todas as variáveis nas estimativas, sendo definido com dois parâmetros: *p* e *q*. O parâmetro *p* é o "*lag"* do modelo VAR e parâmetro *q* indica quantos níveis de ruídos brancos entrarão na equação (LÜTKEPOHL, 2005). Entende-se por ruído branco as variações nos dados que não podem ser explicadas por nenhum modelo de regressão (DATE, 2020).

Como resultado da aplicação dos algoritmos citados, é possível encontrar a interação entre diversas variáveis. Assim, eles são úteis para descrever o

comportamento dinâmico dos dados e também fornecer melhores resultados preditivos.

Na próxima seção, será retratado como foram utilizados os conceitos apresentados ao longo deste artigo, e os materiais e dados utilizados para as análises em *Python*.

3 MATERIAL E MÉTODO

Sendo o escopo deste trabalho a utilização de ferramentas de *Data Science* para a análise de dados espaciais, com um enfoque na posição do Brasil ao longo do tempo, o resultado final obtido será utilizado para que sejam geradas conclusões úteis para a formulação de estratégias que garantam o investimento e a ampliação de estudos e de iniciativas tanto públicas quanto privadas no setor, com ênfase no emprego de satélites.

Assim, a metodologia utilizada para tal intento foi a de pesquisa explicativa quantitativa, empregada de forma bibliográfica e experimental (GIL, 2017). Isso se deu através da aplicação prática dos conhecimentos apresentados anteriormente na revisão bibliográfica, para o desenvolvimento de modelos de análise de dados em Python através de ferramentas de *Data Science* e *Machine Learning*.

O passo a passo do desenvolvimento da pesquisa seguiu as cinco etapas de um projeto de *Data Science*, e serão detalhadas a seguir (com exceção da comunicação e visualização dos resultados, que será abordada na seção 4):

3.1 ELABORAÇÃO DAS PERGUNTAS

Inicialmente foram feitas perguntas relacionadas ao tema que seria abordado, dentre elas: Como o setor espacial evoluiu no Brasil ao longo dos últimos anos? Quais fatores impactaram nessa evolução? Qual a tendência para os próximos anos? O que poderia ser feito para que o Brasil avançasse no setor?

A partir do *brainstorming* realizado, restringiu-se o assunto que seria discutido e analisado, facilitando a busca por dados adequados. Essa etapa teve grande importância na definição concreta do escopo da pesquisa e, ao final, concluiu que o presente estudo analisaria os principais fatores que interferem na produção nacional

de satélites, grande indicativo da atividade espacial brasileira, bem como seu posicionamento estratégico em níveis globais.

3.2 OBTENÇÃO DOS DADOS

Nessa fase do projeto foram definidos os dados e indicadores que seriam mais adequados para utilização como fatores de análise e de comparação em séries temporais. Para tanto, tomou-se como referência principal o trabalho de Leloglu e Kocaoglan (2008), que avalia as dificuldades enfrentadas por países em desenvolvimento no estabelecimento de sua indústria espacial. Assim como foi considerado pelos autores, o número de satélites operado por cada país foi utilizado como variável representativa da atividade espacial. Além disso, como o referido trabalho trata de países em desenvolvimento, o IDH também foi uma das variáveis consideradas. Por fim, como os autores apontam que as maiores dificuldades enfrentadas pelos países estão relacionadas à falta de mão de obra qualificada e à dificuldade das empresas nacionais em competirem com empresas internacionais se não houver incentivo governamental, foram consideradas mais cinco variáveis para compor o conjunto de dados a seguir:

- Número de satélites brasileiros: obtido no *Outer Space Objects Index* (Índice de Objetos do Espaço Sideral), disponibilizado pela *United Nations Office for Outer Space Affairs,* compreendendo o intervalo temporal de 1985 a 2021 (UNITED NATIONS, 2015);
- IDH do Brasil, de 1990 a 2019, disponibilizado pela *Our World in Data* (ROSER, 2014);
- Número de pesquisadores (mestres e doutores) brasileiros: obtido nos dados dos discentes dos programas de pós-graduação *Strictu Sensu* no Brasil, disponibilizado pela Plataforma de Dados Abertos CAPES, compreendendo o intervalo temporal de 2004 a 2020 (COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR, 2021). Esse conjunto de dados foi complementado com o quantitativo de mestres e doutores obtido do artigo “Brasil precisa dobrar número de doutores para atingir o nível mais baixo dos países

desenvolvidos” da Carta Campinas, que cobre o período complementar, de 1996 a 2003 (BRASIL, 2019);

- Empresas que implementaram inovações e receberam apoio do governo, de acordo com a Pesquisa Industrial de Inovação Tecnológica – PINTEC, de 1996 a 2017 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE, 2017a, 2017b);
- Orçamento dos programas 2056 e 2207 do governo federal, respectivamente a Política Espacial e o Programa Espacial Brasileiro, disponibilizados pelo Portal Transparência (BRASIL, 2017a, 2017b).

É importante ressaltar que ao longo da pesquisa, foi observada grande dificuldade em encontrar *datasets* abertos que possuíssem informações englobando um intervalo de tempo abrangente. Em muitos casos, as lacunas temporais de um conjunto de dados não coincidiam com a de outros, gerando “buracos” assimétricos nas planilhas. Apesar do obstáculo inicialmente encontrado, foi possível, por meio da manipulação de alguns dados em CSV (*comma-separated values*) e união de arquivos - passos que serão abordados a seguir - a obtenção de dados úteis.

3.3 EXPLORAÇÃO DOS DADOS

Nessa etapa foram realizadas as manipulações e também tomadas as decisões sobre o que seria feito com os dados incorretos ou incompletos. Todos os *datasets* obtidos tiveram que sofrer conversões de valores, inclusão e exclusão de linhas e colunas e até mesmo mesclagem de fontes antes da sua utilização prática no projeto. Assim, a exploração dos dados é uma segunda etapa de grande importância no método aplicado à projetos de ciência de dados, já que a partir dela os dados brutos começam a ser tratados.

Por exemplo, a lista de satélites brasileiros foi processada de modo a verificar quantos foram lançados em cada ano. Para tanto, criou-se uma programação em *Python* que percorria cada linha do *dataset*, analisando a data de lançamento de cada satélite, excluindo as informações de dia e mês, a fim de verificar o ano do lançamento e aumentar em uma unidade o número de satélites produzidos naquele ano.

Outro tipo de processamento realizado se refere aos dados de discentes dos programas de pós-graduação do Brasil. O *dataset* fornecido pela CAPES é dividido por ano e apresenta a lista de cada discente, com informações como CPF, nível do programa (mestrado ou doutorado), a situação (matriculado ou titulado), entre outras. A contabilização do número de mestres e doutores por ano foi obtida por meio de uma consulta ao *dataset* referente a cada ano, através de um código de importação que pode ser consultado no apêndice F, selecionando a situação “titulado” e o nível do programa em questão. Destaca-se que para se ter uma avaliação mais precisa da quantidade de mão de obra qualificada para o setor espacial, seria necessário verificar a área curso do discente ou o título da dissertação, buscando aqueles com palavras relacionadas ao setor (como satélite, espaço, foguete, etc). Contudo, a avaliação por área é complexa, uma vez que o setor espacial pode receber discentes provenientes de diversas áreas, como Matemática, Física, diversos tipos de Engenharias, entre outros. A avaliação por título da dissertação também não é simples, pois várias palavras poderiam ser consideradas como representativas do setor, sem de fato estarem relacionadas ao espaço, como por exemplo a palavra “motor” ou “telecomunicações”. Assim, optou-se por considerar o número de titulados como um todo, sendo esse o valor utilizado para a quantidade de mão de obra qualificada, assumindo que aqueles que iriam trabalhar no setor espacial seguiriam evolução parecida ao longo do tempo.

De maneira semelhante ao processamento aplicado aos dados de pesquisadores, também foram realizadas algumas considerações para os dados das empresas que implementaram inovações e receberam apoio do governo, disponibilizados pela PINTEC. Não foram filtradas especificamente as empresas do setor espacial, já que esse caso recairia na mesma situação dos pesquisadores e a dificuldade já mencionada em distinguir-se quais áreas ou setores seriam relacionados. Assim, considerou-se que o setor espacial seguiria também o mesmo padrão de evolução ao longo do tempo. Além disso, é importante ressaltar que os dados coletados pela PINTEC possuem periodicidade trienal, desse modo, foi repetido o número das empresas do último ano do triênio nos anos anteriores, o que fez o gráfico adquirir um aspecto de “escadinha”.

Foi verificado que o dado referente ao orçamento da Política Espacial (programa 2056) e do Programa Espacial Brasileiro (programa 2207) era o que abrangia menor intervalo de tempo. Um intervalo de tempo tão pequeno pode prejudicar o uso do aprendizado de máquina, pois o computador não tem dados suficientes para aprender. Dessa forma, uma das opções seria extrapolar os dados para os anos faltantes, com algum tipo de algoritmo de extrapolação, como linear ou polinomial. Contudo, essa estratégia pode criar dados fictícios que não representam a realidade, como é o caso do orçamento da Política Espacial e do Programa Espacial Brasileiro, que só foram incorporados à LOA (Lei Orçamentária Anual) a partir de 2014. Assim, optou-se por utilizar esse dado sem extrapolação, com valores iguais a zero nos anos faltantes.

