

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Núcleo de Educação à Distância

Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

José Alberto de Siqueira Brandão

**ANÁLISE PREDITIVA MEDIANTE SÉRIES TEMPORAIS:**

**Estudo sobre a carga horária de um programa de formação no setor público**

Recife

2020

José Alberto de Siqueira Brandão

**ANÁLISE PREDITIVA MEDIANTE SÉRIES TEMPORAIS:**  
**Estudo sobre a carga horária de um programa de formação no setor público**

Trabalho de Conclusão de Curso apresentado  
ao Núcleo de Educação à Distância - Curso de  
Especialização em Ciência de Dados e Big  
Data como requisito parcial à obtenção do título  
de especialista.

Área de concentração: Ciência de Dados

Recife  
2020

## SUMÁRIO

1. Introdução.....	4
1.1. Contextualização .....	4
1.1. O problema proposto .....	5
2. Coleta de Dados .....	7
3. Processamento/Tratamento de Dados .....	9
4. Análise e Exploração dos Dados .....	12
5. Criação de Modelos ARIMA.....	20
6. Apresentação dos Resultados .....	24
7. Links .....	26
REFERÊNCIAS.....	27
APÊNDICE.....	29

## **1 Introdução**

### **1.1 Contextualização**

As últimas décadas têm demandado significativos esforços de organizações públicas, nas mais diferentes esferas, no sentido de proporcionar qualificação profissional para servidores. Esta iniciativa se coaduna com a necessidade de atualização frente aos desafios impostos pelas mudanças de forma de atuação do poder público, que passa a incorporar ferramentas e modelo de gestão que se aproximam daquelas aplicadas na iniciativa privada (PERNAMBUCO, 2009a). Novas carreiras de servidores públicos foram criadas exatamente com o intuito de proporcionar melhor qualidade na prestação de serviços à população.

Um dos entes públicos que mais tem se destacado em relação a isto tem sido o Estado de Pernambuco. São conhecidas nacionalmente as políticas públicas do Pacto pela Vida, na área de segurança, e do Pacto pela Educação, por exemplo. O suporte legal requerido para a realização de tais iniciativas adveio, sobretudo, da criação de um modelo integrado de gestão, introduzido pela Lei Complementar nº 141/2009, e pelo estabelecimento de carreiras de gestor governamental, especificamente nas áreas administrativa, de planejamento, orçamento e gestão e de controle interno, mediante Leis Complementares nº 117, 118 e 119/2008, respectivamente (PERNAMBUCO, 2009b, 2009c, 2009d).

As três carreiras foram estruturadas mediante a realização de concursos públicos que possibilitaram maior capacidade das entidades estatais em possuírem respaldo gerencial para o alcance de significativos resultados. Uma das similaridades observadas nestas carreiras é a necessidade dos servidores vinculados a elas em se manter atualizados e preparados, haja vista a obrigatoriedade de participação em ações de capacitação para a percepção de parte significativa da remuneração, na forma de um adicional de incentivo à qualificação profissional (AIQP), correspondente a 50% da remuneração, bem como para a obtenção de progressão na carreira e de bônus do desempenho anual (BDA).

Para estruturar o processo de cômputo da carga horária foi instituído o Programa de Formação Continuada (PFC) no ano de 2011. Desta maneira, para que haja a percepção do AIQP e para fins de progressão torna-se necessário o alcance de, no mínimo, 60 (sessenta) horas de capacitação do gestor governamental no ano

anterior mediante registros no PFC. Existem exceções legais que estão vinculadas a ocupação de cargos diretivos nas organizações, além de eventuais afastamentos ou licenças, mas a carga horária indicada é necessária para a maior parte dos gestores governamentais.

Observa-se, portanto, a necessidade da ocorrência de um processo gerencial sobre os dados relacionados às ações de capacitação realizadas pelos gestores governamentais. Para que isto fosse possível, tornou-se necessária a criação de estrutura voltada para a gestão da carga horária dos servidores envolvidos, com a organização dos dados de modo a que melhor pudessem ser dimensionados os recursos necessários para a concretização das capacitações.

## **1.2 O problema proposto**

Em função da necessidade de melhor estruturar as ações de capacitação voltadas para o alcance destas horas, foi criado pelo Decreto nº 37.828/2012, no âmbito da Secretaria de Planejamento e Gestão – SEPLAG, o Instituto de Gestão Pública de Pernambuco – IGPE (PERNAMBUCO, 2012). Este Instituto é o órgão responsável, dentre outras atividades, por promover a aquisição, produção e fruição de conhecimentos de cunho acadêmico, supervisionar a estruturação de cursos de formação e de aperfeiçoamento profissional, e as atividades de treinamento e desenvolvimento técnico nas áreas relacionadas com a gestão pública estadual, voltados para o aprimoramento da gestão e para a formação continuada dos Analistas de Planejamento, Orçamento e Gestão, bem como promover e organizar conferências, simpósios, seminários, palestras e outros eventos assemelhados.

Uma das iniciativas desenvolvidas pelo IGPE é o PFC, cujo objetivo principal é proporcionar aos gestores governamentais ações de capacitação voltadas para sua atualização e constante aprendizagem de novos métodos e técnicas desenvolvidos no setor público. Ademais, o IGPE também promove a curadoria de ações de capacitação realizadas por outras instituições e que possam ter a participação de gestores governamentais, cuja carga horária possa ser reconhecida para fins de percepção dos incentivos anteriormente explicitados.

A carga horária necessária pode ser alcançada pelos gestores governamentais mediante a participação em ações de capacitação presenciais ou à distância, bem como também pode ser considerado a participação oriunda de

instrutoria em ações voltadas para o tema gestão pública. Percebe-se, então, a necessidade de um controle significativo sobre os dados relacionados aos processos de capacitação de servidores, haja vista o impacto financeiro que pode ser obtido caso ocorra o alcance dos quantitativos definidos. Ressalte-se que são diferentes meios para a obtenção da carga horária, por meio de participação enquanto discente, enquanto instrutor, ou ainda por meio de processos de validação de carga horária de ações de capacitação promovidas por entidades parceiras.

Apesar da responsabilidade atribuída ao IGPE de acompanhar a realização da carga horária em ações de capacitação dos gestores governamentais, ainda não havia instrumento formal de predição de carga horária a ser realizada. Deste modo, o planejamento da oferta e de validação de ações de capacitação encontrava-se baseado exclusivamente em repetições dos dados alcançados no ano anterior. Ou seja, não havia o estabelecimento de metas fundamentadas em dados, mas a simples repetição da expectativa anteriormente realizada.

O impacto organizacional da ausência de um estudo preditivo com base em dados implica diretamente na indefinição de metas coerentes, ocasionando falhas no dimensionamento de recursos necessários (equipe, instrutores, salas de aula, materiais didáticos, etc.) e, conseqüentemente, em eventuais lacunas na oferta de capacitações. De igual modo, não permite que sejam observadas eventuais sazonalidades, que poderiam impactar na baixa aderência de gestores governamentais em determinadas ações de capacitação em virtude de períodos em que estivessem ocupados com trabalhos específicos ou que estivessem em gozo de férias, por exemplo.

Resulta daí a importância para a realização de um estudo preditivo acerca da carga horária a ser ofertada pelo PFC. Sua consecução possibilitará dimensionar os recursos necessários para a realização do PFC nos próximos anos, contribuindo para a elaboração do plano estratégico e para a definição de metas a serem alcançadas pela unidade responsável. Pretende-se analisar a variação da carga horária válida ao longo do tempo, de modo a possibilitar uma projeção de demanda de novas ações de capacitação para os próximos anos, considerando que se trata de um programa de formação continuada.

Considerando a disponibilidade de dados confiáveis no IGPE e na SEPLAG, optou-se por desenvolver um estudo de série histórica sobre a carga horária ofertada e validada pelo PFC aos gestores governamentais, especificamente no

âmbito da SEPLAG. Os dados a serem analisados são oriundos da participação de gestores governamentais vinculados à Secretaria de Planejamento e Gestão de Pernambuco – SEPLAG/PE em ações de capacitação ao longo do período compreendido entre 2014 e 2019.

Os dados estudados correspondem ao período de seis anos, até o mais recente ano com dados disponíveis. A análise se ateve ao âmbito dos servidores ativos na SEPLAG/PE, desconsiderando eventuais cessões, licenças, afastamentos e exonerações que tenham ocorrido ao longo do período estudado. Deste modo, definiu-se como objetivo do estudo dimensionar a carga horária de ações de capacitação para o PFC/SEPLAG com base em um estudo de série temporal para os anos de 2020 e 2021.

## **2 Coleta de Dados**

O processo de coleta de dados foi iniciado com o acesso aos arquivos da SEPLAG, contendo os registros das ações de capacitação realizados pelos gestores governamentais, disponibilizados em formato de planilha eletrônica. Cada arquivo corresponde ao período de um ano, possuindo dados sobre ações de capacitação que foram promovidas pelo IGPE, ou que tenham sido promovidas por outras instituições, de modo presencial ou à distância, e que tenham sido validadas. Além disso, os arquivos contém registro de carga horária obtida mediante instrutoria em capacitações com a temática voltada para a área de gestão pública. Para fins de relacionamentos entre planilhas foi adotado o campo-chave CPF, considerando ser um documento disponível nos bancos de dados da SEPLAG e amplamente reconhecido como identificador de pessoas no âmbito nacional.

Cabe destacar ainda que as ações de capacitação promovidas pelo IGPE são abertas ao público interessado, ou seja, pode englobar interessados que não sejam necessariamente gestores governamentais. Entretanto, o foco deste estudo se volta para este tipo de servidor, considerando que são aqueles afetados pelos efeitos de ordem econômica, com a percepção de bonificações e/ou adicionais atrelados às ações de capacitação.

Cada linha da planilha corresponde ao registro de um servidor em uma ação de capacitação, de modo que podem ocorrer repetições específicas no campo

relativo ao nome do servidor. No entanto, não pode haver lacunas no preenchimento dos campos, haja vista a necessidade de todos os campos para efeito de validação da carga horária relacionada a cada ação de capacitação.

As planilhas anuais contemplam 18 colunas (ou campos). O quadro 1 foi elaborado para que seja possível uma melhor visualização da estrutura de dados, contemplando o nome do campo, sua descrição e o tipo de dados nele contido.

