DOI: 10.48011/asba.v2i1.990

Modelagem dinâmica para previsão dos casos novos de COVID-19 no Estado do Paraná

Sheila R. Oro * **Liliane Hellmann** * **Tereza R. Mafioleti** *
Camila N. B. Di Domênico * **Guilherme L. Campos** **

* *Departamento de Física, Estatística e Matemática, Universidade Tecnológica Federal do Paraná, PR, (e-mail: sheilaro@utfpr.edu.br; lilianehellmann@utfpr.edu.br; mafioleti@utfpr.edu.br; camiladomenico@utfpr.edu.br).*
** *Graduando em Engenharia Química, Universidade Tecnológica Federal do Paraná, PR (e-mail: guilherme_lopescampos@hotmail.com)*

Abstract: Since March 2020 Paraná registered cases of an infectious disease caused by the new coronavirus SARS-COV-2, the COVID-19. The speed of disease spread is not constant and requires a regular review of estimates for the future number of infected. In this regard, the use of statistical methods can help in assessment of the disease evolution and projections of disease spread. This study aimed at modeling and forecast the number of new daily cases of COVID-19, in the State of Paraná, by Autoregressive Distributed Lag Model. The ocourrence of delays in new cases reports, in cases under investigation and in the transmissibility was considered. The results indicated that the adjusted model is able to predict the number of new daily cases of the disease, with a 14-day accuracy.

Resumo: Desde o mês de março de 2020 o Paraná tem registrado casos de uma doença infecciosa causada pelo novo coronavírus SARS-COV-2, a COVID-19. A velocidade da propagação da doença não é constante, o que requer frequente revisão dos cálculos das estimativas para o número futuro de infectados. Neste sentido, o uso de métodos estatísticos pode auxiliar na avaliação da evolução e nas projeções da propagação da doença. Este estudo, objetivou realizar a modelagem e previsão do número de casos novos diários de COVID-19, no Estado do Paraná, por meio dos modelos autorregressivos de defasagens distribuídas, considerando o atraso na divulgação dos casos novos, dos casos em investigação e a taxa de transmissão. Os resultados indicaram que o modelo ajustado é capaz de prever o número de casos novos diário da doença, com boa precisão para 14 dias.

Keywords: Time Series; Autoregressive Distributed Lag Model; Coronavirus; Effective reproduction number.

Palavras-chaves: Séries temporais; Modelos autorregressivos de defasagens distribuídas; Coronavírus; Número reprodutivo efetivo;

1. INTRODUÇÃO

Desde o final do ano de 2019, o mundo acompanha com preocupação a disseminação da COVID-19, causada pelo novo coronavírus SARS-COV-2. O estado de pandemia foi decretado pela Organização Mundial da Saúde em março de 2020. Conforme a doença foi se espalhando pelo mundo, os governantes passaram a implementar medidas de mitigação, para evitar o colapso dos sistemas de saúde e minimizar o número de vítimas, além de se preocupar com os danos econômicos, sociais e culturais para a população.

O número de pessoas contaminadas e a velocidade de propagação da doença justifica esta preocupação. O primeiro caso confirmado no mundo, foi em 31 de dezembro de 2019, na China e, em 09 de agosto de 2020, mais de 20 milhões de pessoas haviam sido infectadas por este coronavírus (WHO, 2020). No Brasil, o primeiro caso de COVID-19 foi registrado no dia 25 de fevereiro de 2020 e, em 08 de agosto, mais de 3 milhões de pessoas já tinham recebido o diagnóstico no país (MS, 2020). O Paraná apresentou o primeiro caso no dia 12 de março e, em 14 de agosto, havia ultrapassado 100 mil pessoas com diagnóstico confirmado desta doença (SSEPR, 2020).