Todos os dados foram utilizados como variáveis do sistema (endógenas), compondo um *dataframe*. Segundo Torgo (2003), um *dataframe* também pode ser visto como uma tabela gerada a partir de uma base de dados, em que cada linha corresponde a um registro ou linha da base. Já cada coluna corresponde às propriedades, ou campos, a serem armazenadas para cada registro.

Assim, foi criado um *dataframe* vazio, com as colunas representando os dados obtidos, e as linhas constituindo os anos dos quais estes foram associados. Posteriormente, cada coluna foi completada com os arquivos encontrados e foi gerada a tabela nomeada “dfNovo”. Como houve lacunas nos campos Pesquisadores, IDH e Empresas nos anos de 1990 a 1995, 2020 a 2021 e 2018 a 2021 criou-se um novo *dataframe* apenas com as colunas completas, nomeado “dfSelecao”, que pode ser consultado no Apêndice A.

A partir da Tabela A.01 “dfSelecao” os dados puderam ser de fato modelados, conforme será discutido a seguir.

3.4 MODELAGEM DOS DADOS

Para modelagem dos dados, optou-se pelo método VAR e VARMA, por serem indicados para previsão de séries temporais multivariáveis. Uma vez em posse do *dataframe* “dfSelecao”, foram realizados alguns passos sobre os dados tratados, que forneceram informações relevantes e significativas na formulação de resultados.

O primeiro passo foi calcular a correlação de *Pearson*, que indica a interdependência entre variáveis. É importante mencionar que, segundo Salles (2018), a correlação de *Pearson* mede a associação linear entre variáveis contínuas. É o valor que indica o quanto a relação entre as variáveis pode ser descrita por uma reta, variando de -1 a 1. Valores positivos indicam que o aumento de uma variável acompanha o aumento da outra (na média), enquanto valores negativos indicam que à medida que uma variável cresce, a outra decresce.

Em seguida, após obtida a correlação entre os fatores e realizada a impressão dos gráficos de evolução temporal de cada indicativo, foi iniciada a preparação dos dados para aplicar o aprendizado de máquina, ou *Machine Learning*, com a criação do *dataframe* “dfTrain”. Essa tabela é uma seleção dos dados de 1996 a 2016 do *dataframe* “dfSelecao”, os quais foram utilizados para o treinamento ou aprendizado pelos modelos VAR e VARMA.

Os dados do intervalo temporal de 2017 a 2021 foram utilizados para validar o resultado obtido. Ressalta-se que o *dataframe* não estava completo nos anos de 2020 e 2021 (não havia informação sobre IDH, pesquisadores ou empresas). Contudo, isso não foi um problema, uma vez que a única variável de interesse para validação e previsão era o número de satélites, que estava disponível em todo o intervalo de validação.

Em seguida, foi realizada a verificação da estacionariedade das séries temporais, pelo método publicado por Prabhakaran (2019), uma vez que os modelos VAR e VARMA são indicados para séries estacionárias.

O passo seguinte foi aplicar o modelo VAR variando o *lag* entre 1, 2 e 3 e o modelo VARMA variando p de 0 a 2 e q de 0 a 3, não sendo aplicado com ambos parâmetros zerados.

Os modelos testados geraram previsões para as variáveis nos anos de 2017 a 2021, que foram salvos em um novo *dataframe*, chamado de “pred”. O próximo passo foi confrontar os dados reais com os previstos para a variável de interesse (número de satélites) e calcular o Erro Quadrático Médio (EQM).

Por fim, foram impressos os gráficos comparativos de cada dado real com o previsto pelo modelo de melhor resultado (menor EQM) e realizada a previsão de evolução dos dados até 2026. Além disso, também foi possível determinar uma

fórmula regressiva geral capaz de estimar a variação da quantidade de satélites brasileiros em órbita em um determinado ano, prevendo a capacidade aeroespacial nacional.

4 RESULTADOS E DISCUSSÕES

A última etapa prevista no passo a passo para a formulação e execução de um projeto de *Data Science* é a comunicação e visualização dos resultados obtidos. Assim, essa seção compilará as conclusões por meio da aplicação dos algoritmos no código desenvolvido, em *Python.*

Inicialmente foi obtida a correlação de *Pearson* entre os dados analisados: número de satélites, índice de desenvolvimento humano (IDH), quantidade de pesquisadores, empresas e orçamento e percebeu-se que a interação entre as 5 variáveis é positiva, podendo ser observada na Tabela 01:

Tabela 01 – Correlação de *Pearson* das variáveis

Variáveis	Nº Satélites	IDH	Pesquisadores	Empresas	Orçamento
Nº Satélites	1.000000	0.472233	0.525401	0.162508	0.581583
IDH	0.472233	1.000000	0.983478	0.780304	0.670148
Pesquisadores	0.525401	0.983478	1.000000	0.707187	0.735441
Empresas	0.162508	0.780304	0.707187	1.000000	0.316258
Orçamento	0.581583	0.670148	0.735441	0.316258	1.000000

Fonte: O autor

O valor positivo denota que quando as variáveis são analisadas duas a duas, o crescimento de uma implica, na média, no crescimento da outra. Pode-se verificar também que o orçamento é a variável que apresenta maior correlação com o número de satélites.

Apesar da correlação ser usada para verificar a relação das variáveis duas a duas, ela não é útil para prever a interação de todas as variáveis do sistema. Com esse intuito, os modelos VAR e VARMA podem ser aplicados.

Para tanto, como citado anteriormente, foi feito o teste de verificação de estacionariedade das séries temporais, concluindo que apenas a série “Nº de

satélites” era estacionária. Assim, com o objetivo de tornar as demais estacionárias, realizou-se uma transformação de todas as variáveis, criando novas séries temporais com as diferenças entre dois anos subsequentes das séries originais. Por exemplo, na série das diferenças “Nº de satélites”, o valor em 1997 era calculado como o número de satélites em 1997 menos o número de satélites em 1996. Ao testar a estacionariedade nas novas séries, foi constatado que todas elas passaram a ser estacionárias. Assim, os modelos VAR e VARMA puderam ser aplicados, com a ressalva de que os resultados obtidos deveriam passar pela transformada inversa da diferença, para se obter os valores reais. Os apêndices B e C mostram os gráficos das séries temporais antes e depois da transformação realizada, ressaltando que todas as variáveis passaram a ter comportamento estacionário.

Em seguida, após a aplicação dos dois modelos com diferentes parâmetros, foi possível obter o erro quadrático médio (EQM) para a variação do número de satélites, conforme Tabelas D.01 e D.02 (apêndice D).

Por meio das Tabelas D.01 e D.02, foi possível verificar que o melhor modelo testado foi o VARMA com parâmetros *p* igual a 1 e *q* igual a 1, obtendo-se EQM igual a 1.1637028765427473. Tal resultado mostra que os valores das variáveis do ano x são previstos com base nos valores do ano x-1 (parâmetro *p* igual a 1). No modelo VAR, o melhor resultado também foi obtido com *lag* igual a 1, e EQM de 1.3724309808428260. Em seguida, os resultados passaram pela transformada inversa, de modo que foram obtidos os valores das variáveis para cada ano, e não mais a diferença entre eles.

A representação visual comparativa dos valores reais *versus* valores previstos pelos melhores parâmetros dos modelos VAR e VARMA para o número de satélites foi impressa na Figura E.01, do apêndice E. Os gráficos das demais variáveis também se encontram no apêndice citado.

A variável em que a previsão mais se afastou dos dados de validação foi orçamento. Um dos motivos que pode ter contribuído para esse resultado se deve ao fato de que o modelo “aprendeu” com dados de orçamento zerados em muito anos, tendo poucos dados diferentes de zero no conjunto de treinamento (apenas 3 valores). Esse problema leva a erros nas previsões das demais variáveis, já que os valores previstos em 2025, por exemplo, usam como base o orçamento previsto de 2024.

Além disso, o orçamento alocado é determinado pela Lei Orçamentária Anual (LOA), que pode sofrer grandes variações ao longo dos anos devido a fatores econômicos (SIGA BRASIL, 2021). Dessa forma, na previsão da modelagem, o Brasil ficaria estagnado com aproximadamente dois novos satélites por ano a longo prazo.

Além disso, é importante ressaltar que os modelos geram números não inteiros, sendo que a variável número de satélites é inteira. Assim, obteve-se uma estimativa por arredondamento. Para regressões com séries temporais inteiras existem métodos estatísticos ainda mais complexos, como o MINAR (*Multivariate Integer-Valued Autoregressive Model*), porém estes ainda não estão implementados nas API mais conhecidas do *Python*, como é o caso do VAR e VARMA que estão implementados na API *statsmodels*.