Quadro 1. Estrutura do dataset

Nome da coluna/campo	Descrição	Tipo
Servidor	Nome completo do servidor	<i>string</i> / polinomial não categórico
CPF	Número de cadastro de pessoa física	<i>integer</i> / quantitativo discreto
TipoServ	Identificação do tipo de servidor, se gestor governamental ou não.	<i>string</i> / binominal assimétrico
Lotacao	Local onde o servidor atuava na época da ação de capacitação	<i>string</i> / polinomial não categórico
LotacaoA	Local onde o servidor atua	<i>string</i> / polinomial não categórico
Situacao	Status do servidor à época da ação de capacitação	<i>string</i> / polinomial categórico
Promotor	Entidade promotora da ação de capacitação	<i>string</i> / polinomial não categórico
Acao	Nome da ação de capacitação	<i>string</i> / polinomial não categórico
Forma	Origem da carga horária de capacitação (ofertada, validada, instrutoria)	<i>string</i> / polinomial categórico
Tipo	Definição do tipo de ação de capacitação (curso, seminário, etc.)	<i>string</i> / polinomial categórico
Modalidade	Definição do espaço de realização da ação de capacitação (presencial, à distância)	<i>string</i> / binominal simétrico
Turno	Definição do tempo de realização da ação de capacitação	<i>string</i> / polinomial categórico
Datal	Data de início da ação de capacitação	<i>integer</i> / quantitativo discreto
DataF	Data de término da ação de capacitação	<i>integer</i> / quantitativo discreto
Ano	Ano de realização da ação de capacitação	<i>integer</i> / quantitativo discreto
Mês	Identificação do mês em que a ação de capacitação foi encerrada	<i>integer</i> / quantitativo discreto
CHT	Carga horária total da ação de capacitação	<i>integer</i> / quantitativo discreto
CHV	Carga horária válida da ação de capacitação	<i>integer</i> / quantitativo discreto

Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Considerando-se que a maior parte dos campos é do tipo *string* e que o foco do estudo encontra-se relacionado à distribuição da carga horária, optou-se por trabalhar com a carga horária válida (CHV) para efeito dos cálculos da série temporal. De todo modo, apenas para efeito de classificação, optou-se ainda por representar outros dados de modo meramente descritivo.



### 3 Processamento/Tratamento de Dados

Uma das primeiras preocupações após a coleta dos dados foi verificar a consistência e a fidedignidade dos mesmos. Como se tratam de dados que alimentam a concessão de incentivos no setor público, estes dados precisam ser divulgados anualmente em publicação no diário oficial. Desta maneira, não podem existir falhas nos registros, bem como a ausência de qualquer dado referente às ações de capacitação.

Para efeito de verificação do status dos participantes, também foi necessária a realização de uma consulta em planilha eletrônica disponibilizada pela Gerência de Pessoas (Gespe) da Seplag. Nesta planilha constam os registros anuais de movimentação de pessoal, como cessões, licenças, afastamentos e exonerações, que possibilitam a verificação de aspectos importantes para o entendimento da participação dos GGPOGs em ações de capacitação. Com base nela são atualizados os campos relativos à lotação atual e situação do gestor governamental. Neste sentido, entende-se que o servidor cedido, por exemplo, não pode estar computando carga horária para efeito de fruição de bônus ou de adicional de qualificação.

Além disso, foi necessário realizar o ajuste demandado pela orientação do trabalho de conclusão de curso relativa à edição dos dados para que não fosse possível a identificação de dados pessoais. Neste sentido, optou-se por realizar um ajuste no campo “servidor” nos próprios arquivos em excel, inserindo um conjunto de caracteres para representar cada nome. No mesmo sentido, procedeu-se a submissão de parte dos números que identificam o CPF de cada servidor de modo que não fosse possível sua identificação imediata e que não ocorresse repetição dos números representativos.

A partir deste ponto foi adotada a alternativa de desenvolver o estudo com a linguagem R. A opção por esta linguagem se deu em função da mesma ser *open source*, disponibilizando gratuitamente diversos pacotes para programação estatística, inclusive sobre a temática de séries temporais (AQUINO, 2014). Para a leitura e manipulação de planilhas eletrônicas foi utilizado o pacote *readxl()* (LONG; TEETOR, 2019).

Após a checagem inicial dos dados relativos ao ano de 2019, considerando que os demais anos já haviam sido finalizados, coube a realização do processo de

empilhamento dos dados, ou seja, realizar o agrupamento das diversas planilhas anuais em um único banco de dados, considerando que todas possuem os mesmos campos. Depois que foi feito o empilhamento das planilhas anuais, mediante a adoção da função *rbind()* (VENABLES; SMITH, 2020), observou-se que o banco de dados contava com 5.834 registros (ou linhas).

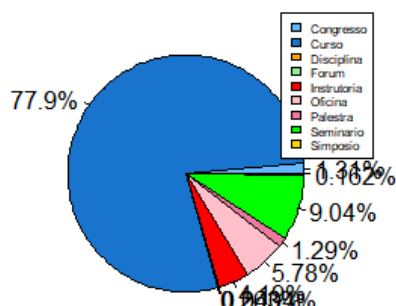
No entanto, ainda era necessário fazer o processo de descarte dos dados que não eram de interesse do estudo, tais como a participação de servidores que não sejam gestores governamentais e a retirada das ações de capacitação que não possuíssem carga horária válida. Neste sentido, realizou-se o processo de filtragem, considerando os campos TipoServ e CPF (zerados para quem não é gestor governamental) como parâmetros para exclusão dos servidores que não fossem gestores governamentais, mediante a função *subset()* (FERREIRA, 2018), gerando redução dos registros válidos, que passou a compor um conjunto de 4.946 registros realizados por 191 servidores com as características anteriormente indicadas.

Sobre estes registros individuais foi realizado o trabalho de estatística descritiva, conforme Long e Teetor (2019), especialmente em função da variável escolhida para análise, ou seja, carga horária válida. Foram realizados os cálculos de medidas de posição, com a função *summary()* calculando média (14,56h.), mediana (16h.), valor mínimo (0h.), valor máximo (90h.), 1º quartil (4h.) e 3º quartil (20h.). Também foram calculadas as medidas de dispersão, como variância (145,97h.), desvio-padrão (12,08h.) e coeficiente de variação (82,97). Apesar dos dados indicarem uma concentração de respostas em torno da média, foi verificada a presença de alguns *outliers*, com registros individuais de carga horária válida superiores acima de 40 horas.

Para melhor compreensão dos registros obtidos inicialmente foram elaboradas representações gráficas sobre os principais conjuntos de dados. As figuras 1 e 2 apresentam, respectivamente, os percentuais quanto ao tipo e quanto à forma da ação de capacitação. É possível observar grande proeminência do tipo curso (77,9%), seguido da participação em fóruns (9,04%) e em oficinas (5,78%), cabendo aos outros tipos um percentual equivalente a 7,28%, demonstrando a preferência por ações de capacitação em cursos tradicionais. No tocante à forma de capacitação verifica-se que a maior parte da carga horária é obtida mediante cursos ofertados pelo próprio PFC (61,6%), enquanto boa parte (34,4%) da carga horária é

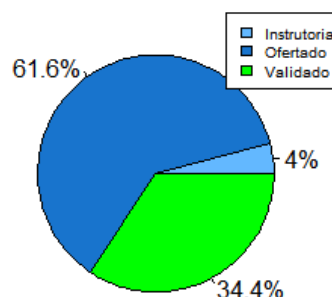
alcançada em processos de validação, ou seja, cursos ofertados por parceiros e pequena parcela é obtida mediante instrutoria (4%).

Figura 1 – Tipo de capacitação



Fonte: elaborado pelo autor.

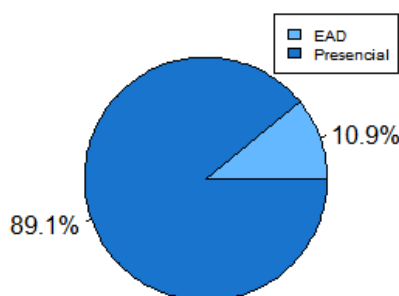
Figura 2 – Forma de capacitação



Fonte: elaborado pelo autor.

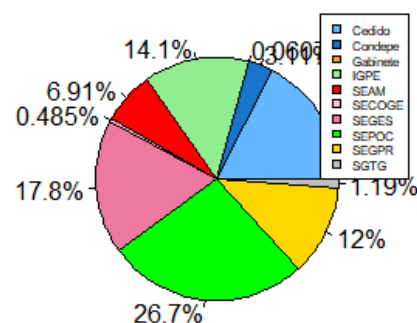
As figuras 3 e 4 demonstram, respectivamente, os percentuais quanto à modalidade da ação de capacitação e quanto à lotação do participante da ação de capacitação. Verifica-se a predominância da participação em cursos presenciais (89,1%) em detrimento dos cursos na modalidade de educação à distância – EAD (10,9%), reforçando a perspectiva de capacitações ainda tradicionais. Quanto à lotação é possível observar uma distribuição entre as unidades, com maior incidência de participantes do IGPE (26,7%), da SEGES (17,8%), da SEPOC (14,1%) e da SEGPR (12%).

Figura 3 – Modalidade de capacitação



Fonte: elaborado pelo autor.

Figura 4 – Lotação do participante



Fonte: elaborado pelo autor.

Procedeu-se ainda a acumulação dos dados mensais e anuais relativos à carga horária válida, de modo a se obter os quantitativos mensais. Neste momento foi verificada a ocorrência de carga horária válida nula (zero) em alguns meses entre os anos de 2011 e 2013, o que levou a definição pela construção da série temporal a partir do ano de 2014. Ainda assim foi observada a ausência de registros no mês de

abril de 2014. Para construir a série temporal optou-se por registrar neste mês o valor de carga horária válida equivalente ao menor registro mensal da série. Deste modo, atribui-se o valor de 116 horas válidas para o mês com registro nulo, retirando-se do mês subsequente esta carga horária, produzindo-se o menor impacto possível e permitindo a construção da série temporal.