A expansão acelerada do número de pessoas acometidas por esta doença, juntamente com o desconhecimento das suas características de disseminação motivaram inúmeras pesquisas. Muitos estudos epidemiológicos utilizam modelos matemáticos (Bastos and Cajueiro, 2020; Franco and Dutra, 2020; Silva, 2020; Nadler et al., 2020; Palladino et al., 2020; Wangping et al., 2020) ou estatísticos (Aslam, 2020; Ribeiro et al., 2020; Abdulmajeed et al., 2020) para avaliar a evolução da doença em diversos países e realizar projeções da propagação.

A taxa de transmissão de doenças infecciosas, como a COVID-19, varia à medida que algumas pessoas são infectadas, se recuperam e tornam-se resistentes ao vírus. Por isso, esta taxa deve ser reconsiderada constantemente, uma

vez que o tamanho da população suscetível se altera com o tempo. Uma forma de estimá-la é por meio do número reprodutivo efetivo (Re) que, de acordo com Wallinga and Teunis (2004), para o caso j, é a soma de todos os casos i, ponderada pela probabilidade relativa de que o caso i tenha sido infectado pelo caso j, conforme a equação:

$$Re_j = \sum_i p_{ij} \tag{1}$$

onde p_{ij} é a probabilidade relativa do caso i ter sido infectado pelo caso j, considerando a sua diferença no tempo do início dos sintomas $t_i - t_j$.

Esta probabilidade é normalizada pela chance de que o caso i tenha sido infectado por qualquer outro caso k. Sendo assim, p_{ij} pode ser expresso em termos da distribuição de probabilidade do intervalo de geração τ, conforme a equação:

$$p_{ij} = \frac{\tau(t_i - tj)}{\sum_{i \neq k}(t_i - tj)} \tag{2}$$

Este método de estimação de Re foi implementado computacionalmente por (Cori et al., 2013) e está disponível no *software* R por meio do pacote *EpiEstim* (Cori, 2020). Este pacote tem sido utilizado por diversos estudiosos em suas pesquisas acerca de previsões relacionadas com a disseminação da COVID-19 (Abbott et al., 2020; Hu, 2020; Dana et al., 2020; Najafi et al., 2020; Al Wahaibi et al., 2020; Menendez, 2020)

As quantidades de casos novos de COVID-19 e de casos em investigação registradas periodicamente, constituem duas séries temporais. Em geral, estas séries são não estacionárias. Neste caso, o uso de modelos dinâmicos, tais como o Autorregressivo de Defasagens Distribuídas (ARDL), que consideram o retardo nos valores observados e existência de fatores condicionantes, têm obtido resultados promissores quando aplicados a séries temporais na área da saúde (Al-Mulali et al., 2016; Sarkodie and Owusu, 2020).

Os modelos ARDL utilizam os valores presentes e passados da variável resposta (dependente) e suas preditoras (independentes). Um modelo ARDL(r, s) com r defasagens para a variável dependente (Z_t), s para a independente (x_t) e com um termo constante (μ) pode ser representado por:

$$Z_t = \mu + \sum_{i=1}^{r} \alpha_i y_{t-i} + \sum_{i=0}^{s} \beta_i x_{t-i} + \epsilon_i \tag{3}$$

Este modelo 3 supõe resíduos independentes, normalmente distribuídos e com variância constante. Para sua especificação é preciso determinar a ordem das defasagens para todas as variáveis, a estimação dos coeficientes.

Como as informações fornecidas pelos órgãos de saúde são limitadas, pois não é possível saber exatamente o dia em que houve o contágio, muitas vezes, os pesquisadores baseiam-se na data de divulgação dos resultados dos exames para extrair os dados e realizar a modelagem. Mas, geralmente, os números informados estão defasados, devido à diferença entre a data do resultado do exame, seu lançamento no sistema gerenciador de dados, o horário da coleta das informações e sua divulgação. Nesse contexto, o principal objetivo deste trabalho é obter um modelo dinâmico capaz de estimar o número de casos diários de COVID-19 no estado do Paraná, Brasil, considerando os possíveis atrasos dos dados divulgados pela Secretaria Estadual da Saúde, o número de casos investigados e as estimativas da taxa de transmissão. A acurácia do modelo é medida pela Média Percentual Absoluta do Erro (MAPE).