Por fim, os modelos geram equações para cada uma das variáveis no ano x. Para a variável de interesse (Nº de Satélites), foi possível obter a fórmula regressiva pelo VAR (1) e pelo VARMA (2):

$$Qs_x = -0.086350 - (0.287650)Qs_{x-1} - (0.001666)IDH_{x-1} + (0.042658)Qp_{x-1} + (0.024671)E_{x-1} + (0.019110)O_{x-1} + \varepsilon Qs_x \quad (1)$$

$$Qs_x = -0.1108 - (0.6043)Qs_{x-1} + (0.0217)IDH_{x-1} - (0.0034)Qp_{x-1} + (0.0813)E_{x-1} + (0.0015)O_{x-1} - (0.8443)\varepsilon Qs_{x-1} - (0.3357)\varepsilon IDH_{x-1} + (0.5183)\varepsilon Qp_{x-1} - (0.5423)\varepsilon E_{x-1} + (0.0277)\varepsilon O_{x-1} + \varepsilon Qs_x \quad (2)$$

Sendo:

Qs_x= Variação da quantidade de satélites no ano x;

IDH_{x-1}= Variação do IDH no ano x-1;

Qp_{x-1}= Variação da quantidade de pesquisadores no ano x-1;

E_{x-1}= Variação da quantidade de empresas no ano x-1;

O_{x-1}= Variação do orçamento no ano x-1;

εQs_{x-1}= Ruído branco da variação da quantidade de satélites no ano x-1;

εIDH_{x-1}= Ruído branco da variação do IDH no ano x-1;

εQp_{x-1}= Ruído branco da variação da quantidade de pesquisadores no ano x-1;

εE_{x-1}= Ruído branco da variação da quantidade de empresas no ano x-1;

εO_{x-1}= Ruído branco da variação do orçamento no ano x-1;
εQs_x= Ruído branco da variação da quantidade de satélites no ano x;
$x \geq 1997$.

Analisando a fórmula (1), do modelo VAR, cuja complexidade é inferior comparada ao VARMA, percebe-se que os coeficientes de Qs_{x-1} e IDH_{x-1} são negativos. Essa característica pode ser percebida empiricamente pois observa-se que se em um ano a quantidade de satélites produzida for muito grande, a tendência no ano seguinte é reduzi-la, mantendo-se a estacionariedade constatada no teste de Prabhakaran (2019). Isso pode ocorrer por motivos como saturação tecnológica, redução orçamentária, entre outros. O aumento no IDH também influencia negativamente na quantidade de satélites produzidos, já que para que o índice aumente, a qualidade de vida da população também se eleva, e isso ocorre principalmente quando há investimento do governo em saúde e educação. Para tanto, mais capital deve ser direcionado para essas áreas, e a curto prazo reduz o orçamento que poderia ser empregado em outros programas do governo, como o 2056 e 2207, respectivamente a Política Espacial e o Programa Espacial Brasileiro.

Já os coeficientes positivos das variáveis Qp_{x-1}, E_{x-1} e O_{x-1} permitem inferir que a equação obtida reflete o conceito da Hélice Tríplice da Inovação no setor espacial. Sendo a quantidade de pesquisadores, de empresas e o orçamento as três variáveis que se interrelacionam para a produção de inovação no setor, elas se encaixam na tríade universidade-indústria-governo, e, segundo Etzkowitz e Zhou (2017, p.24), são a “chave para o crescimento econômico e o desenvolvimento social baseados no conhecimento”, afetando positivamente na quantidade anual de satélites.

Ademais, ainda pela fórmula obtida pelo modelo VAR, para que se possa aumentar o número de satélites em uma unidade do ano de 2021 para 2022, *coeteris paribus*, o governo deveria elevar seu investimento em aproximadamente 87 milhões no orçamento, ou ampliar o investimento em 67 mil empresas, ou seria necessário um incremento em 39 mil pesquisadores. O resultado obtido confirma o conceito da Hélice Tríplice, pois demonstra que para que um dos setores cresça isoladamente, é demandado muito investimento específico. Esse investimento poderia ser melhor

empregado se dissipado pelos três elos, desenvolvendo a área espacial como um todo.

Já a fórmula (2), do modelo VARMA, apesar de gerar resultados com menor EQM, por se tratar de um método mais complexo e que considera vários ruídos brancos, a explicação das variáveis não é tão intuitiva quanto a do VAR. Assim, sua análise não será realizada no escopo deste estudo.

Por fim, sendo as fórmulas Qs_x os produto finais da pesquisa desenvolvida ao longo do presente artigo, foi possível obter um medidor quantitativo para dados cuja relação não poderia ser percebida sem prévia exploração e modelagem.

5 CONSIDERAÇÕES FINAIS

No decorrer dos anos que sucederam a corrida espacial da segunda metade do século XX, muitos foram os investimentos internacionais em pesquisa e desenvolvimento realizados no setor. Entretanto, como apresentado ao longo do artigo em questão, o Brasil, mesmo com diretrizes aeroespaciais bem definidas, seja na END ou na PND, e com programas estruturados, como o PESE, PNAE e PNDE, ainda não possui uma posição de destaque no setor. Tendo como base a problemática retratada, o estudo realizado objetivou analisar dados estratégicos para a aplicação aeroespacial brasileira como instrumento para formulação de diretrizes mais específicas e atuais.

Inicialmente, foi realizada uma revisão bibliográfica, dividida em duas áreas: a espacial e a de ciência de dados. O primeiro assunto foi abordado a fim de observar o avanço obtido nesse campo, tendo sido possível verificar que os maiores entraves para o deslanchar das aplicações tanto civis quanto militares está na disponibilidade de mão-de-obra qualificada e de empresas com capital e tecnologias mais modernas. Além disso, também foi possível aventar que o indicador que seria mais adequado para a mensuração do potencial espacial é a produção de satélites, que representa majoritariamente a indústria espacial e a tecnologia de ponta. O IDH, a quantidade de pesquisadores disponíveis nacionalmente, a quantidade de empresas catalogadas como implementadoras de inovações e o orçamento alocado para este fim, portanto, foram variáveis utilizadas na modelagem e que contribuíram para um resultado final

coerente. Na segunda parte da revisão bibliográfica, foram introduzidos alguns conceitos importantes para a compreensão da metodologia utilizada na pesquisa, como as cinco etapas de um projeto de *Data Science*, séries temporais e os algoritmos VAR e VARMA.

Posteriormente, foram testados os dois modelos apresentados, com diferentes parâmetros, e o que apresentou menor EQM para o número de satélites foi o VARMA. Entretanto, foi possível observar que, no resultado, o número de novos satélites do Brasil a longo prazo ficaria estagnado em aproximadamente dois, não acompanhando o crescimento das outras variáveis. Também foi possível analisar a fórmula gerada para a variação da quantidade de satélites pelo modelo VAR, que possibilitou a inferência de conclusões sofre a influência das demais variáveis no resultado final.

Tendo a hipótese de que a quantidade de satélites do Brasil diz muito sobre sua posição no contexto espacial dual (militar e civil), para que as previsões obtidas sejam ainda mais precisas, poderiam ser utilizados outros conjuntos de dados públicos, como investimento em pesquisa e desenvolvimento, telecomunicações e economia espacial global. Além disso, outra melhoria possível para a exploração e tratamento de dados, seria a diferenciação tanto dos pesquisadores, quanto das empresas apenas nas áreas afetas ao setor espacial. No modelo, estão sendo considerados todos os pesquisadores brasileiros e empresas do PINTEC, o que pode ser um motivo pelo qual a previsão permaneceu estagnada ao longo dos anos. Ademais, a obtenção de dados com intervalos temporais maiores, também permitiriam um melhor aprendizado de máquina, que resultaria em um modelo mais fidedigno. Tais lacunas observadas na produção da pesquisa são suscitadoras de novos estudos e podem ser aproveitadas como complementares a este.

Por fim, é oportuno ressaltar que a tendência de crescimento no setor espacial prevista para o brasil até 2026 não é otimista. Entretanto, ao tomarmos conhecimento acerca de quais são os fatores e variáveis mais relevantes para tal crescimento – obtidos através da fórmula Qs_x – é possível estabelecer diretrizes a níveis estratégico, tático e operacional específicas. Além disso, também é factível direcionar os recursos orçamentários governamentais para fontes que de fato são aplicadas no setor espacial, principalmente na tríade universidade-indústria-governo, elevando o Brasil

da posição de mero consumidor para inovador e desenvolvedor de tecnologias espaciais de vanguarda.

REFERÊNCIAS

ANDRADE, Umberto. Satélites Brasileiros: por que o Brasil precisa ocupar seu lugar no espaço. **Revista ADESG**: defesa e desenvolvimento, Rio de Janeiro, ano 40, ed. 289, p. 4-5, jan./abr. 2015. Disponível em: http://adesg.org.br/wp-content/uploads/2018/08/Revista-ADESG-289-Satelites-Brasileiros.pdf. Acesso em: 29 set. 2020.

BRASIL. Controladoria-Geral Da União. Política espacial. *In*: **Portal da Transparência**. [*S. l.*], 2017a. Disponível em: http://www.portaltransparencia.gov.br/programas-e-acoes/programa-orcamentario/2056-politica-espacial. Acesso em: 22 abr. 2021.