A agregação dos dados possibilitou a identificação da quantidade mensal de carga horária válida, com a construção de variável contemplando 72 registros mensais, o equivalente a quantidade de meses compreendida entre janeiro de 2014 e dezembro de 2019. O cômputo total da carga horária válida ao longo dos meses implicou em um total de 72.025 horas em ações de capacitação. Foi criado um novo campo com o registro acumulado com a data retroagindo ao primeiro dia de cada mês, no formato ano-mês-dia. Deste modo, foi possível observar a evolução dos dados mensais ao longo dos anos, gerando a seguinte tabela:

Tabela 1 – Distribuição da carga horária válida ao longo dos meses e anos

Ano/Mês	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total Ano	$\Delta\%$
2014	880	412	199	116	935	760	690	840	929	735	1509	152	8157	-
2015	334	1136	721	583	810	1632	1136	2052	772	684	1043	260	11163	137%
2016	572	531	2319	500	896	1024	1054	1381	773	1761	1268	568	12647	113%
2017	450	315	1802	475	1618	545	674	886	800	1699	1419	472	11155	88%
2018	116	1437	1289	1122	1553	1118	919	825	1775	820	1291	432	12697	114%
2019	368	748	1528	1190	1785	1862	1054	998	1690	1191	2018	1774	16206	128%
Total Mês	2720	4579	7858	3986	7597	6941	5527	6982	6739	6890	8548	3658	72025	116%

Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Apesar dos destaques sobre os resultados obtidos com o processamento e o tratamento dos dados, bem como a análise da estatística descritiva que foi importante para melhor compreensão dos elementos componentes da base, o cerne deste trabalho reside no processo de análise da série temporal. A próxima seção abordará especificamente a discussão sobre a temática central do trabalho, com a exploração da série temporal visando prever os quantitativos de carga horária válida a serem indicados para os próximos períodos.

#### 4 Análise e Exploração dos Dados

A partir do processamento e tratamento dos dados, conforme descrito na seção anterior, foi possível observar a construção da série temporal, organizada em

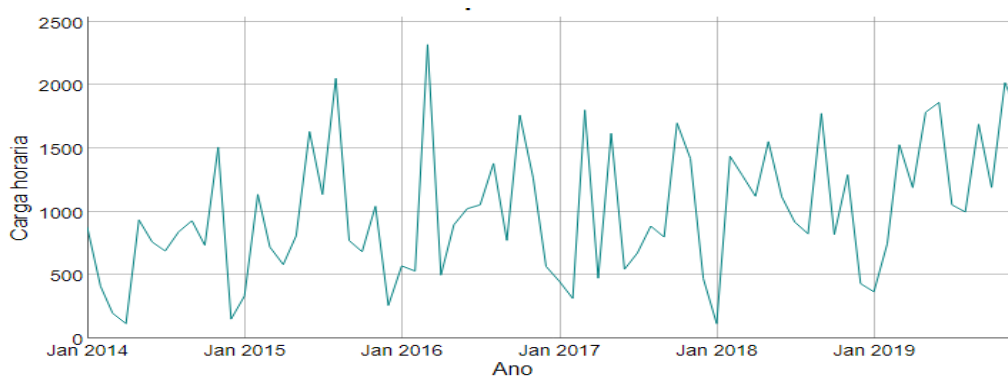
72 meses, conforme apresentado na tabela 1. Na presente seção será demonstrada a análise e exploração dos dados contidos na série temporal. A etapa inicial é identificar características e comportamentos dos dados nela contidos.

A leitura da tabela 1 já possibilitou a observação de muitas características da série temporal. Verificou-se que a média anual de carga horária válida ao longo do período estudado é de 12.004 horas, havendo um ápice no ano de 2019, quando foram registradas 16.206 horas, e o menor quantitativo em 2014, quando foram registradas 8.157 horas.

Ainda considerando os dados dispostos na tabela 1 é possível verificar o comportamento ao longo dos meses. Neste sentido, verificou-se que os meses de novembro e março apresentam, respectivamente, as maiores médias de carga horária. É importante destacar que os servidores precisam cumprir a carga horária mínima estabelecida dentro do mesmo ano, sendo esta a motivação para a construção da hipótese de maior observância de carga horária nestes meses. Destaca-se ainda que, como a Seplag é responsável pelo orçamento do Estado, que precisa ser finalizado até o mês de novembro, os servidores procuram concluir sua obrigação antes deste período, de modo a não comprometer nem a necessidade de trabalho e nem a necessidade de formação. Observa-se ainda uma menor incidência de carga horária válida nos meses de janeiro e dezembro, período em que se concentra a maior parte das demandas por férias.

Para melhor visualizar estes dados, optou-se pela construção de gráficos de linha, que possibilitam a visualização de determinado fenômeno ao longo do tempo. Inicialmente foi elaborada uma representação que demonstrasse a evolução da carga horária válida em função dos meses, conforme se observa na figura 5:

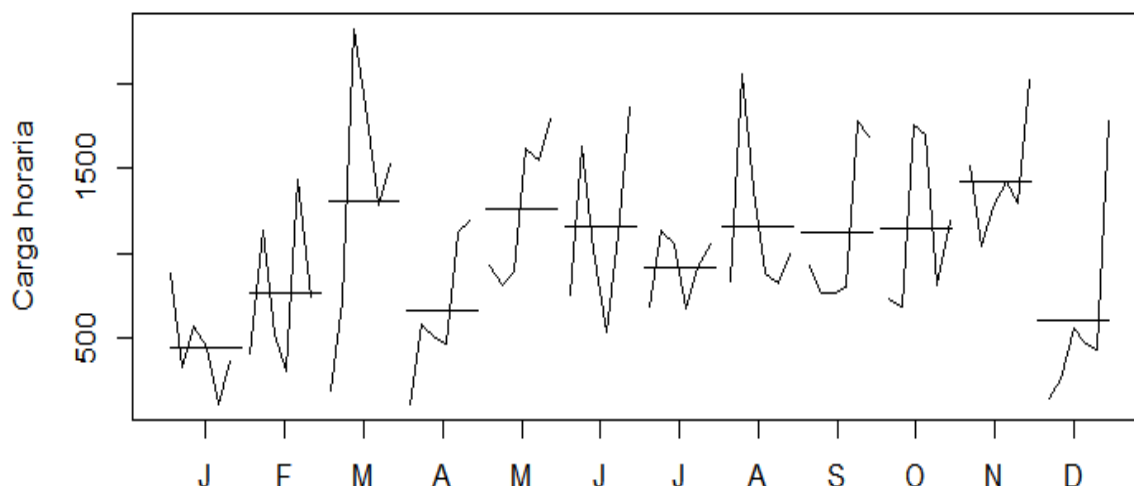
Figura 5 – Evolução da carga horária válida de capacitação ao longo do período 2014-2019



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

A análise gráfica sobre a figura 5 permite identificar a oscilação da carga horária válida, com a presença de meses com maior incidência (picos) e meses com menor incidência (vales). Observa-se que o mês de março de 2016 apresenta o maior pico, evidenciado pela realização de curso de formação englobando todos os servidores com características estudadas neste trabalho. Por outro lado, verifica-se a existência de vales ao longo da série temporal, sendo destacado o mês de janeiro de 2018. Para melhor compreender esta oscilação em função dos meses foi elaborado um gráfico, mediante a função *monthplot()* (FERREIRA, 2018), demonstrando a oscilação da carga horária válida distribuída por mês ao longo dos anos, conforme se observa na figura 6:

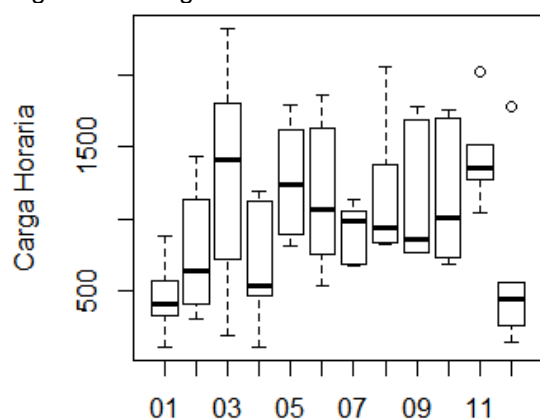
Figura 6 – Evolução da carga horária válida de capacitação ao longo do período 2014-2019



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

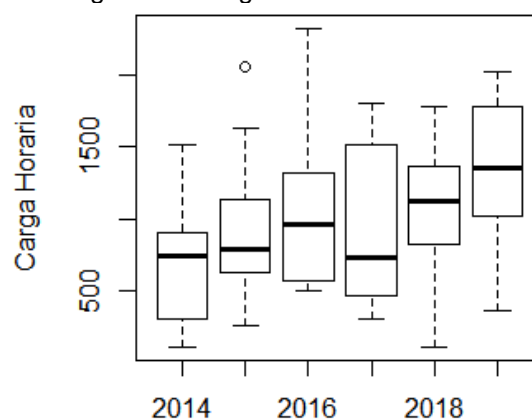
A análise gráfica da figura 6 indica a evolução da carga horária válida em cada mês ao longo dos anos. Deste modo, percebe-se que o mês de novembro é aquele que possui maior média de carga horária válida, especialmente em função da necessidade de se completar o cômputo das horas até o final do ano. Em sequência aparece o mês de março, ainda que também represente aquele em que ocorre maior amplitude ou variação dos dados, ratificando a atipicidade dos registros neste mês. Por outro lado, os meses com menores resultados são os meses de janeiro e de dezembro, provavelmente em função do período de férias. As figuras 7 e 8, a seguir representadas, demonstram no formato de *boxplot* os dados relativos aos meses e anos:

Figura 7 – Carga horária mensal 2014-2019



Fonte: elaborado pelo autor.

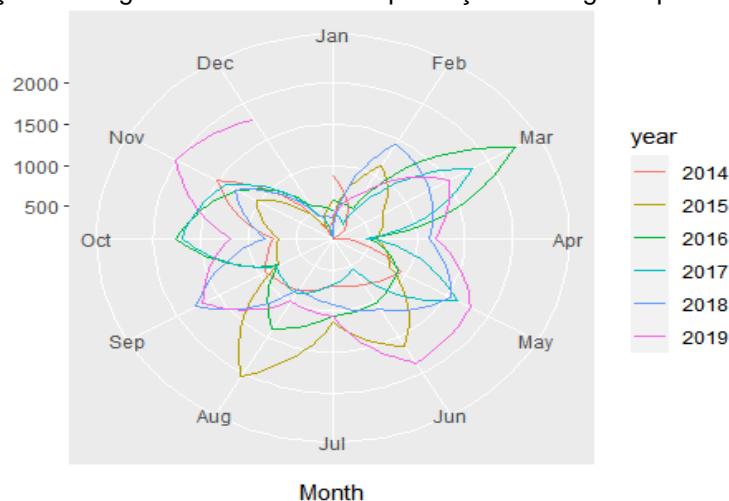
Figura 8 – Carga horária anual 2014-2019



Fonte: elaborado pelo autor.