2. MATERIAIS E MÉTODOS

Este estudo considerou o número de casos novos diários de COVID-19 e de casos investigados, no Estado do Paraná, divulgados pela Secretaria de Estado da Saúde do Paraná nos boletins epidemiológicos emitidos no período de 10 de maio a 14 de agosto de 2020 (SSEPR, 2020). Os casos em investigação representam os suspeitos de COVID-19 que tiveram amostras coletadas e processadas pelo Lacen-PR e laboratórios privados habilitados no Estado.

A evolução de casos novos e em investigação desta doença no Paraná, entre a vigésima e a trigésima segunda semanas epidemiológicas (10 de maio a 01 de agosto), é apresentada na Figura 1. Percebe-se uma alteração na tendência crescente aproximadamente a partir da vigésima quarta semana (31 de maio), a partir de quando os números passaram a evoluir de forma acentuada, sem indicação de estabilização.

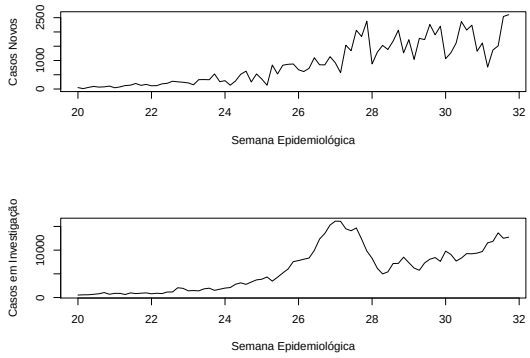

Figura 1. Casos novos e em investigação de COVID-19 entre 10 de maio e 31 de julho de 2019, no Paraná-BR.

A taxa de transmissão da doença também foi considerada neste estudo, por meio do número reprodutivo efetivo (Re). Devido à dificuldade em encontrar uma série temporal diária referente a esta taxa, valores de Re foram estimados com base no número de casos novos diários no período, por meio do pacote *EpiEstim*, disponível no *software* R (Cori, 2020). A Tabela 1 apresenta as medidas de posição das taxas de transmissão diárias estimadas para os meses de maio, junho e julho. De acordo com a Tabela 1, os valores estimados para Re variaram entre $0,90$ e $1,63$, sendo que o valor máximo pode ser considerado atípico dentre os demais do conjunto , conforme evidenciado no histograma (Fig. 2). Com base nos quartis ($Q1$ e $Q3$), em aproximadamente metade dos dias analisados a taxa oscilou entre 1,02 e 1,20. Além disso, pode-se supor que Re apresentou uma distribuição aproximadamente simétrica no período analisado, pois a média e a mediana

ficaram muito próximas 1, 12 e 1, 10, respectivamente, e as classes de maior frequência ficaram na porção central do histograma, se desconsiderarmos o valor atípico. Como o primeiro quartil estimado ($Q1$) foi de 1,02, isto evidencia que na maior parte dos dias (75%) a taxa ficou acima de 1, indicando aumento na disseminação da doença.

Tabela 1. Medidas descritivas da taxa de transmissão diária estimada da COVID-19, entre maio e julho de 2020, no Paraná-BR.

Mín	Q1	Med	Média	Q3	Máx
0,90	1,02	1,10	1,12	1,20	1,63

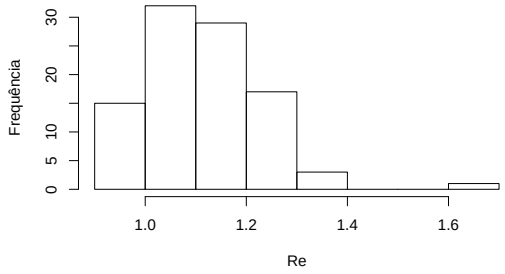

Figura 2. Distribuição da taxa de transmissão (Re) de COVID-19 no Paraná, entre os meses de maio e julho de 2020.