BRASIL. Controladoria-Geral Da União. Política espacial. *In*: **Portal da Transparência**. [*S. l.*], 2017b. Disponível em: http://www.portaltransparencia.gov.br/programas-e-acoes/programa-orcamentario/2207-programa-espacial-brasileiro. Acesso em: 22 abr. 2021.

BRASIL. **Decreto nº 1.332, de 8 de dezembro de 1994**. Aprova a atualização da Política de Desenvolvimento das Atividades Espaciais - PNDAE. [*S. l.*], 1994. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/1990-1994/d1332.htm#:~:text=A%20Pol%C3%ADtica%20Nacional%20de%20Desenvolvimento,em%20benef%C3%ADcio%20da%20sociedade%20brasileira. Acesso em: 29 set. 2020.

BRASIL. **Decreto nº 6.703, de 18 de dezembro de 2008**. Aprova a Estratégia Nacional de Defesa, e dá outras providências. Brasília: Subchefia para Assuntos Jurídicos, 2008. Disponível em: https://www.gov.br/defesa/pt-br/assuntos/copy_of_estado-e-defesa/pnd_end_congresso_.pdf. Acesso em: 30 ago. 2020.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. Agência Espacial Brasileira. 2012. **Programa Nacional de Atividades Espaciais**: PNAE: 2012 - 2021, Brasília, 2012. Disponível em: http://antigo.aeb.gov.br/wp-content/uploads/2018/05/PNAE-Portugues.pdf. Acesso em: 28 set. 2020.

BRASIL. Ministério da Defesa. Estratégia Nacional de Defesa. *In*: **Governo Federal**: Ministério da Defesa. [*S. l.*], 3 dez. 2013. Atualizado em 24/07/2020. Disponível em: https://www.gov.br/defesa/pt-br/assuntos/copy_of_estado-e-defesa/estrategia-nacional-de-defesa. Acesso em: 28 set. 2020.

BRASIL precisa dobrar número de doutores para atingir o nível mais baixo dos países desenvolvidos. *In*: **Carta Campinas**. [*S. l.*], 24 maio 2019. Disponível em: https://cartacampinas.com.br/2019/05/brasil-precisa-dobrar-numero-de-doutores-

para-atingir-o-nivel-mais-baixo-dos-paises-desenvolvidos/. Acesso em: 13 ago. 2021.

CLOWER, Erica. Introduction to the fundamentals of vector autoregressive models. *In*: **Aptech Systems**. [*S. l.*], 15 abr. 2021. Disponível em: https://www.aptech.com/blog/introduction-to-the-fundamentals-of-vector-autoregressive-models/. Acesso em: 19 ago. 2021.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. Conjuntos de dados. *In*: **Dados abertos CAPES**. [*S. l.*], 2021. Disponível em: https://dadosabertos.capes.gov.br/dataset. Acesso em: 23 jul. 2021.

DATE, Sachin. The white noise model. *In:* TDS Editors. **Towards Data Science.** Canadá, 28 ago. 2020. Disponível em: https://towardsdatascience.com/the-white-noise-model-1388dbd0a7d. Acesso em: 01 nov. 2021.

EHLERS, Ricardo S. **Análise de séries temporais**. 5. ed. atual. [*S. l.*: *s. n.*], 2009. 114 p. Disponível em: https://sites.icmc.usp.br/ehlers/stemp/stemp.pdf. Acesso em: 25 jul. 2021.

ETZKOWITZ, Henry; ZHOU, Chunyan. Hélice tríplice: inovação e empreendedorismo universidade-indústria-governo. **Estudos Avançados**, [*s. l.*], ano 90, n. 31, p. 23-48, maio/ago. 2017. DOI 10.1590/s0103-40142017.3190003. Disponível em: https://doi.org/10.1590/s0103-40142017.3190003. Acesso em: 29 ago. 2021.

EUROPEAN SPACE POLICY INSTITUTE (ESPI). Economy & Business. *In:* **ESPI Yearbook 2019**: space policies, issues and trends. Schwarzenbergplatz, Vienna, Austria: [s. n.], May 2020. ISBN 2076-6688. Disponível em: https://espi.or.at/downloads/send/79-espi-yearbook/510-espi-yearbook-2019. Acesso em: 17 maio 2021.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017. 173 p. ISBN 978-85-97-01261-3.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Pesquisa de Inovação: tabela 5924 - empresas que implementaram inovações, que receberam apoio do governo, por tipo de programa de apoio do governo e atividades da indústria, do setor de eletricidade e gás e dos serviços selecionados. *In*: **Sistema IBGE de Recuperação Automática - SIDRA**: pesquisa de inovação - PINTEC. [*S. l.*], 2017a. Disponível em: https://sidra.ibge.gov.br/tabela/5924#notas-tabela. Acesso em: 6 ago. 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Pesquisa de Inovação: tabela 6492 - empresas que implementaram inovações, que receberam apoio do governo, por tipo de programa de apoio do governo e atividades da indústria e dos serviços selecionados (CNAE 1.0). *In*: **Sistema IBGE de Recuperação Automática - SIDRA**: pesquisa de inovação - PINTEC. [*S. l.*], 2017b.

Disponível em: https://sidra.ibge.gov.br/tabela/6492#notas-tabela. Acesso em: 6 ago. 2021.

LALL, Sanjaya. Technological capabilities and industrialization. **World Development**, Institute of Economics and Statistics, Oxford, USA, v. 20, ed. 2, p. 165-186, 22 mar. 1992.

LELOGLU, U. M.; KOCAOGLAN, E. Establishing space industry in developing countries: opportunities and difficulties. **Advances in Space Research**, Science Direct, v. 42, p. 1879-1886, 2008.

LÜTKEPOHL, H. Vector autoregressive moving average processes. *In*: **New Introduction to Multiple Time Series Analysis**. 1. ed. Berlin, Heidelberg: Springer, 2005. cap. 11, p. 419-446. Disponível em: https://doi.org/10.1007/978-3-540-27752-1_11. Acesso em: 28 ago. 2021.

MATOS, P. D. O. Sistemas espaciais voltados para defesa. *In:* (IPEA), I. D. P. E. A. **Mapeamento da base industrial de defesa**. *[S. l.]:* [s.n.], 2016. Cap. 7, p. 509-595.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MOREIRA, J. M.; CARVALHO, A.; HORVÁTH, T. **A general introduction to data analytics.** Hoboken, NJ: Wiley, 2018.

OZDEMIR, Sinan. **Principles of data science.** Birmingham: Packt Publishing, 2016.

PRABHAKARAN, Selva. Vector Autoregression (VAR): comprehensive guide with examples in Python. *In*: **Machine Learning +**. [*S. l.*], 7 jul. 2019. Disponível em: https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/. Acesso em: 27 ago. 2021.

ROLLEMBERG, R. Relatório. *In:* **A Política Espacial Brasileira**. Câmara dos Deputados. Brasília, p. 29. 2010.

ROSER, Max. Human Development Index (HDI). *In*: **Our world in data**. [*S. l.*], 2014. Disponível em: https://ourworldindata.org/human-development-index. Acesso em: 22 abr. 2021.

SÁ, Carlos Minelli de. O Brasil conquistando o espaço. **Revista ADESG**: defesa e desenvolvimento, Rio de Janeiro, ano 40, ed. 289, p. 6-9, Janeiro/Abril 2015. Disponível em: http://adesg.org.br/wp-content/uploads/2018/08/Revista-ADESG-289-Satelites-Brasileiros.pdf. Acesso em: 29 set. 2020.

SALLES, Rodrigo. Correlação: direto ao ponto. *In*: **BrData**. [*S. l.*], 7 jun. 2018. Disponível em: https://medium.com/brdata/correla%C3%A7%C3%A3o-direto-ao-ponto-9ec1d48735fb. Acesso em: 26 jul. 2021.

SAS INSTITUTE. Machine Learning: o que é e qual sua importância?. *In*: **SAS Institute**. [*S. l.*], 16 jul. 2021. Disponível em: https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Acesso em: 15 ago. 2021.

SATELLITE INDUSTRY ASSOCIATION (SIA). **2010 State of the Satellite Industry Report**. Washington, DC, Aug. 2010. Disponível em: https://sia.org/PDF/2011%20State%20of%20Satellite%20Industry%20Report%20(June%202011).pdf. Acesso em: 10 out. 2020.

SCHMIDT, Flávia de Holanda. Desafios e oportunidades para uma indústria espacial emergente: o caso do Brasil. **Discussion Papers 1667**, Instituto de Pesquisa Econômica Aplicada – IPEA, 2011.

SIGA BRASIL. Painel do orçamento anual: principais informações de acompanhamento da Lei Orçamentária do setor espacial brasileiro no ano vigente. *In*: AGÊNCIA ESPACIAL BRASILEIRA. **Observatório do setor espacial brasileiro**. [*S. l.*], 10 ago. 2021. Disponível em: https://observatorio.aeb.gov.br/dados-e-indicadores/tema-governo/tema-orcamento/acompanhamento-da-loa-vigente. Acesso em: 20 ago. 2021.