Verifica-se na figura 7 que no ano de 2019 ocorreu um significativo crescimento dos resultados de carga horária relativos aos meses de novembro e dezembro, implicando na verificação de referências fora do padrão ou *outliers*. Esta ascensão pode ter ocorrido em função da mudança no entendimento jurídico acerca da necessidade de formação do corpo gerencial e de servidores cedidos, passando a ser compulsória a carga horária para que houvesse a progressão na carreira e a percepção de bônus para estes servidores. Deste modo, ocorreu uma sobrecarga de demanda, implicando no aumento significativo do quantitativo de carga horária válida nestes períodos. Outra maneira de observar a distribuição dos dados mensais e, especialmente as oscilações não regulares é através do gráfico de radar. Para ampliar a percepção sobre estas ocorrências, a análise da figura 9 possibilita o entendimento sobre tais variações.

Figura 9 – Distribuição da carga horária válida de capacitação ao longo do período 2014-2019

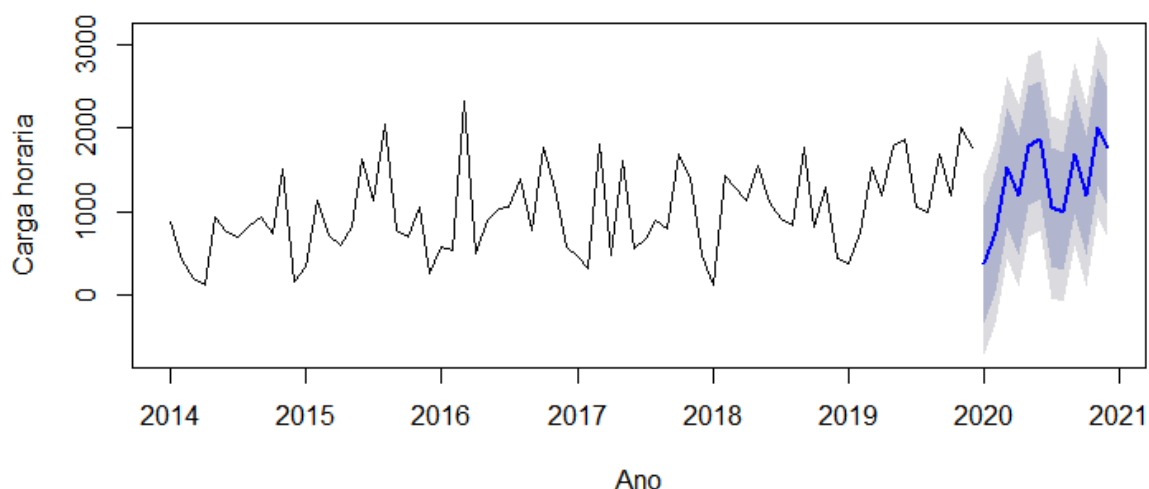


Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Com base nos dados mensais apresentados foi possível a construção da série temporal 2014-2019. A sua constituição ocorreu mediante a adoção da função `ts()` do pacote `xts` (AQUINO, 2014), gerando dados conforme anteriormente apresentados na tabela 1.

O trabalho do IGPE com relação aos dados mensais do PFC era concluído nesta etapa. A projeção da demanda para os anos seguintes se baseava em um modelo sazonal ingênuo (SNAIVE), cujo quantitativo de carga horária projetada dependia exclusivamente da observância ligeiramente anterior, eventualmente acrescentando algum percentual de expectativa de crescimento. Para que fosse possível entender esta etapa foi elaborada a figura 10 com a representação gráfica do modelo adotado. Observa-se em 2020 uma repetição do comportamento da linha do gráfico indicado para o ano de 2019. Percebe-se o erro neste tipo de projeção, especialmente por desconsiderar a tendência de crescimento observada ao longo dos anos.

Figura 10 – Projeção da Série Temporal 2014-2019 para o ano de 2020



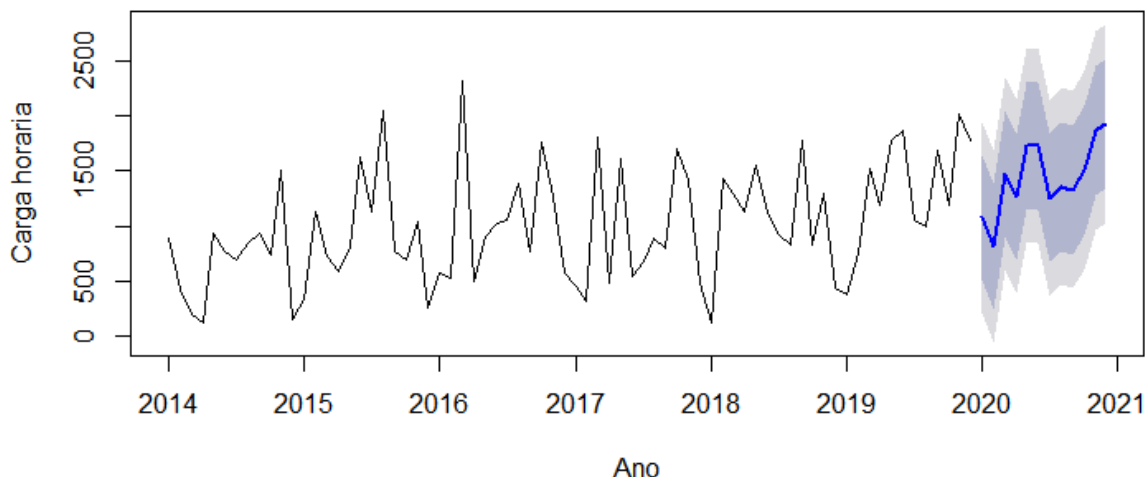
Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

A distribuição da série temporal original já possibilita a realização de algumas projeções conforme modelos mais avançados que o SNAIVE. Uma das possibilidades é a aplicação do modelo ARIMA, ou seja, um modelo autorregressivo integrado de médias móveis (*autoregressive integrated moving average* ou ARIMA, na sigla em inglês). A aplicação do modelo automático de ARIMA (`auto.arima`) à série temporal original implicou em um resultado melhor do que o resultado obtido com o SNAIVE, conforme se verifica na figura 11. No entanto, quando se verifica o critério de informação de



Akaike (AIC), que é uma medida de qualidade do modelo, verificou-se um resultado bastante elevado (1.080,37), sendo indicada a continuidade do estudo com a finalidade de obtenção de um resultado de AIC mais baixo.

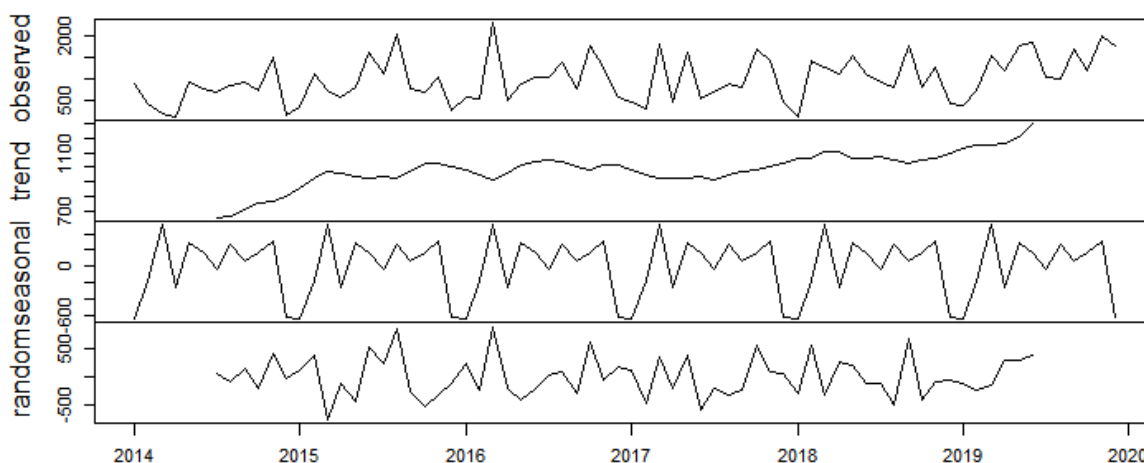
Figura 11 – Projeção da Série Temporal 2014-2019 para o ano de 2020 conforme modelo auto-arima



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Outra representação importante no desenvolvimento da análise gráfica da série temporal é o que destaca a decomposição da série temporal, exposta na figura 12. Este recurso, acionado pela função *decompose()* (FERREIRA, 2018), possibilita a visualização da série em virtude dos dados observados, além da construção de uma linha de tendência, da identificação de sazonalidade e de um componente de aleatoriedade dos dados.

Figura 12 – Decomposição da série temporal 2014-2019



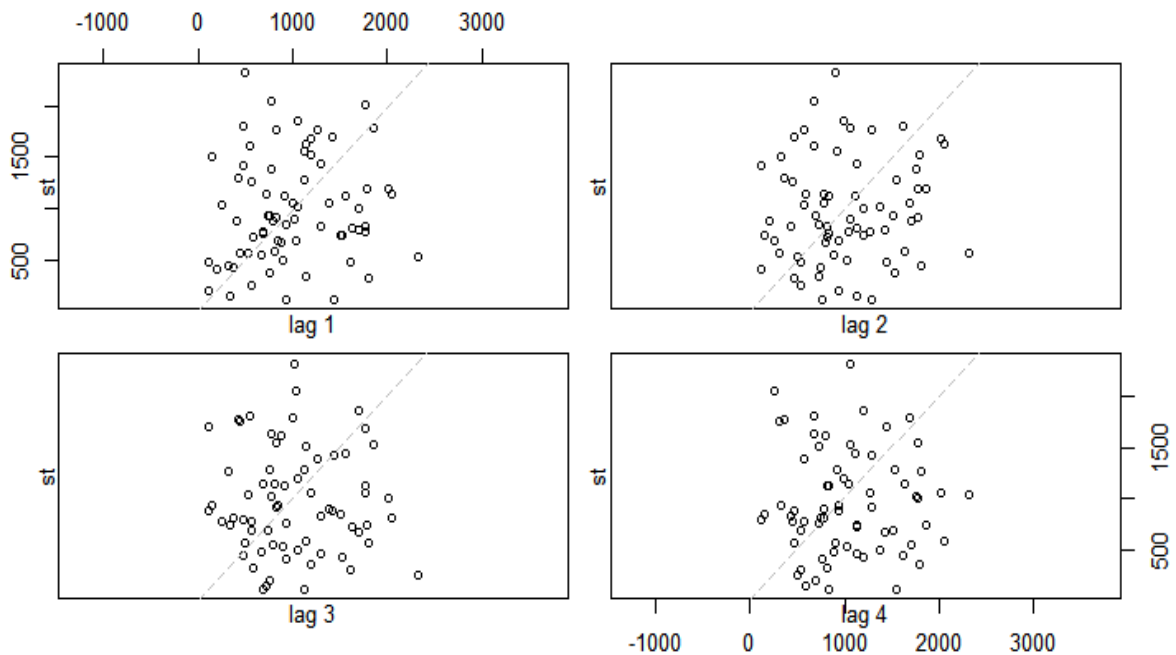
Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Com base neste gráfico é possível confirmar a tendência de crescimento suave da série temporal. Também é observável a presença de ciclos anuais, com

picos mais elevados ocorrendo no mês de março e as maiores depressões ocorrendo no mês de janeiro, o que reforça as interpretações anteriormente realizadas especialmente no tocante à existência de sazonalidade na ocorrência dos registros de carga horária.

