

Previsão da evolução dos casos de COVID-19 no Brasil e no estado do Rio de Janeiro

O LEGOS trata o problema de prever a evolução dos casos confirmados de COVID-19 a partir de duas abordagens que serão apresentadas a seguir. Para chegar aos resultados por ambas abordagens utilizamos as séries históricas de todos os países, disponibilizadas pela Universidade Johns Hopkins Whiting School of Engineering na plataforma github (c.f. <https://github.com/CSSEGISandData/COVID-19>).

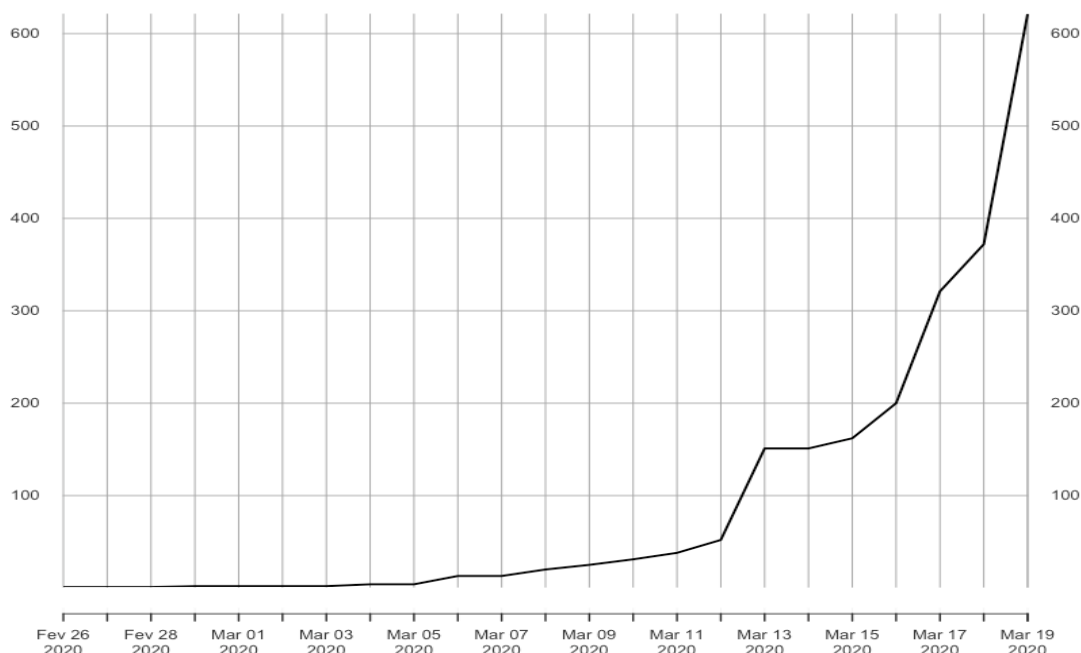
1. Abordagem 1: prever a evolução dos casos confirmados de COVID-19 no Brasil a partir do curto histórico de dados brasileiros.
2. Abordagem 2: prever a evolução dos casos confirmados de COVID-19 no Brasil a partir da evolução dos casos confirmados de todos os demais países.

O acesso direto aos dados de casos confirmados, casos de óbito e casos de recuperação podem ser encontrados online diretamente pelos links:

- [Casos confirmados](#)
- [Casos de óbito](#)
- [Casos de recuperação](#)

A série histórica de casos confirmados no Brasil são apresentadas na tabela e figura abaixo.