Para fins de verificação da capacidade preditiva do modelo, separou-se o conjunto de dados (casos novos, casos em investigação e taxa de transmissão) em duas partes. Na primeira, utilizada para o ajuste do modelo, foram considerados os dados da 20ª até a 32ª semana epidemiológica (10 de maio a 31 de julho). A segunda, composta pelos números registrados no período de 01 a 14 de agosto, foram reservados para a validação do modelo.

A sequência das atividades realizadas no desenvolvimento deste estudo estão apresentadas na Figura 3. Todas as análises, modelagens e testes foram feitas com o auxílio do *software* R (R Core Team, 2019). As etapas de coleta de dados e estimação da taxa Re já foram descritas no início desta seção. A verificação da estacionariedade das séries temporais foi feita por meio do teste de Dickey-Fuller aumentado (ADF) disponível no pacote *tseries* (Trapletti and Hornik, 2019). A estimação das defasagens foi obtida pela abordagem denominada *Bounds testing*, segundo o critério de informação de Akaike (AIC), disponível no pacote *dLagM* (Demirhan, 2020), com número máximo de 7 defasagens para cada variável e, na sequência, foi ajustado o modelo ARDL com os parâmetros selecionados e verificados os pressupostos para os resíduos. Os novos dados dos casos em investigação e novas taxas de transmissão foram simulados por meio de reamostragem, com reposição. Com os valores simulados e o modelo ajustado, foram realizadas as previsões para os casos novos de COVID-19. Os valores previstos foram comparados com os reais (atuais) por meio do teste de Dunnett. A validação do modelo foi feita por meio da medida de acurácia obtida pela Média Percentual Absoluta do Erro (MAPE). Os detalhes a respeito do método de modelagem ARDL e dos testes utilizados podem ser encontrados em Nkoro et al. (2016).

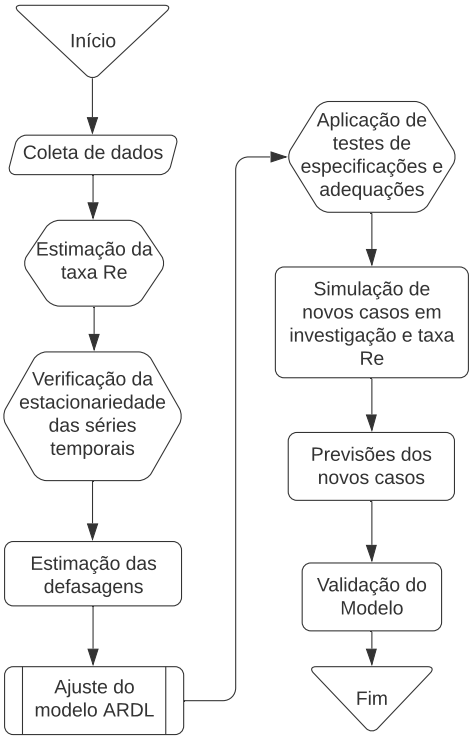

Figura 3. Etapas da pesquisa.

3. RESULTADOS E DISCUSSÃO

Nesta seção são apresentados os resultados obtidos para a modelagem e previsão da série temporal do número de casos diários da COVID-19 no estado do Paraná, pelo modelo ajustado.

O teste ADF aplicado para as séries diárias de casos novos de COVID-19 e de casos em investigação, indicou que ambas podem ser caracterizadas como um passeio aleatório com possível deslocamento e tendência, ao nível de significância de até 1% ($prob < 0.0001$), sendo integradas de ordem 1. Este resultado viabilizou a utilização dos modelos Autorregressivos de Defasagens Distribuídas (ARDL).