Outra forma gráfica de análise é a que relaciona a série temporal original com a série temporal defasada conforme uma quantidade determinada de *lags*. A figura 13, obtida mediante a função *lag.plot()* (ROME, 2020), demonstra esta relação em função de 4 *lags*. Nela é possível identificar que, apesar de haver certa dispersão dos dados, existe uma relação linear positiva entre a série temporal original e os dados dispostos nas séries temporais defasadas.

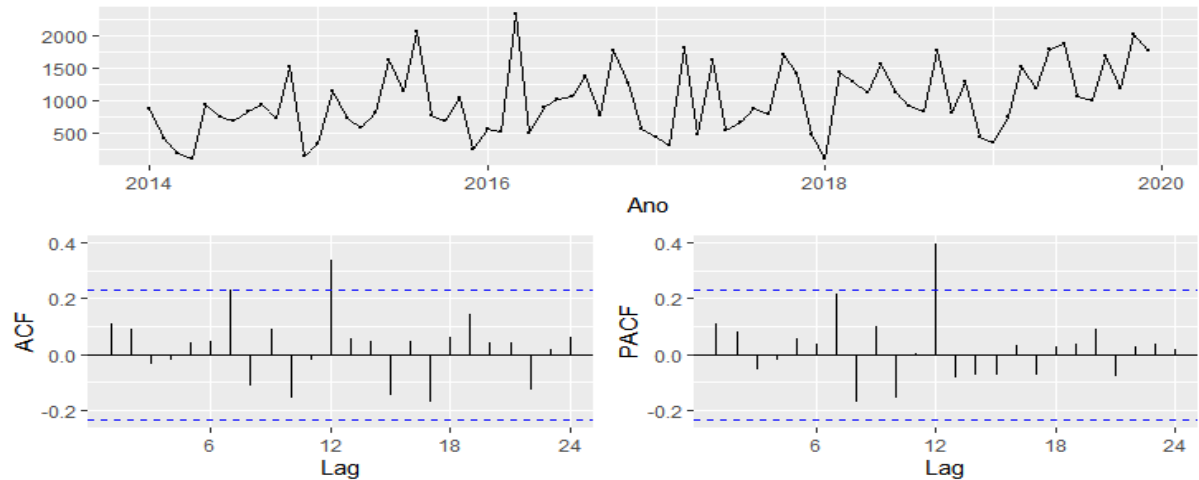
Figura 13 – Defasagem da série temporal 2014-2019 em 4 *lags*



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Além disso, é possível a realização de análise gráfica mediante a observância de autocorrelação, mediante a função *acf()*, e de autocorrelação parcial, através da função *pacf()* (FERREIRA, 2018). A figura 14 abaixo apresentada indica que a série temporal escolhida apresenta características de estacionariedade, ou seja, vem se mantendo durante quase todo o período indicado dentro da faixa entre -0,2 e 0,2, no entorno de uma média constante, não havendo variância significativa. A exceção se verifica na aplicação do lag 12, onde ocorre uma extrapolação desta faixa, alcançando quase o dobro do limite estabelecido.

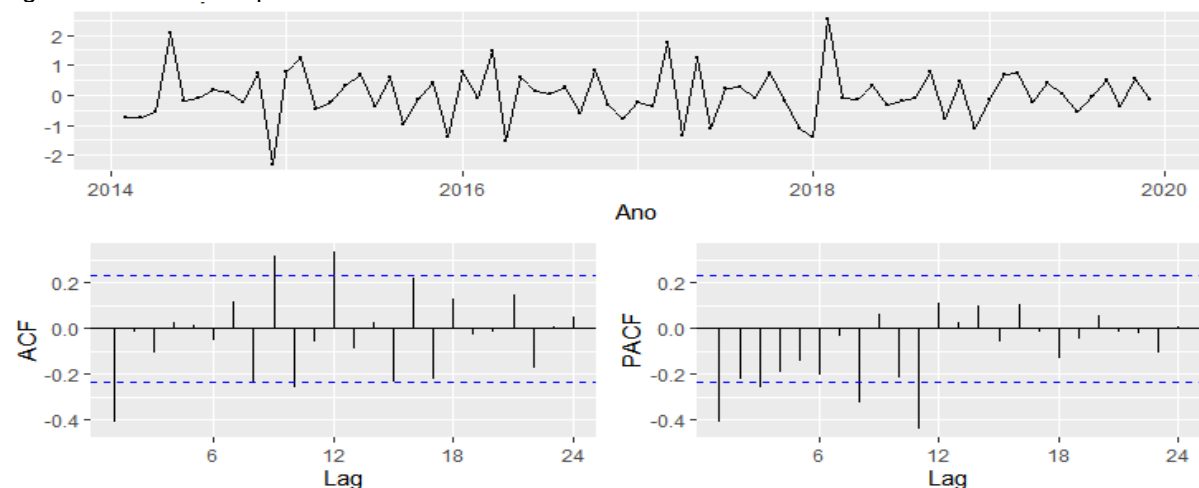
Figura 14 – Série temporal 2014-2019 com ACF e PACF



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Além da análise gráfica é importante adotar outros meios para a verificação da presença de estacionariedade de modo que haja uma confirmação sobre como atuar sobre os dados. Um conjunto de testes com raiz unitária pode ser aplicado para a verificação de estacionariedade em uma série temporal. Para efeito deste estudo foi aplicado o teste de Dickey-Fuller Aumentado, que apesar de apresentar p-valor baixo, não rejeitou a hipótese de não-estacionariedade ao apresentar resultado da estatística de teste ( $\tau_2 = 1.219294$ ) superior ao valor crítico ao nível de confiança de 95% (-1,95) para o tamanho da amostra. Desta maneira, diante do conflito com a análise gráfica, optou-se por realizar diferenciação na série temporal para que fosse confirmada a estacionariedade. Após a aplicação desta diferenciação, verificou-se a situação gráfica demonstrada na figura 15.

Figura 15 – Série temporal 2014-2019 diferenciada com ACF e PACF



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

A análise da figura 15 possibilita enxergar uma maior estacionariedade da série temporal mediante a aplicação da diferenciação. No entanto, o PACF apresenta três pontos de extrapolação dos limites estabelecidos. Ao se aplicar o teste de Dickey-Fuller Aumentado com uma diferenciação na série temporal e com a transformação logarítmica, obteve-se o resultado da estatística de teste satisfatório (-6.825953), tornando-se possível a rejeição da hipótese de não-estacionariedade. Com isto, ficou possibilitada a construção da modelagem conforme o modelo ARIMA, tema discutido na próxima seção.

## 5 Criação do Modelo ARIMA

Diante da série temporal diferenciada e transformada logaritmicamente tornou-se possível sua caracterização enquanto série estacionária. Optou-se pela realização de um modelo ARIMA, ou seja, um modelo autorregressivo integrado de médias móveis. A confecção do ARIMA possibilita um melhor entendimento sobre os dados, permitindo a realização de previsões de pontos futuros na série temporal.

A construção do modelo ARIMA demanda a realização de cinco etapas, conforme Granger e Newbold (apud Ferreira, 2018). A primeira etapa é a especificação da classe geral da estrutura, ou seja, a definição do número de defasagens ou autorregressão ( $p$ ), do grau de diferenciação ( $q$ ) e de médias móveis ( $d$ ) a ser adotado. Para iniciar este processo, optou-se por aplicar a função `auto.arima()` (ROMA, 2020) para a série temporal transformada. O resultado indicou uma significativa redução no AIC em relação à série original, passando a ficar com o resultado de 135,73. O modelo sugerido foi de ARIMA (2,0,1)(1,0,0)[12].

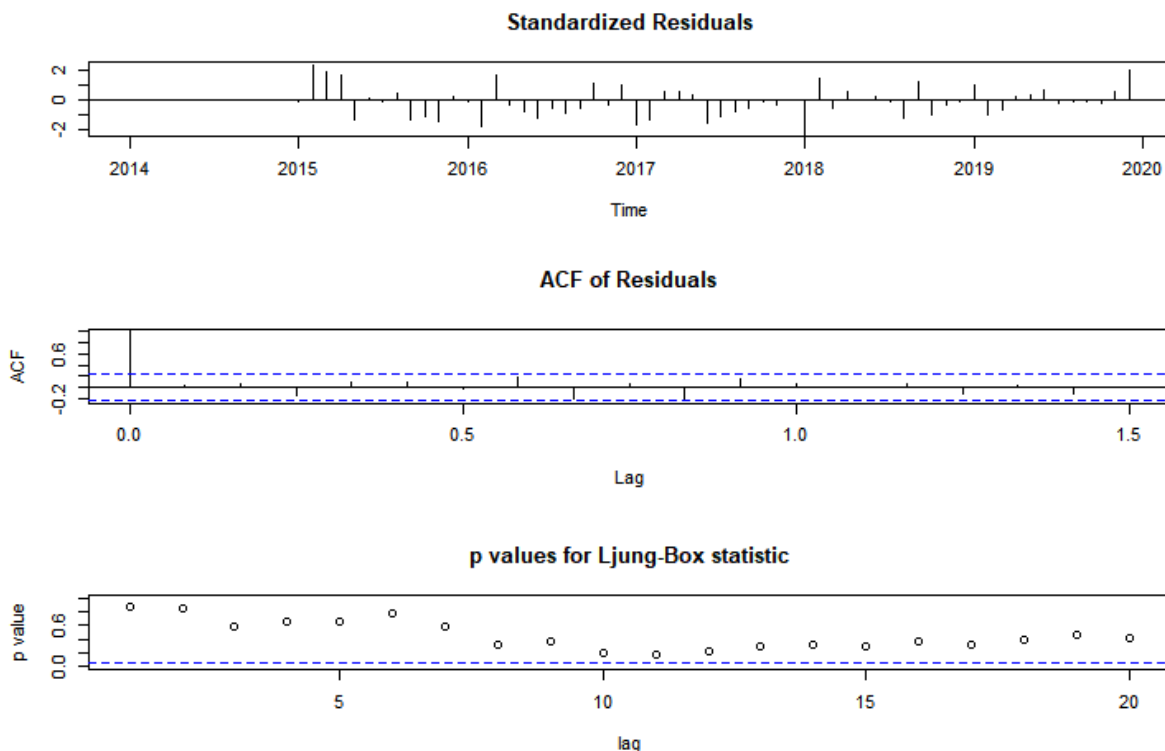
A segunda etapa é a identificação da série mediante a adoção da função de autocorrelação (ACF) e da função de autocorrelação parcial (PACF) e outros critérios. Com base na análise da figura 15, foi identificado que a série temporal não precisa de autorregressão, com a aplicação de uma diferenciação, e aplicação de 3 médias móveis em função das extrapolações identificadas anteriormente, tudo isso com periodicidade anual.