SILVA FILHO, Durval Henriques da. Considerações sobre a comercialização do Centro de Lançamento de Alcântara. **Espaço e desenvolvimento**, [*s. l.*], ed. 7, p. 75-85, Outubro 1999. Disponível em: http://seer.cgee.org.br/index.php/parcerias_estrategicas/article/viewFile/81/74. Acesso em: 28 set. 2020.

SINGH, Aishwarya. **A multivariate time series guide to forecasting and modeling (with Python codes)**. Analytics Vidhya, 27 set. 2018. Disponível em: https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-*python*-codes/. Acesso em: 20 abr. 2021.

TORGO, Luis. Data frames. *In*: **Departamento de Ciência de Computadores**: Faculdade de Ciências da Universidade do Porto. [*S. l.*], 3 out. 2003. Disponível em: https://www.dcc.fc.up.pt/~ltorgo/SebentaR/HTML/node16.html#:~:text=Um%20data%20frame%20%C3%A9%20semelhante,registo%20(linha)%20da%20tabela.&text=notas%20%C3%A9%20um%20data%20frame,de%20dados%2C%20neste%20momento). Acesso em: 15 ago. 2021.

UNITED NATIONS. Outer space objects index. *In*: **United Nations Office for Outer Space Affairs**. [*S. l.*], 2015. Disponível em: https://www.unoosa.org/oosa/osoindex/search-ng.jspx?lf_id=. Acesso em: 7 maio 2021.

VITÓRIA, Penélope. Qual a melhor linguagem para ciência de dados? *In*: **IMasters**. [*S. l.*], 25 fev. 2019. Disponível em: https://imasters.com.br/data/qual-melhor-linguagem-para-ciencia-de-dados. Acesso em: 29 set. 2020.

APÊNDICE A – Tabela “dfSelecao”

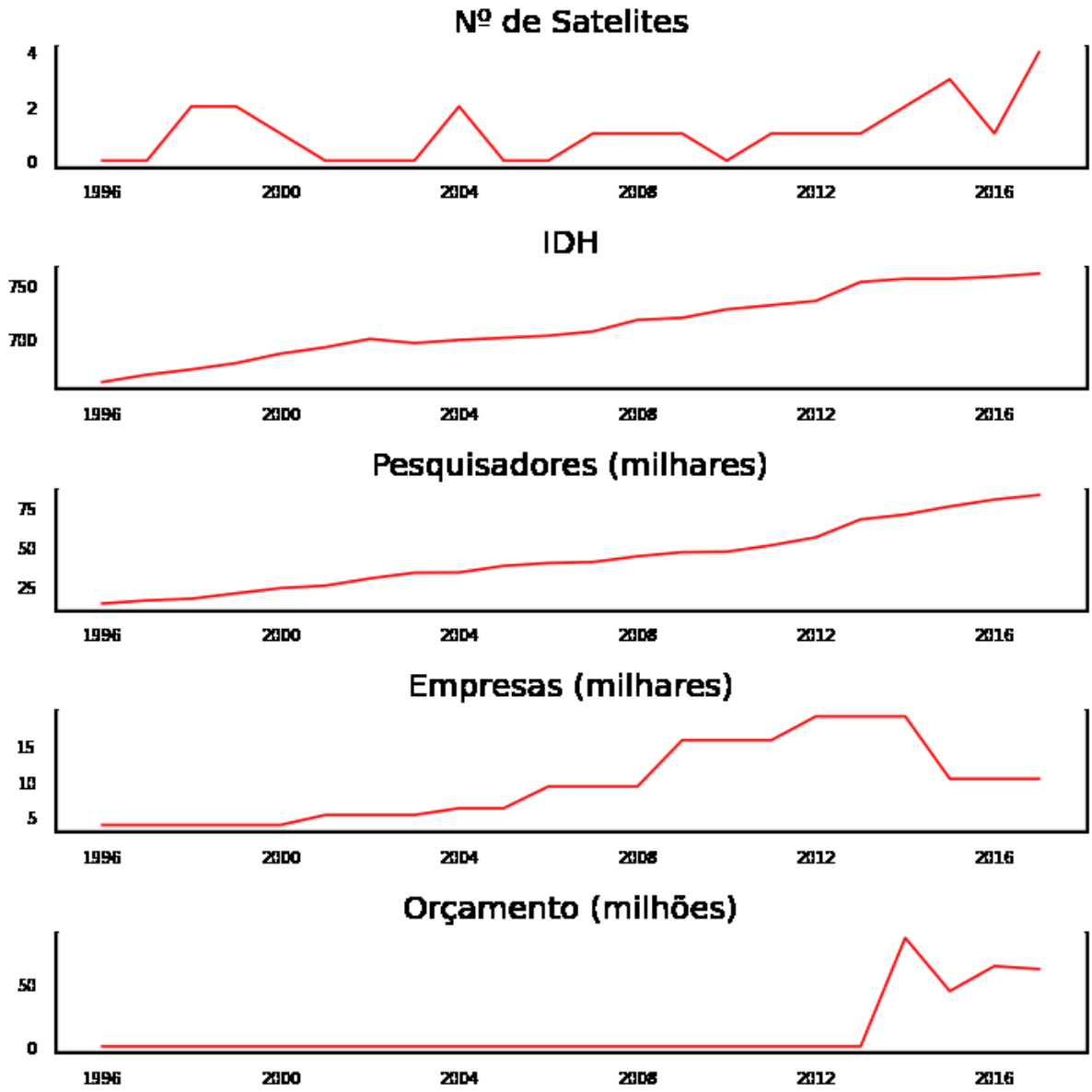
Tabela A.01 – Tabela “dfSelecao”.

Ano	Nº Satélites	IDH	Pesquisadores (Milhares)	Empresas (Milhares)	Orçamento (Milhões)
1996	0.0	658.0	13.336	3.831	0.000000
1997	0.0	665.0	15.314	3.831	0.000000
1998	2.0	670.0	16.432	3.831	0.000000
1999	2.0	676.0	19.925	3.831	0.000000
2000	1.0	685.0	23.270	3.831	0.000000
2001	0.0	691.0	24.838	5.233	0.000000
2002	0.0	699.0	29.491	5.233	0.000000
2003	0.0	695.0	33.177	5.233	0.000000
2004	2.0	698.0	33.365	6.169	0.000000
2005	0.0	700.0	37.589	6.169	0.000000
2006	0.0	702.0	39.408	9.214	0.000000
2007	1.0	706.0	40.082	9.214	0.000000
2008	1.0	717.0	43.751	9.214	0.000000
2009	1.0	719.0	46.366	15.696	0.000000
2010	0.0	727.0	46.704	15.696	0.000000
2011	1.0	731.0	50.856	15.696	0.000000
2012	1.0	735.0	55.998	19.029	0.000000
2013	1.0	753.0	67.534	19.029	0.000000
2014	2.0	756.0	70.668	19.029	86.314208
2015	3.0	756.0	75.885	10.290	43.917288
2016	1.0	758.0	80.278	10.290	63.830993
2017	4.0	761.0	83.281	10.290	61.490709

Fonte: O autor.