O modelo selecionado pela abordagem *Bounds testing*, foi o $ARDL(7, 7, 7)$, isto é, 7 defasagens para cada variável. Este resultado sugere que a quantidade de novos casos diários de COVID-19 (CN) no Paraná depende dos próprios números registrados na última semana, da quantidade de pessoas com suspeita da doença (Su) e da taxa de transmissão dos últimos 7 dias (Re). Este foi o modelo ARDL que apresentou o menor valor para o AIC $(1033, 146)$, dentre todos os testados, com uma capacidade de explicar cerca de 95% da variação presente no número diário de casos novos de COVID-19.

O modelo ajustado, representado pela Equação (4), obteve resultados satisfatórios quanto aos pressupostos dos resíduos, uma vez que a hipótese de normalidade foi aceita ($prob > 0, 20$), assim como a de homogeneidade ($prob > 0, 08$), ao nível de significância de 5% e pela análise gráfica, os resíduos foram considerados independentes (Figura 4), pois flutuaram aleatoriamente em torno da média zero.

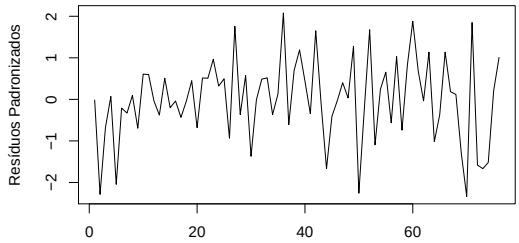

Figura 4. Resíduos padronizados do modelo.

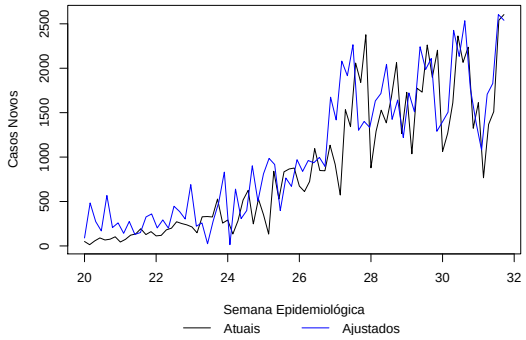

Figura 5. Comparação entre as séries temporais dos casos novos e dos valores ajustados pelo modelo.

$$
\begin{aligned}
CN_t = & -564,60 - 0,04Su_t + 3320Re_t + 0,24CN_{t-1} + \\
& 0,27CN_{t-2} - 0,17CN_{t-3} - 0,31CN_{t-4} + \\
& 0,11CN_{t-5} + 0,41CN_{t-6} + 0,39CN_{t-7} - \\
& 0,01Su_{t-1} + 0,06Su_{t-2} + 0,06Su_{t-3} - \\
& 0,12Su_{t-4} + 0,12Su_{t-5} - 0,01Su_{t-6} - \\
& 0,01Su_{t-7} - 3067Re_{t-1} + 105,7Re_{t-2} + \\
& 359,7Re_{t-3} + 1111Re_{t-4} - 1223Re_{t-5} - \\
& 299,3Re_{t-6} + 226Re_{t-7}
\end{aligned}
\tag{4}
$$

Os valores ajustados pelo modelo $ARDL(7,7,7)$ para os casos novos de COVID-19 no Paraná, entre a 20ª e a 32ª semanas epidemiológicas, encontram-se representados no gráfico 5. Percebe-se que o modelo foi capaz de representar satisfatoriamente a série real ($MAPE = 29,38\%$), com tendência a superestimação e um erro.