A terceira etapa é a estimação, quando são testadas as possibilidades de ARIMA, mediante a função `arima()` do pacote *forecast* (FERREIRA, 2018). Foi

adotado o método de máxima verossimilhança (ML) por ser o padrão usado na linguagem R. É importante expor que foram realizados testes com outras possibilidades de modelo ARIMA, com alterações nas variáveis. No entanto, o modelo ARIMA (0,1,3)(0,1,3)[12] obteve o menor resultado de AIC, critério válido para comparação entre modelos de série temporal, alcançando o patamar de 115,22.

A quarta etapa corresponde ao diagnóstico do modelo, com a análise dos resíduos e aplicação de testes de verificação para checar se o modelo sugerido é adequado. Nesta etapa é preciso verificar as características dos resíduos, de modo a se comprovar a ausência de autocorrelação linear, a ausência de heterocedasticidade condicional e a verificação da normalidade. Um visão global dos resíduos é realizada com a aplicação da função *tsdiag()* (FERREIRA, 2018) conforme se observa na figura 16.

Figura 16 – Diagnóstico dos resíduos do modelo ARIMA (0,1,3)(0,1,3)[12]



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

A primeira imagem da figura 16 retrata a distribuição dos resíduos padronizados, com dados que estão distribuídos em torno da média zero, sem nenhuma extrapolação dos limites, indicando pertencer a uma distribuição normal. A

segunda imagem da figura 16 representa a ACF dos resíduos, apresentando resultados não discrepantes, o que sinaliza ausência de autocorrelação linear nos resíduos. A terceira imagem da figura 16 representa o p-valor da estatística Ljung-Box para diferentes defasagens, com os resultados se posicionando sempre acima do limite estabelecido, de modo a possibilitar o entendimento de não rejeição da hipótese nula de não existência de dependência serial.

Além da aplicação deste diagnóstico, foram realizados outros testes. Inicialmente aplicou-se o teste de Ljung-Box para um número maior de lags (24). O resultado indicou p-valor igual a 0,193, confirmando a ausência de autocorrelação linear nos resíduos.

Outro teste aplicado foi o multiplicador de Lagrange para heterocedasticidade condicional autorregressiva (ARCH LM), visando avaliar a estacionariedade da variância. O teste também foi aplicado para a mesma quantidade de lags (24) e resultou em um p-valor de 0,1185, indicando a não rejeição da hipótese nula a 95% de confiança, indicando a estacionariedade da variância.

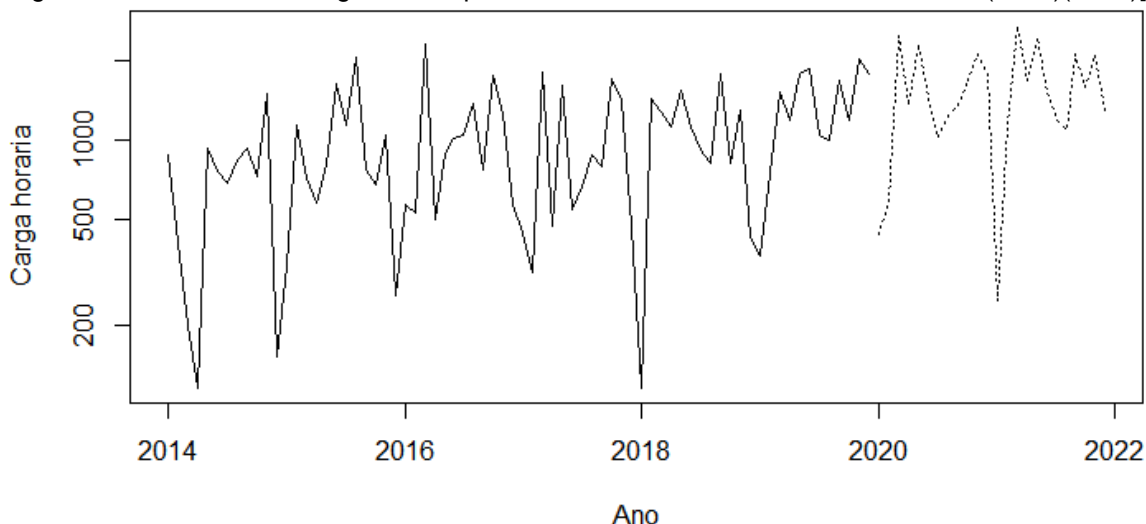
Para confirmar a normalidade do resíduo foi aplicado o teste de Jarque e Bera, baseado no pacote *normtest()* (FERREIRA, 2018). A aplicação deste teste indicou o p-valor de 0,358, direcionando para o entendimento de não rejeição da hipótese nula de normalidade a 95%.

Ademais, para efeito de comparação do modelo ARIMA (0,1,3)(0,1,3)[12] com o modelo auto.arima foram utilizadas as medidas AIC (quanto menor, melhor) e logLik (quanto maior, melhor) (ROME, 2020). A aplicação de ambas indicou um resultado melhor para o modelo construído pelo autor, sendo o AIC menor (115,22 contra 135,73) e o logLik maior (-50,61 contra -62,86).

Finalmente, Granger e Newbold (apud FERREIRA, 2018) recomendam uma quinta etapa que é a elaboração do modelo definitivo. Estes autores recomendam a consideração das medidas RMSE (raiz quadrada do erro médio) e MAPE (erro percentual absoluto médio), obtidas na apuração da acurácia dos modelos. Quando comparados os resultados, observa-se que para ambas as medidas o modelo ARIMA (0,1,3)(0,1,3)[12] apresenta resultados menores que o auto.arima. Deste modo, conforme as medidas utilizadas para avaliação dos modelos, é possível confirmar a predominância do modelo construído pela autoria em detrimento do modelo auto-arima.

Com fundamento na definição do modelo definitivo, torna-se necessária a sua aplicação para extração da previsão desejada. Inicialmente foi elaborada a figura 17 que contempla a projeção realizada conforme o modelo ARIMA (0,1,3)(0,1,3)[12].

Figura 17 – Previsão da carga horária para 2020 e 2021 conforme modelo ARIMA (0,1,3)(0,1,3)[12]



Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Para efeito de comparação dos resultados e posterior avaliação do modelo aplicado, resolveu-se construir uma tabela para comparação das projeções realizadas pelo modelo SNAIVE, pelo auto.arima e pelo modelo ARIMA (0,1,3)(0,1,3)[12]. A tabela 2, apresentada a seguir, demonstra os resultados alcançados em cada modelo analisado.

Tabela 2 – Projeção da carga horária válida conforme modelos analisados

Modelo	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total Ano
SNAIVE	368	748	1528	1190	1785	1862	1054	998	1690	1191	2018	1774	16206
AUTO.ARIMA	1077	814	1479	1267	1728	1732	1253	1358	1330	1510	1868	1927	17343
ARIMA 2020	442	588	2476	1363	2312	1461	1031	1244	1355	1691	2116	1793	17872
ARIMA 2021	249	1154	2709	1692	2437	1524	1191	1099	2110	1594	2082	1305	19146

Fonte: elaborado pelo autor com base nos dados do PFC/Seplag.

Após a apresentação do cenário preditivo, compreende-se nítida divergência entre os modelos. No entanto, a validação dos modelos se dará mediante a comparação com os dados que forem disponibilizados ao longo do ano de 2020. A próxima seção do estudo é dedicada a uma discussão sobre os resultados alçados com a realização do trabalho.

## 6 Apresentação dos Resultados

Discorrer sobre o porvir em um contexto de complexidade e fluidez pode soar como uma insensatez. Isto parece se confirmar diante do atual cenário, em que a mutação de um vírus na China impacta profundamente a economia mundial, transformando planos e projetos robustos, construídos com significativo esforço em artefatos rapidamente ultrapassados, que apenas retratam histórias não vividas.

A construção de um estudo baseado em séries temporais no cenário atual parece representar um destes artefatos. Idealizado ainda em dezembro de 2019, quando ainda não havia evidência do surgimento da pandemia do COVID-19, o presente trabalho implicou em mais um desafio a ser superado pelo pesquisador, ainda neófito no campo de programação computacional. O contexto disruptivo atual, que pode facilmente se caracterizar como o que Taleb (2015) denominou de “cisne negro”, ou seja, fenômenos improváveis que ocorrem gerando profundo impacto, implicando na necessidade de estudos complementares, caracterizando o desafio apontado por Ferreira (2018) de se analisar a construção de séries temporais com alterações conjunturais ou quebras estruturais em determinado instante de tempo, o que não era o foco inicial do presente estudo e nem caberia no curto espaço de tempo disponível para sua realização.

Diante disto, resta o clamor pela compreensão de que os resultados do trabalho que foi desenvolvido ao longo de quatro meses possa ser avaliado como um esboço sobre a temática das séries temporais. Para que fosse possível demonstrar a linha de raciocínio que possibilitou a construção do estudo foi elaborada a figura 18 que representa um *canvas*, conforme o modelo de Dorard (2020). Nele é observada uma síntese dos principais elementos componentes do projeto que foi desenvolvido, cuja ênfase desta seção se resume à discussão dos resultados alcançados.