APÊNDICE B – Gráficos das séries temporais antes da transformação das diferenças

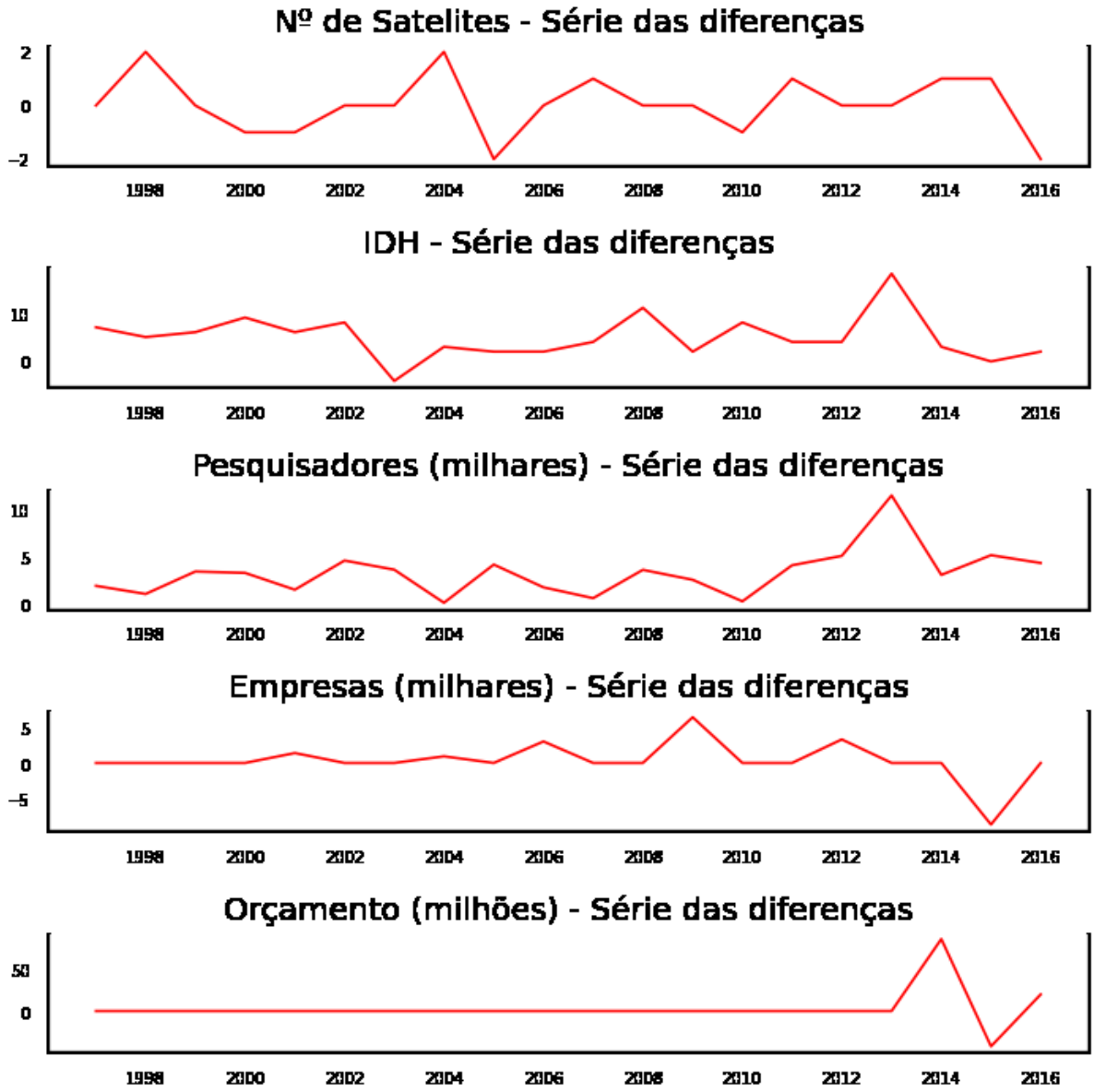
Figura B.01 – Séries temporais antes da transformação das diferenças entre dois anos subsequentes das séries originais.



Fonte: O autor.

APÊNDICE C – Gráficos das séries temporais após a transformação das diferenças

Figura C.01 – Séries temporais após a transformação das diferenças entre dois anos subsequentes das séries originais.



Fonte: O autor.

APÊNDICE D – Erros Quadráticos Médios (EQM)

Tabela D.01 – EQM obtido pelo método VAR.

Lag	Erro
1	1.3724309808428260
2	2.0044090983996417
3	3.9465172633207946

Fonte: O autor.

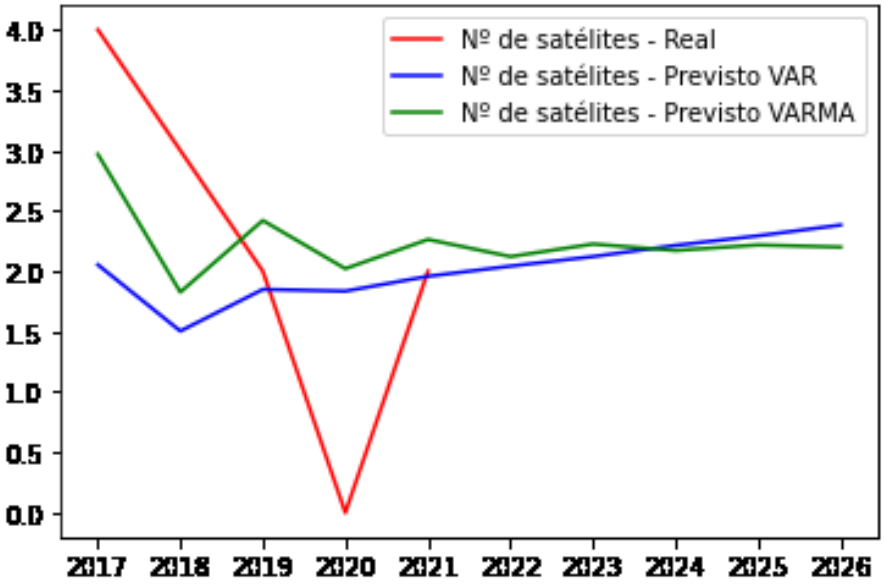
Tabela D.02 – EQM obtido pelo método VARMA.

Parâmetro *p* (*lag*)	Parâmetro *q*	Erro
0	1	1.4130797190080693
0	2	1.7721947856056148
0	3	2.1219106776036725
1	0	1.3770381597081125
1	1	1.1637028765427473
1	2	1.2693347016512142
1	3	2.1528705846913760
2	0	2.3963744590222590
2	1	2.4764491944694180
2	2	2.0904615260268558
2	3	2.4315053773478827

Fonte: O autor.

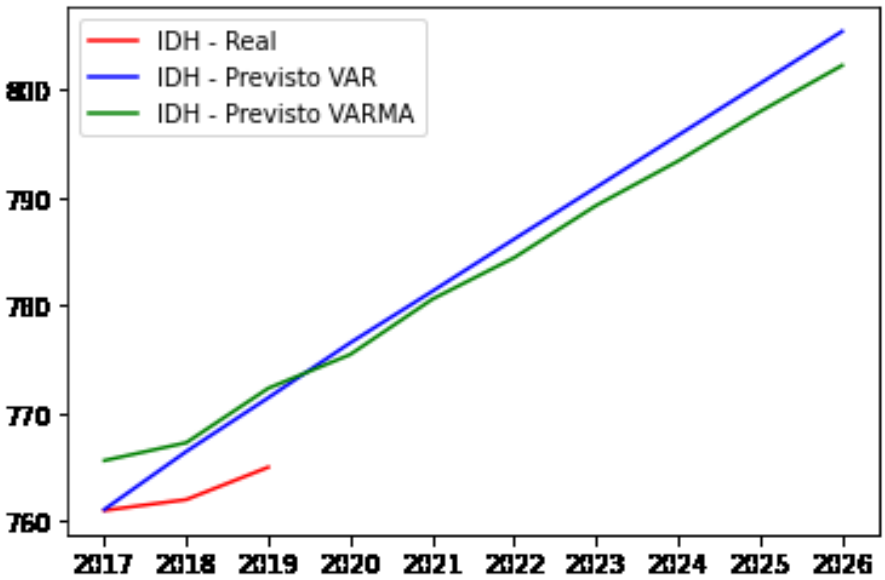
APÊNDICE E – Resultados pelos algoritmos VAR e VARMA

Figura E.01 – Número de satélites real *versus* previsto pelos algoritmos VAR e VARMA.



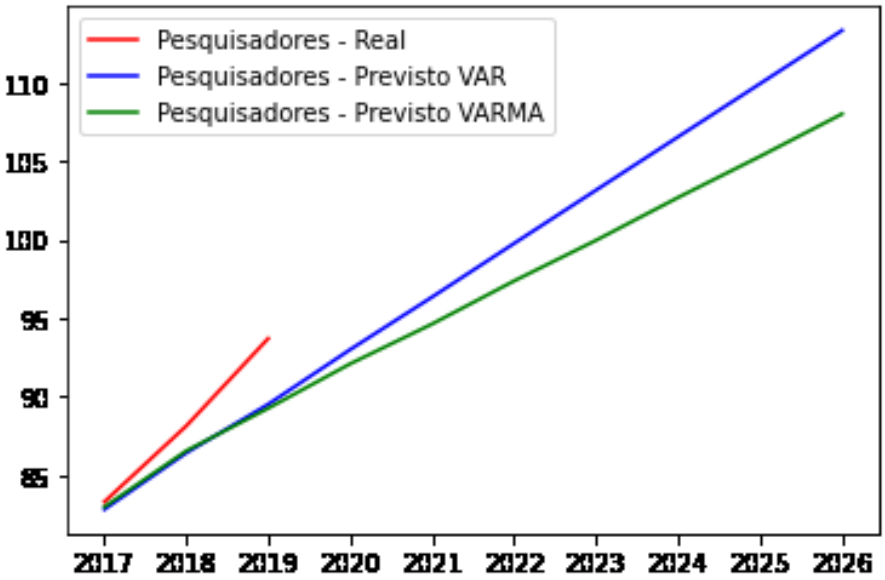
Fonte: O autor.

Figura E.02 – IDH real *versus* previsto pelos algoritmos VAR e VARMA.