Sendo assim, foram feitas 10 simulações para as previsões usando o modelo $ARDL(7,7,7)$ para o horizonte de 14 dias, fixando a origem da previsão em 01 de agosto de 2020. Como pode ser observado na Figura 6 os números obtidos nas simulações (previstos) variaram aproximadamente dentro da mesma faixa de valores que os casos registrados no período (atuais). As medidas descritivas apresentadas na Tabela 2, sugerem semelhança entre as médias dos valores previstos e a média dos casos novos registrados. As medidas dos desvios-padrão (DP) e coeficientes de variação (CV) ficaram próximas, permitindo supor homogeneidade. Pelo teste de Dunnett, confirmou-se a semelhança entre os valores previstos para os casos novos e os atuais, pois

Tabela 2. Medidas descritivas dos casos novos diários de COVID-19 e dos valores previstos nas simulações

Série	**Média**	**DP**	**CV (%)**
Casos Novos	1847,36	474,66	25,7
Simulação 1	2079,63	384,43	18,50
Simulação 2	2119,85	463,89	21,88
Simulação 3	2035,69	458,86	22,54
Simulação 4	2019,16	413,98	20,50
Simulação 5	2030,47	491,89	24,23
Simulação 6	2013,18	394,12	19,58
Simulação 7	2015,17	563,45	27,96
Simulação 8	2012,7	413,99	20,57
Simulação 9	2021,6	427,16	21,13
Simulação 10	1972,55	407,52	20,66

não foi possível rejeitar a hipótese de igualdade de médias ($prob > 0,51$ para todas as comparações).

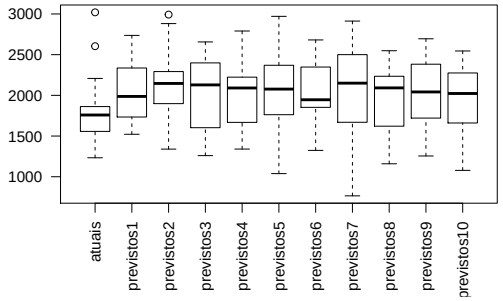

Figura 6. Comparação entre os casos novos de COVID-19 e os valores previstos nas 10 simulações.

O desempenho preditivo do modelo ARDL, foi considerado satisfatório, pois as previsões das simulações erraram, em média, de $25,2\%$ a $36,33\%$, de acordo com as medidas MAPE, indicando boa precisão. Além disso, os resultados destas previsões podem ser estimativas mais próximas do número real de pessoas diagnostigadas com a doença em cada dia, por considerar o retardo no registro destes números, decorrente do atraso no processo de testagem.

4. CONSIDERAÇÕES FINAIS

Os modelos autorregressivos de defasagens distribuídas, mostraram-se eficientes para estimar e prever o número de casos diários de COVID-19 no Estado do Paraná, Brasil, a partir os casos registrados no passado, do número de casos em investigação e da taxa de transmissão.

Os resultados obtidos neste estudo podem ser úteis aos gestores dos órgãos de controle de saúde pública para auxiliar na tomada de decisões e no estabelecimento de medidas de mitigação.

REFERÊNCIAS

Abbott, S., Hellewell, J., Thompson, R.N., Sherratt, K., Gibbs, H.P., Bosse, N.I., Munday, J.D., Meakin, S., Doughty, E.L., Chun, J.Y., et al. (2020). Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112), 112.

Abdulmajeed, K., Adeleke, M., and Popoola, L. (2020). Online forecasting of covid-19 cases in nigeria using limited data. *Data in Brief*, 105683.

Al-Mulali, U., Solarin, S.A., and Ozturk, I. (2016). Investigating the presence of the environmental kuznets curve (ekc) hypothesis in kenya: an autoregressive distributed lag (ardl) approach. *Natural Hazards*, 80(3), 1729–1747.

Al Wahaibi, A., Al Manji, A., Al Maani, A., Al Rawahi, B., Al Harthy, K., Alyaquobi, F., Al-Jardani, A., Petersen, E., and Al-Abri, S. (2020). Covid-19 epidemic monitoring after non-pharmaceutical interventions: the use of time-varying reproduction number in a country with a large migrant population. *International Journal of Infectious Diseases*.

Aslam, M. (2020). Using the kalman filter with arima for the covid-19 pandemic dataset of pakistan. *Data in Brief*, 105854.

Bastos, S.B. and Cajueiro, D.O. (2020). Modeling and forecasting the covid-19 pandemic in brazil. *arXiv preprint arXiv:2003.14288*.