A princípio, cabe lembrar que o objetivo do presente estudo foi o de dimensionar a carga horária de ações de capacitação para o PFC/Seplag com base em um estudo de série temporal para os anos de 2020 e 2021. Também é importante destacar que não havia previsão na condução do referido programa, mas tão somente a repetição dos quantitativos de carga horária realizadas no ano anterior, semelhante ao modelo SNAIVE.













Figura 18 – Canvas do projeto

**The Machine Learning Canvas (v0.4)**

Designed for:

Designed by: José Alberto de S. Brandão Date: 30/03/2020 Iteration: 3

<b>Decisions</b>  How are predictions used to make decisions that provide the proposed value to the end-user?  Atualmente não são feitas previsões, mas tão somente a repetição dos quantitativos de carga horária realizadas no ano anterior.	<b>ML task</b>  Input, output to predict, type of problem.  As entradas correspondem ao registro de cada ação de capacitação realizada pelos gestores governamentais vinculados à SEPLAG/PE, com respectivas cargas horárias. As saídas devem representar a estimativa do quantitativo anual acumulado de carga horária para os próximos anos.	<b>Value Propositions</b>  What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?  Pretende-se realizar um melhor dimensionamento da carga horária das ações de capacitação, possibilitando a observância de períodos de maior e de menor intensidade na ocorrência destes eventos de modo a melhor ajustar os recursos necessários à sua execução, possibilitando a melhoria da qualidade percebida em relação aos processos relativos ao PFC.	<b>Data Sources</b>  Which raw data sources can we use (internal and external)?  As principais fontes de dados brutos que podem ser acessados são as planilhas contendo os registros individuais das ações de capacitação realizadas pelos GGPOGs, como também a planilha disponibilizada mensalmente pela Gerência de Pessoas, que contempla a movimentação de pessoal, registrando eventuais cessões, licenças, afastamentos e exonerações.	<b>Collecting Data</b>  How do we get new data to learn from (inputs and outputs)?  Os novos dados para aprendizagem emanarão da comparação entre a predição fornecida pelo modelo de séries temporais e a comparação com os resultados efetivamente alcançados a partir do ano de 2020.	
<b>Making Predictions</b>  When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?  Apesar de novos dados poderem chegar a qualquer momento, em função dos processos de validação de ações de capacitação, a compilação dos dados só é realizada efetivamente ao término de cada ano, para efeito de elaboração de relatórios.	<b>Offline Evaluation</b>  Methods and metrics to evaluate the system before deployment.  Antes da implantação da análise de série temporal não existem modelos para avaliação do sistema. Existe apenas um processo de avaliação da qualidade das ações de capacitação ofertadas.	<b>Features</b>  Input representations extracted from raw data sources.  Os dados estão disponíveis no formato de planilhas eletrônicas anuais.			<b>Building Models</b>  When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?  Os dados precisam ser condensados ao longo do período de um ano, de modo que o prazo para que seja possível ajustar a análise da série temporal é de pelo menos um ano.
<b>Live Evaluation and Monitoring</b>  Methods and metrics to evaluate the system after deployment, and to quantify value creation.  A partir da construção da série temporal seria possível verificar sua acurácia ao compará-la com os dados efetivamente obtidos. Conforme houvesse aproximação dos resultados alcançados em relação ao previsto seria possível a confirmação do valor criado com a modelagem. Também seria possível observar o volume de recursos dispendidos antes e depois do modelo, de modo a se verificar se foi possível alcançar ganhos econômicos com a modelagem.					

machinelearningcanvas.com by Louis Dorard, Ph.D.

Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License



Fonte: elaborada pelo autor com base no modelo Dorard (2020).

Para construção da série temporal é importante conhecer o processo de entrada de dados, que ocorre mediante a disponibilização de arquivos em planilha eletrônica, anualmente consolidados. Tais planilhas correspondem ao registro de cada ação de capacitação realizada pelos gestores governamentais vinculados à SEPLAG/PE, com respectivas cargas horárias.

Com o desenvolvimento do estudo foi possível compreender os mecanismos geradores da série temporal, com a inserção contínua de dados sobre ações de capacitação realizadas por GGPOGs. Com isso, foram observadas as características da existência de uma série temporal, ou seja, a presença dos elementos de tendência, a sazonalidade, caracterizada com a oscilação entre picos e vales ao longo do tempo, e o ciclo anual identificado neste processo. Desta maneira, foi possível a descrição do comportamento da série e a predição do comportamento para os anos de 2020 e 2021.

A consecução do estudo de série temporal possibilita um melhor dimensionamento da carga horária das ações de capacitação, permitindo a observância de períodos de maior e de menor intensidade na ocorrência destes eventos. Desta maneira, tornou-se possível realizar um melhor ajuste dos recursos

necessários à execução do PFC/Seplag, como por exemplo a redefinição da escala de férias dos servidores que trabalham neste programa. Com isso, espera-se no médio prazo que haja a melhoria da qualidade percebida em relação aos processos relativos ao PFC.

A partir da construção da série temporal e da possível comparação do modelo proposto com os dados reais a serem obtidos ao longo dos anos de 2020 e 2021 poderá haver um processo de aprendizagem do modelo, com a simulação antecipada de demanda. Com isso, retrata-se o desafio a ser desenvolvido *a posteriori* no sentido de realizar esta atividade de monitoramento contínuo dos novos dados e inserção no modelo para reavaliá-lo.

Deste modo, será possível confirmar a acurácia do modelo em relação às práticas atualmente desenvolvidas. Havendo aproximação dos resultados alcançados em relação ao previsto será possível a confirmação de valor criado com a modelagem. Também será possível dimensionar o volume de recursos despendidos antes e depois do modelo, de modo a se verificar se foi possível alcançar ganhos econômicos com a modelagem.

## 7 Links

Conforme solicitado, seguem abaixo os links para acesso ao vídeo de apresentação e para o repositório contendo o *script* criado no R para elaboração da série temporal e os dados utilizados no projeto.

<https://github.com/albertobrandao/CursoCDBD/>

## REFERÊNCIAS

AQUINO, J. **R para cientistas sociais**. Ilhéus: EDITUS, 2014.

BOX, G.; JENKINS, G. **Time series analysis**: forecasting and control. San Francisco: Holden-Day Brockwell, 2008.

DORARD, L. **The machine learning canvas**. Disponível em: <https://www.louisdorard.com/machine-learning-canvas>. Acesso em: 26/03/2020.

FERREIRA, P. (Org.). **Análise de séries temporais em R**: um curso introdutório. Rio de Janeiro: Elsevier FGV IBRE, 2018.

GROLEMUND, G.; WICKHAM, H. **R for data science**: import, tidy, transform, visualize, and model data. Sebastopol: O'Reilly, 2017.

LONG, J.; TEETOR, P. **R cookbook**: proven recipes for data analysis statistics & graphics. Sebastopol: O'Reilly, 2019.

PERNAMBUCO. Assembléia Legislativa. Decreto nº. 37.828, de 02 de fevereiro de 2012. Cria o Instituto de Gestão Pública de Pernambuco - "GESTÃO/PE" Instituto de Gestão Pública de Pernambuco Governador Eduardo Campos na estrutura organizacional da Secretaria de Planejamento e Gestão do Estado de Pernambuco. Diário Oficial do Estado de Pernambuco, Recife, PE, 26 abr. 2012.

\_\_\_\_\_. Lei complementar nº. 117, de 26 de junho de 2008. Dispõe sobre a criação da Carreira de Gestão Administrativa e seus cargos, fixa sua remuneração, e dá outras providências. Diário Oficial do Estado de Pernambuco, Recife, PE, 3 set. 2009b.

\_\_\_\_\_. Lei complementar nº. 118, de 26 de junho de 2008. Dispõe sobre a criação da Carreira de Planejamento, Orçamento e Gestão e seus cargos, fixa sua remuneração, e dá outras providências. Diário Oficial do Estado de Pernambuco, Recife, PE, 3 set. 2009c.

\_\_\_\_\_. Lei complementar nº. 119, de 26 de junho de 2008. Dispõe sobre a criação da Carreira de Controle Interno e seus cargos, fixa sua remuneração, e dá outras providências. Diário Oficial do Estado de Pernambuco, Recife, PE, 3 set. 2009d.

\_\_\_\_\_. Lei complementar nº. 141, de 3 de setembro de 2009. Dispõe sobre o modelo integrado de gestão do poder executivo do Estado de Pernambuco. Diário Oficial do Estado de Pernambuco, Recife, PE, 3 set. 2009a.

PERNAMBUCO. Secretaria de Planejamento e Gestão - SEPLAG. **Fundamentos da gestão pública**: programação com R. Disponível em: <http://gc.seplag.pe.gov.br/s/bp61pakunvgn9c4bg7ug/fundamentos-da-gestao-publica/d/bpe6fakunvgn9c4bgbh0/r>>. Acesso em 18/03/2020.

PETER, J.; DAVIS, R. **Introduction to time series and forecasting**. Springer-Verlag, 2002.

ROME, C. **Time series markdown**. Disponível em: < [http://rstudio-pubs-static.s3.amazonaws.com/387852\\_8d49b434b09a47ce959c253ef6e6607b.html](http://rstudio-pubs-static.s3.amazonaws.com/387852_8d49b434b09a47ce959c253ef6e6607b.html)>. Acesso em 18/03/2020.

TALEB, N. **A lógica do cisne negro**: o impacto do altamente improvável. Rio de Janeiro: Best Seller, 2015.

VENABLES, W.; SMITH, D.; R Core Team. **An Introduction to R**. Sebastopol: O'Reilly, 2020.