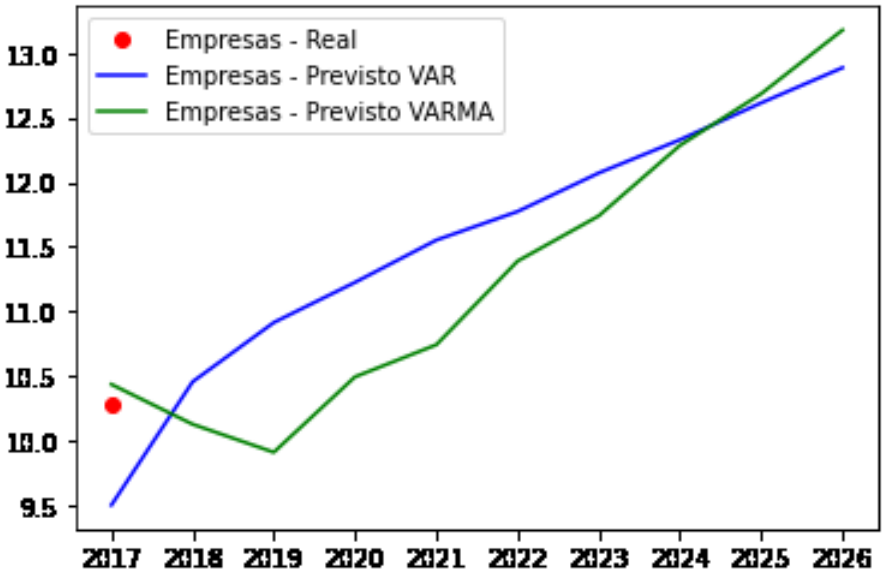
Fonte: O autor.

Figura E.03 – Quantidade de pesquisadores real *versus* prevista pelos algoritmos VAR e VARMA



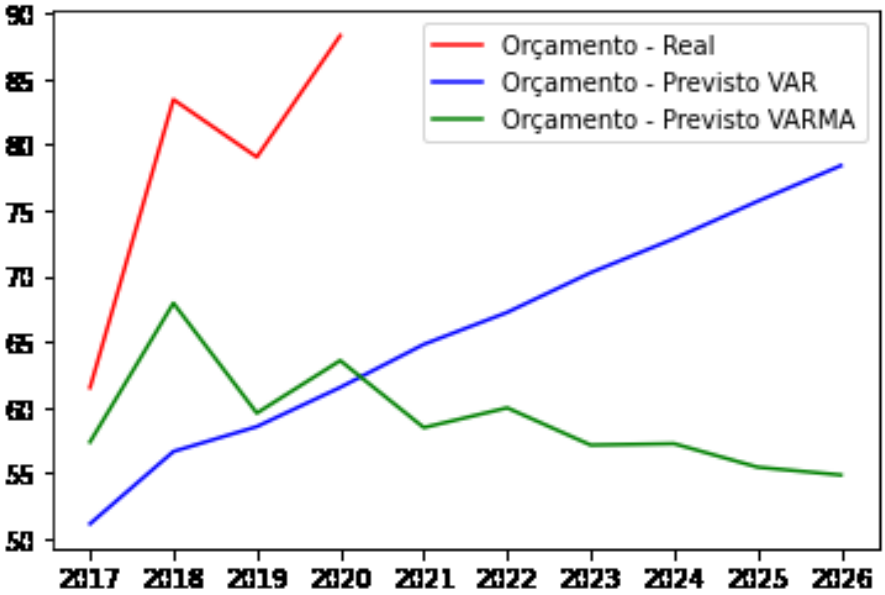
Fonte: O autor.

Figura E.04 – Quantidade de empresas real *versus* prevista pelos algoritmos VAR e VARMA.



Fonte: O autor.

Figura E.05 – Orçamento real *versus* previsto pelos algoritmos VAR e VARMA.



Fonte: O autor.

APÊNDICE F – Códigos do projeto em *Python*

Código geral do projeto:

In [1]:

```
import datetime
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from math import *
from sklearn.metrics import mean_squared_error
import warnings
warnings.filterwarnings('ignore')

low_memory=False
%matplotlib inline
pd.options.display.max_columns = 80
pd.options.display.max_rows = 90

filename = 'satelitesDoBrasil.csv'
#arquivo retirado de https://www.unoosa.org/oosa/osoindex/index.jspx?lf_id=
df = pd.read_csv(filename, sep=';')
df.rename(columns={'Date of Launch':'DateOfLaunch'}, inplace=True)
print(df)

filename2 = 'IDH - Brasil.csv'
dfIDH = pd.read_csv(filename2, sep=";")
print(dfIDH)



filename3 = 'mestresDoutores.csv'
dfMestresDoutores = pd.read_csv(filename3, sep=';')
print(dfMestresDoutores)



filename4 = 'OrcamentoPoliticaEspacial.csv'
dfOrcPolitico = pd.read_csv(filename4, sep=';')
print(dfOrcPolitico)


filename5 = 'pintec.csv'
dfPintec = pd.read_csv(filename5, sep=';')
print(dfPintec)
```

In [2]:

```
linhas=[]
for i in range(32):
    linhas.append(datetime.datetime(1990+i,1,1))#primeiro ano com informação
(IDH - Brasil)
    i=i+1
```

```
columns_names=['N° de Satelites','IDH','Pesquisadores (milhares)','Empresas
(milhares)','Orçamento (milhões)']

dfNovo = pd.DataFrame(index=linhas,columns=columns_names)
dfNovo
```

In [3]:

```
dfNovo=dfNovo.replace(np.nan,0.0)
for i in range(len(dfNovo)):
    aux=dfNovo.index[i]
    ano=aux.year
    for j in range(len(df)):
        dataLancamento=df.loc[j]['DateOfLaunch']
        anoLancamento=int(dataLancamento.split("/")[2])
        if(anoLancamento==(ano)):
            dfNovo.loc[aux,'N°    de    Satelites']=dfNovo.loc[aux,'N°    de
Satelites']+1

    for j in range(len(dfIDH)):
        if((dfIDH.loc[j]['Date'])==ano):
            dfNovo.loc[aux,'IDH']=dfIDH.loc[j,'HDI']

    for j in range(len(dfMestresDoutores)):
        if((dfMestresDoutores.loc[j]['Date'])==ano):
            #número de pesquisadores em milhares
            dfNovo.loc[aux,'Pesquisadores
(milhares)']=((dfMestresDoutores.loc[j,'Mestres']+dfMestresDoutores.loc[j,'
Doutores']))/1000

    for j in range(len(dfPintec)):
        if((dfPintec.loc[j]['Date'])==ano):
            #valor de empresas em milhares
            dfNovo.loc[aux,'Empresas  (milhares)']=dfPintec.loc[j,'Empresas
trienio']/1000

    for j in range(len(dfOrcPolitico)):
        if((dfOrcPolitico.loc[j]['Date'])==ano):
            valor=dfOrcPolitico.loc[j,'VALOR TOTAL PAGO NO ANO']
            valor=valor.replace(',','.')
            #valor do orçamento em milhões de reais
            dfNovo.loc[aux,'Orçamento (milhões)']=float(valor)/1000000
            dfOrcPolitico.loc[j,'VALOR        TOTAL        PAGO        NO
ANO']=float(valor)/1000000

print(dfNovo)
```

In [4]:

```
#selecionar dados completos de 1996 a 2017
dfSelecao=dfNovo[6:28]
dfSelecao
```

In [5]:

```
print(dfSelecao.corr())
```

In [6]:

```
# Plot
fig, axes = plt.subplots(nrows=5, ncols=1, dpi=120, figsize=(6,6))
for i, ax in enumerate(axes.flatten()):
    data = dfSelecao[dfSelecao.columns[i]]
    ax.plot(data, color='red', linewidth=1)
    # Decorations
    ax.set_title(dfSelecao.columns[i])
    ax.xaxis.set_ticks_position('none')
    ax.yaxis.set_ticks_position('none')
    ax.spines["top"].set_alpha(0)
    ax.tick_params(labelsize=6)

plt.tight_layout();
```

In [7]:

```
#treinar até 2016 e validar de 2017 a 2021
#como a variável de interesse para se fazer a previsão é o número de satélites,
não há problemas em não ter dados das demais variáveis em 2020 e 2021
train = dfNovo[6:27]
valid = dfNovo[27:32]
```

In [8]:

```
#método para testar se a série é estacionária, uma vez que os modelos usados
para fazer previsão são indicados para séries estacionárias
#retirado de https://www.machinelearningplus.com/time-series/vector-
autoregression-examples-python/
from statsmodels.tsa.stattools import adfuller

def adfuller_test(series, signif=0.05, name='', verbose=False):
    """Perform ADFuller to test for Stationarity of given series and print
report"""
    r = adfuller(series, autolag='AIC')
    output = {'test_statistic':round(r[0], 4), 'pvalue':round(r[1], 4),
'n_lags':round(r[2], 4), 'n_obs':r[3]}
    p_value = output['pvalue']
    def adjust(val, length= 6): return str(val).ljust(length)

    # Print Summary
    print(f'    Augmented Dickey-Fuller Test on "{name}"', "\n   ", '-'*47)
    print(f' Null Hypothesis: Data has unit root. Non-Stationary.')
    print(f' Significance Level    = {signif}')
    print(f' Test Statistic        = {output["test_statistic"]}')
    print(f' No. Lags Chosen       = {output["n_lags"]}')

    for key,val in r[4].items():
        print(f' Critical value {adjust(key)} = {round(val, 3)}')

    if p_value <= signif:
        print(f" => P-Value = {p_value}. Rejecting Null Hypothesis.")
```

```
        print(f" => Series is Stationary.")
    else:
        print(f" => P-Value = {p_value}. Weak evidence to reject the Null
Hypothesis.")
        print(f" => Series is Non-Stationary.")
```