Cori, A. (2020). *EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves*. URL `https://CRAN.R-project.org/package=EpiEstim`. R package version 2.2-3.

Cori, A., Ferguson, N., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178. doi:10.1093/aje/kwt133.

Dana, S., Simas, A.B., Filardi, B.A., Rodriguez, R.N., da Costa Valiengo, L.L., and Gallucci-Neto, J. (2020). Brazilian modeling of covid-19 (bram-cod): a bayesian monte carlo approach for covid-19 spread in a limited data set context. *medRxiv*.

Demirhan, H. (2020). dLagM: An R package for distributed lag models and ARDL bounds testing. *PLoS ONE*, 15(2), e0228812. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228812`.

Franco, C.M.R. and Dutra, R.F. (2020). Modelos matemáticos em epidemiologia e aplicação na evolução da covid-19 no brasil e no estado da paraíba. *Educação, Ciência e Saúde*, 7(1).

Hu, F.C. (2020). The estimated time-varying reproduction numbers during the ongoing pandemic of the coronavirus disease 2019 (covid-19) in 12 selected countries outside china. *medRxiv*.

Menendez, J. (2020). A poor-man's approach to the effective reproduction number: the covid-19 case. *medRxiv*.

MS (2020). Painel coronavírus. *Ministério da Saúde*. URL `www.https://covid.saude.gov.br/`.

Nadler, P., Wang, S., Arcucci, R., Yang, X., and Guo, Y. (2020). An epidemiological modelling approach for covid19 via data assimilation. *arXiv preprint arXiv:2004.12130*.

Najafi, F., Izadi, N., Hashemi-Nazari, S.S., Khosravi-Shadmani, F., Nikbakht, R., and Shakiba, E. (2020). Serial interval and time-varying reproduction number estimation for covid-19 in western iran. *New microbes and new infections*, 36, 100715.

Nkoro, E., Uko, A.K., et al. (2016). Autoregressive distributed lag (ardl) cointegration technique: application and interpretation. *Journal of Statistical and Econometric Methods*, 5(4), 63–91.

Palladino, A., Nardelli, V., Atzeni, L.G., Cantatore, N., Cataldo, M., Croccolo, F., Estrada, N., and Tombolini, A. (2020). Modelling the spread of covid19 in italy using a revised version of the sir model. *arXiv preprint arXiv:2005.08724*.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ribeiro, M.H.D.M., da Silva, R.G., Mariani, V.C., and dos Santos Coelho, L. (2020). Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. *Chaos, Solitons & Fractals*, 109853.

Sarkodie, S.A. and Owusu, P.A. (2020). Investigating the cases of novel coronavirus disease (covid-19) in china using dynamic statistical techniques. *Heliyon*, e03747.

Silva, R.F. (2020). Estudos por meio do modelo epidemiológico sir para os casos de infecção pelo covid-19 no paraná. *UTFPR*. URL `http://hpc.ct.utfpr.edu.br/~rsilva/simulacaoSIRcovid19.pdf`.

SSEPR (2020). Informe epidemiológico coronavírus (covid-19). *Secretaria da Saúde do Estado do Paraná*. URL `http://www.saude.pr.gov.br/Pagina/Coronavirus-COVID-19`.

Trapletti, A. and Hornik, K. (2019). *tseries: Time Series Analysis and Computational Finance*. URL `https://CRAN.R-project.org/package=tseries`. R package version 0.10-47.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6), 509–516.

Wangping, J., Ke, H., Yang, S., Wenzhe, C., Shengshu, W., Shanshan, Y., Jianwei, W., Fuyin, K., Penggang, T., Jing, L., et al. (2020). Extended sir prediction of the epidemics trend of covid-19 in italy and compared with hunan, china. *Frontiers in medicine*, 7, 169.

WHO (2020). Explore the data. *World Health Organization*. URL `https://www.who.int/`.