## APÊNDICE – SCRIPT DO TCC EM R

```
##### PONTIFICIA UNIVERSIDADE CATOLICA DE MINAS GERAIS - PUC/MG #####
##### PÓS-GRADUACAO EM CIENCIA DE DADOS E BIG DATA #####
##### TRABALHO DE CONCLUSAO DE CURSO #####
##### AUTOR: JOSE ALBERTO DE SIQUEIRA BRANDAO #####

##### SCRIPT #####

##### ANALISE PREDITIVA DA CARGA HORARIA DE UM PROGRAMA DE FORMACAO NO
SETOR PUBLICO COM USO DE SERIES TEMPORAIS #####

##### INSTALANDO PACOTES REQUERIDOS #####

install.packages('readxl')
install.packages('UsingR')
install.packages('xts')
install.packages('dygraphs')
install.packages('fpp2')
install.packages('forecast')
install.packages('urca')
install.packages('tseries')
install.packages('lmtest')
install.packages('FinTS')
install.packages('normtest')

##### ETL #####

# define o diretorio onde estao os arquivos

setwd('E:/TCC_DATA')

# carga dos arquivos anuais do PFC

library(readxl)

d14 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2014.xlsx", sheet="BD_PFC_(2014)")
d15 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2015.xlsx", sheet="BD_PFC_(2015)")
d16 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2016.xlsx", sheet="BD_PFC_(2016)")
d17 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2017.xlsx", sheet="BD_PFC_(2017)")
d18 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2018.xlsx", sheet="BD_PFC_(2018)")
d19 <- read_excel("E:/TCC_DATA/PFC_ANUAL/PFC_2019.xlsx", sheet="BD_PFC_(2019)")

# empilhamento dos dados por linha criando um dataset ds

ds <- rbind(d14,d15,d16,d17,d18,d19)

# construção de subconjunto com GGOVs

dsg = subset(ds, TIPOSERV != "Não" & CPF != "0")
attach(dsg)

##### ESTATISTICA DESCRITIVA #####

# calcula medidas de posicao

summary(CHV)

# calcula medidas de dispersao (variância e desvio padrao)
```

```

var(CHV)
sd(CHV) # ou sqrt(var(dsg$CHV))

# calculando a amplitude total

ampdsg <- max(CHV)-min(CHV)
ampdsg

# calculando o coeficiente de variacao

sd(CHV)/mean(CHV)*100

##### GRÁFICOS - ESTATISTICA DESCRITIVA #####

# construindo histograma de observacoes por ano

library(UsingR)

# construindo graficos de pizza

par(mfrow=c(1,2))

#Grafico1

chtipo <- table(TIPO)
chtipoval <- signif(chtipo/sum(chtipo)*100, 3)
chtipoval
pie(chtipoval, main = "Tipo de capacitacao", labels = paste(chtipoval,"%",sep=""), col =
c("steelblue1","dodgerblue3","orange", "lightgreen", "red", "pink", "palevioletred2", "green", "gold"))
texttot <- c("Congresso", "Curso", "Disciplina", "Forum", "Instrutoria", "Oficina", "Palestra", "Seminario",
"Simposio")
legend(x = "topright", legend = texttot, fill = c("steelblue1","dodgerblue3","orange", "lightgreen", "red",
"pink", "palevioletred2", "green", "gold"), cex = 0.50)

#Grafico2

chforma <- table(FORMA)
chformaval <- signif(chforma/sum(chforma)*100, 3)
chformaval
pie(chformaval, main = "Forma de capacitacao", labels = paste(chformaval,"%",sep=""), col =
c("steelblue1","dodgerblue3","green"))
textof <- c("Instrutoria", "Ofertado", "Validado")
legend(x = "topright", legend = textof, fill = c("steelblue1","dodgerblue3","green"), cex = 0.65)

#Grafico3

chmodal <- table(MODALIDADE)
chmodalval <- signif(chmodal/sum(chmodal)*100, 3)
chmodalval
pie(chmodalval, main = "Modalidade de capacitacao", labels = paste(chmodalval,"%",sep=""), col =
c("steelblue1","dodgerblue3"))
textom <- c("EAD", "Presencial")
legend(x = "topright", legend = textom, fill = c("steelblue1","dodgerblue3"), cex = 0.65)

#Grafico4

chlota <- table(LOTACAO)
chlotalval <- signif(chlota/sum(chlota)*100, 3)
chlotalval

```

```
pie(chlotval, main = "Lotacao", labels = paste(chlotval,"%",sep=""), col =
c("steelblue1","dodgerblue3","orange", "lightgreen", "red", "pink", "palevioletred2", "green", "gold",
"gray"))
textol <- c("Cedido", "Condepe", "Gabinete", "IGPE", "SEAM", "SECOGE", "SEGES", "SEPOC",
"SEGPR", "SGTG")
legend(x = "topright", legend = textol, fill = c("steelblue1","dodgerblue3","orange", "lightgreen", "red",
"pink", "palevioletred2", "green", "gold", "gray"), cex = 0.50)
```

#### ##### AGRUPAMENTO DOS DADOS #####

```
# acumular dados por mes e ano
```

```
tempo_POSIX <- strptime(DATAI, format = "%d.%m.%Y %H:%M", tz = "GMT")
x <- as.POSIXct(c(DATAI))
da <- strptime(x, "01")
mo <- strptime(x, "%m")
yr <- strptime(x, "%Y")
chv <- runif(3)
dd <- data.frame(da, mo, yr, CHV)

dsg.agr <- aggregate(CHV ~ da + mo + yr, dd, FUN = sum)
detach(dsg)
attach(dsg.agr)
dsg.agr$DMA <- as.POSIXct(paste(dsg.agr$yr, dsg.agr$mo, dsg.agr$da, sep = "-"))
```

```
# gerando gráfico de série temporal mensal e anual
```

```
par(mfrow = c(1, 2))
```

```
boxplot(dsg.agr$CHV~dsg.agr$mo, xlab="Mes", ylab = "Carga Horaria", main ="Carga Horaria Mensal
- 2014-2019")
```

```
boxplot(dsg.agr$CHV~dsg.agr$yr, xlab="Mes", ylab = "Carga Horaria", main ="Carga Horaria Anual -
2014-2019")
```

#### ##### CRIACAO DA SERIE TEMPORAL #####

```
library(xts)
```

```
st <- ts(CHV, start = c(2014,01), end = c(2019,12), frequency = 12)
st
```

```
# grafico em grade da serie tempoal
```

```
par(mfrow = c(1, 1))
```

```
library(dygraphs)
dygraph(st,xlab="Ano", ylab = "Carga horaria",main="Serie Temporal - 2014-2019")
```

```
#gráfico mensal radar
```

```
library(fpp2)
ggseasonplot(st, season.labels = NULL,
  year.labels = FALSE,
  year.labels.left = FALSE,
  main = "Distribuição da Série Temporal por Ano e Mês",
  continuous = FALSE,
  polar = TRUE,
```

```

        labelgap = 0.25)

#gráfico de variacao mensal

monthplot(st, xlab="Ano", ylab = "Carga horaria",main="Variacao Mensal da Serie Temporal - 2014-2019")

##### PROJECAO SERIE TEMPORAL - MODELO SNAIVE #####

library(forecast)

summary(snaive(st,h=12))
plot(snaive(st,h=12),include=200, xlab="Ano", ylab = "Carga horaria",main="Projecao da Serie Temporal conforme Modelo SNAIVE- 2020")

##### MODELAGEM AUTO-ARIMA SERIE TEMPORAL ORIGINAL #####

autoARIMAst <- auto.arima(st)
autoARIMAst
prevARIMAst <- forecast(autoARIMAst, h=12)
prevARIMAst
plot(prevARIMAst, xlab="Ano", ylab = "Carga horaria",main="Projecao da Serie Temporal conforme Modelo ARIMA - 2020")

##### DECOMPOSICAO DA SERIE TEMPORAL #####

dec <- decompose(st)
plot(dec)

lag.plot(st, lags = 4, do.lines = FALSE)

ggtsdisplay(st,xlab="Ano", main="Serie temporal 2014-2019 com ACF e PACF")

##### TESTE DICKEY-FURLEY AUMENTADO APLICADO A SERIE TEMPORAL ORIGINAL #####

library(urca)
library(tseries)

adf.drift <- ur.df(y = st, lags = 24, selectlags = "AIC")
adf.test(st, alternative="stationary", k=0)
acf(adf.drift@res)
adf.drift@teststat
adf.drift@cval #valores tabulados por MacKinnon (1996)
summary(adf.drift)

##### TRANSFORMACAO DA SERIE TEMPORAL (COM DIFERENCIACAO LOGARITMICA) #####

ggtsdisplay(diff(log(st)),xlab="Ano", main="Serie temporal Transformada 2014-2019 com ACF e PACF")

##### TESTE DICKEY-FURLEY AUMENTADO APLICADO A SERIE TEMPORAL TRANSFORMADA (COM DIFERENCIACAO LOGARITMICA) #####

adf.driftd <- ur.df(y = diff(log(st)), type = c("drift"), lags = 24, selectlags = "AIC")
adf.test(diff(log(st)), alternative="stationary", k=0)

```



```
adf.driftd@teststat
adf.drift@cval #valores tabulados por MacKinnon (1996)
summary(adf.driftd)
```

```
##### IDENTIFICACAO DA SERIE TEMPORAL TRANSFORMADA #####
```

```
library(lmtest)
```

```
Box.test(diff(log(st)), lag = 24, type = "Ljung-Box")
```

```
##### MODELAGEM AUTO-ARIMA SERIE TEMPORAL TRANSFORMADA #####
```

```
autoARIMAst<- auto.arima(diff(log(st)))
autoARIMAst<- forecast(autoARIMAst, h=12)
plot(autoARIMAst, xlab="Ano", ylab = "Carga horaria",main="Projecao da Serie Temporal
Transformada conforme Modelo ARIMA - 2020")
```

```
##### ESTIMACAO DA SERIE TEMPORAL #####
```

```
##### MODELAGEM ARIMA SERIE TEMPORAL TRANSFORMADA #####
```

```
fit.modelARIMA <- arima(log(st), c(0,1,3), seasonal = list(order=c(0,1,3), period=12))
summary(fit.modelARIMA)
```

```
##### DIAGNOSTICO DO MODELO ARIMA (0,1,3) DA SERIE TEMPORAL TRANSFORMADA #####
```

```
tsdiag(fit.modelARIMA, gof.lag = 20)
```

```
Box.test(x = fit.modelARIMA$residuals, lag = 24, type = "Ljung-Box", fitdf = 2)
```

```
library(FinTS)
```

```
ArchTest(fit.modelARIMA$residuals,lags = 24)
```

```
library(normtest)
```

```
jB.norm.test(fit.modelARIMA$residuals, nrepl=1000)
```

```
AIC(autoARIMAst)
AIC(fit.modelARIMA)
```

```
logLik(autoARIMAst)
logLik(fit.modelARIMA)
```

```
accuracy(autoARIMAst)
accuracy(fit.modelARIMA)
```

```
##### PREVISAO #####
```

```
pred <- predict(fit.modelARIMA, n.ahead = 24, level = 0.95)
ts.plot(st,exp(pred$pred), log = "y", lty = c(1,3),xlab="Ano", ylab = "Carga horaria",main="Projecao
Final da Serie Temporal 2020/2021 Modelo ARIMA (0,1,3)")
```

```
previsao <- exp(pred$pred)
previsao
```

```
##### FIM #####
```