In [9]:

```
#testar ser a série atual é estacionária (o resultado mostrou que apenas a
série Nº de satélites é estacionária)
for name, column in train.iteritems():
    adfuller_test(column, name=column.name)
    print('\n')
```

In [10]:

```
#testar ser a série de diferenças é estacionária
trainDiff=train.diff().dropna()

for name, column in trainDiff.iteritems():
    adfuller_test(column, name=column.name+'Diff')
    print('\n')
```

In [11]:

```
# Plot
fig, axes = plt.subplots(nrows=5, ncols=1, dpi=120, figsize=(6,6))
for i, ax in enumerate(axes.flatten()):
    data = trainDiff[dfSelecao.columns[i]]
    ax.plot(data, color='red', linewidth=1)
    # Decorations
    ax.set_title(trainDiff.columns[i]+" - Série das diferenças")
    ax.xaxis.set_ticks_position('none')
    ax.yaxis.set_ticks_position('none')
    ax.spines["top"].set_alpha(0)
    ax.tick_params(labelsize=6)

plt.tight_layout();
```

In [12]:

```
#método para fazer a transformada inversa no resultado, ou seja, voltar das
séries de diferenças para série original
def invert_transformation(df_train, df_forecast):
    """Revert back the differencing to get the forecast to original scale."""
    df_fc = df_forecast.copy()
    columns = df_train.columns
    for col in columns:
        df_fc[str(col)+'_forecast'] = df_train[col].iloc[-1] +
df_fc[str(col)].cumsum()
    return df_fc
```

In [13]:

```
#testar o modelo VAR
from statsmodels.tsa.vector_ar.var_model import VAR
```

```
model = VAR(endog=trainDiff)
#testar o modelo com diferentes lags e verificar qual leva ao menor erro
for lag in range (1,4):
    model_fit = model.fit(maxlags=lag)
    # fazer a previsão considerando os dados de validação
    prediction = model_fit.forecast(model_fit.y, steps=len(valid))
    pred = pd.DataFrame(index=valid.index,columns=valid.columns)
    for j in range(len(valid.columns)):
        for i in range(0, len(valid)):
           pred.iloc[i][j] = prediction[i][j]

    df_results = invert_transformation(train, pred)
    #erro quadratico medio da previsao da coluna satélites com o modelo
testado
    erro=sqrt(mean_squared_error(df_results.iloc[:,5], valid.iloc[:,0]))
    print(f'LAG={lag}, ERRO={erro}')
```

In [14]:

```
# detalhamento do melhor resultado do modelo VAR
model = VAR(endog=trainDiff)
model_fit = model.fit(maxlags=1)
print(model_fit.summary())
```

In [15]:

```
#fazer a previsão com melhor resultado do modelo VAR
prediction = model_fit.forecast(model_fit.y, steps=10)
linhas=[]
for i in range(2017,2027):
    linhas.append(datetime.datetime(i,1,1))
pred = pd.DataFrame(index=linhas,columns=valid.columns)
for j in range(len(pred.columns)):
    for i in range(0, len(pred.index)):
        pred.iloc[i][j] = prediction[i][j]

df_results = invert_transformation(train, pred)
data1 = valid[valid.columns[0]]
data2 = df_results[df_results.columns[5]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['N° de satélites - Real','N° de satélites - Previsto'])
```

In [16]:

```
data1 = valid.iloc[0:3,1]
data2 = df_results[df_results.columns[6]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['IDH - Real','IDH - Previsto'])
```

In [17]:

```
data1 = valid.iloc[0:3,2]
data2 = df_results[df_results.columns[7]]
plt.plot(data1,'red',data2,'blue')
```

```
plt.legend(['Pesquisadores - Real','Pesquisadores - Previsto'])
```

In [18]:

```
data1 = valid.iloc[0:1,3]
data2 = df_results[df_results.columns[8]]
plt.plot(data1,'ro',data2,'blue')
plt.legend(['Empresas - Real','Empresas - Previsto'])
```

In [19]:

```
data1 = valid.iloc[0:4,4]
data2 = df_results[df_results.columns[9]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['Orçamento - Real','Orçamento - Previsto'])
```

In [20]:

```
#testar o modelo VARMAX
for p in range(0,3):
    for q in range(0,4):
        if(p+q!=0 and p+q!=6):
            model = sm.tsa.VARMAX(endog=trainDiff, order=(p,q))
            model_fit = model.fit()
            # fazer a previsão considerando os dados de validação
            prediction = model_fit.forecast(steps=len(valid))
            pred = pd.DataFrame(index=valid.index,columns=valid.columns)
            for j in range(len(valid.columns)):
                for i in range(0, len(valid)):
                   pred.iloc[i][j] = prediction.iloc[i][j]
            df_results = invert_transformation(train, pred)
            #erro quadratico medio da previsao da coluna satélites com o
modelo testado
            erro=sqrt(mean_squared_error(df_results.iloc[:,5],
valid.iloc[:,0]))
            print(f'p={p}, q={q}, ERRO={erro}')
```

In [21]:

```
#detalhar o melhor resultado do modelo VARMAX
model = sm.tsa.VARMAX(endog=trainDiff, order=(1,1))
model_fit = model.fit()
print(model_fit.summary())
```

In [22]:

```
#fazer a previsão com melhor resultado do modelo v=VARMAX
prediction = model_fit.forecast(steps=10)
linhas=[]
for i in range(2017,2027):
    linhas.append(datetime.datetime(i,1,1))
pred = pd.DataFrame(index=linhas,columns=valid.columns)
for j in range(len(pred.columns)):
    for i in range(0, len(pred.index)):
        pred.iloc[i][j] = prediction.iloc[i][j]
df_results = invert_transformation(train, pred)
data1 = valid[valid.columns[0]]
```

```
data2 = df_results[df_results.columns[5]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['N° de satélites - Real','N° de satélites - Previsto'])
```

In [23]:
```
data1 = valid.iloc[0:3,1]
data2 = df_results[df_results.columns[6]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['IDH - Real','IDH - Previsto'])
```

In [24]:
```
data1 = valid.iloc[0:3,2]
data2 = df_results[df_results.columns[7]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['Pesquisadores - Real','Pesquisadores - Previsto'])
```

In [25]:
```
data1 = valid.iloc[0:1,3]
data2 = df_results[df_results.columns[8]]
plt.plot(data1,'ro',data2,'blue')
plt.legend(['Empresas - Real','Empresas - Previsto'])
```

In [26]:
```
data1 = valid.iloc[0:4,4]
data2 = df_results[df_results.columns[9]]
plt.plot(data1,'red',data2,'blue')
plt.legend(['Orçamento - Real','Orçamento - Previsto'])
```

Código de importação de dados dos pesquisadores:

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""

import pandas as pd
import numpy as np
from math import *
import csv

exit_file = open('mestresEdoutores.csv','w',newline='', encoding='utf-8')
colnames=["Date","Mestres","Doutores"]
writer=csv.DictWriter(exit_file, fieldnames=colnames)

ano=2004
csv_name=str(ano)+'.csv'
```

```
writer.writeheader()
for i in range(0,9):
        filename = str(ano)+'.csv'
        dfDiscentes= pd.read_csv(filename, sep=";",  low_memory=False)

        mestres=dfDiscentes.query('NM_SITUACAO_DISCENTE==    "TITULADO"    &
(NM_NIVEL_TITULACAO_DISCENTE == "MESTRADO" | NM_NIVEL_TITULACAO_DISCENTE ==
"MESTRADO PROFISSIONAL")').count()[1]
        doutores=dfDiscentes.query('NM_SITUACAO_DISCENTE==    "TITULADO"    &
(NM_NIVEL_TITULACAO_DISCENTE == "DOUTORADO")').count()[1]

        writer.writerow({"Date":   ano,   "Mestres":   mestres,   "Doutores":
doutores})
        ano+=1
#   A   partir   do   ano   de   2013   O   nome   da   coluna   muda   de
'NM_NIVEL_TITULACAO_DISCENTE'para 'DS_GRAU_ACADEMICO_DISCENTE

for i in range(0,7):
        filename = str(ano)+'.csv'
        dfDiscentes= pd.read_csv(filename, sep=";",  low_memory=False)

        mestres=dfDiscentes.query('NM_SITUACAO_DISCENTE==    "TITULADO"    &
(DS_GRAU_ACADEMICO_DISCENTE  ==  "MESTRADO"  |  DS_GRAU_ACADEMICO_DISCENTE  ==
"MESTRADO PROFISSIONAL")').count()[1]
        doutores=dfDiscentes.query('NM_SITUACAO_DISCENTE==    "TITULADO"    &
(DS_GRAU_ACADEMICO_DISCENTE == "DOUTORADO")').count()[1]

        writer.writerow({"Date":   ano,   "Mestres":   mestres,   "Doutores":
doutores})
        ano+=1
```