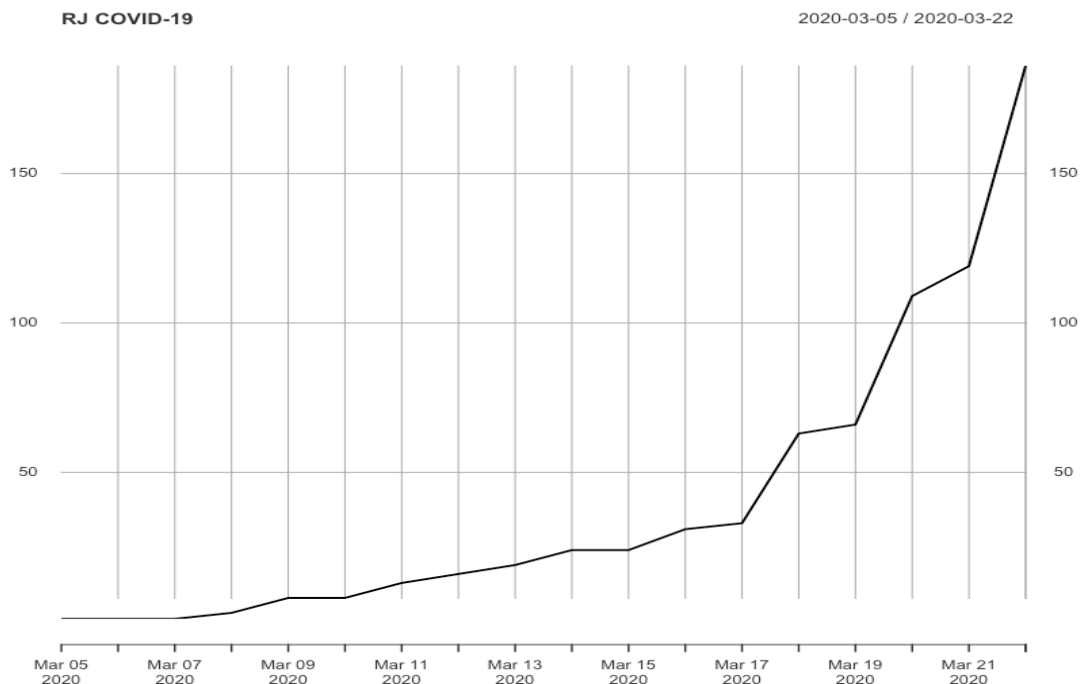
Data	Casos confirmados	Data	Casos confirmados
26/02/2020	1	10/03/2020	31
27/02/2020	1	11/03/2020	38
28/02/2020	1	12/03/2020	52
29/02/2020	2	13/03/2020	151
01/03/2020	2	14/03/2020	151
02/03/2020	2	15/03/2020	162
03/03/2020	2	16/03/2020	200
04/03/2020	4	17/03/2020	321
05/03/2020	4	18/03/2020	372
06/03/2020	13	19/03/2020	621
07/03/2020	13	20/03/2020	793
08/03/2020	20	21/03/2020	1021
09/03/2020	25	22/03/2020	1593



Observando a figura acima podemos perceber que no Brasil há um comportamento não linear e muito menos estacionário (nem estacionário com tendência). Para ratificar esta observação aplicamos teste Augmented Dickey Fuller (ADF) com hipótese nula de que a série é estacionária ou estacionária com tendência a um nível de significância de 5% (ou 95% de intervalo de confiança). Neste teste, obtivemos o valor Dickey-Fuller de 2,85 e p-valor menor que 0,01. Logo, obtivemos resultado estatisticamente significativo de que devemos rejeitar a hipótese nula de estacionaridade e assumir a hipótese alternativa que indica um comportamento explosivo. Este comportamento também foi observado quando realizamos o teste Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tanto para o nível (0,67 e p-valor 0,016) quanto para a tendência (0,21 e p-valor 0,011).

A série histórica de casos confirmados no Estado do Rio de Janeiro considerada para análise é apresentada na tabela abaixo.

Data	Casos confirmados	Data	Casos confirmados
05/03/2020	1	14/03/2020	24
06/03/2020	1	15/03/2020	24
07/03/2020	1	16/03/2020	31
08/03/2020	3	17/03/2020	33
09/03/2020	8	18/03/2020	63
10/03/2020	8	19/03/2020	66
11/03/2020	13	20/03/2020	109
12/03/2020	16	21/03/2020	119
13/03/2020	19	22/03/2020	186



Observando a figura acima podemos perceber que no Estado do Rio de Janeiro há um comportamento não linear e muito menos estacionário (nem estacionário com tendência). Para ratificar esta observação aplicamos teste Augmented Dickey Fuller (ADF) com hipótese nula de que a série é estacionária ou estacionária com tendência a um nível de significância de 5% (ou 95% de intervalo de confiança). Neste teste, obtivemos o valor Dickey-Fuller de 0,37 e p-valor menor que 0,01. Logo, obtivemos resultado estatisticamente significativo de que devemos rejeitar a hipótese nula de estacionaridade e assumir a hipótese alternativa que indica um comportamento explosivo. Este comportamento também foi observado quando realizamos o teste Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tanto para o nível (0,59 e p-valor 0,023) quanto para a tendência (0,17 e p-valor 0,026).

Abordagem 1

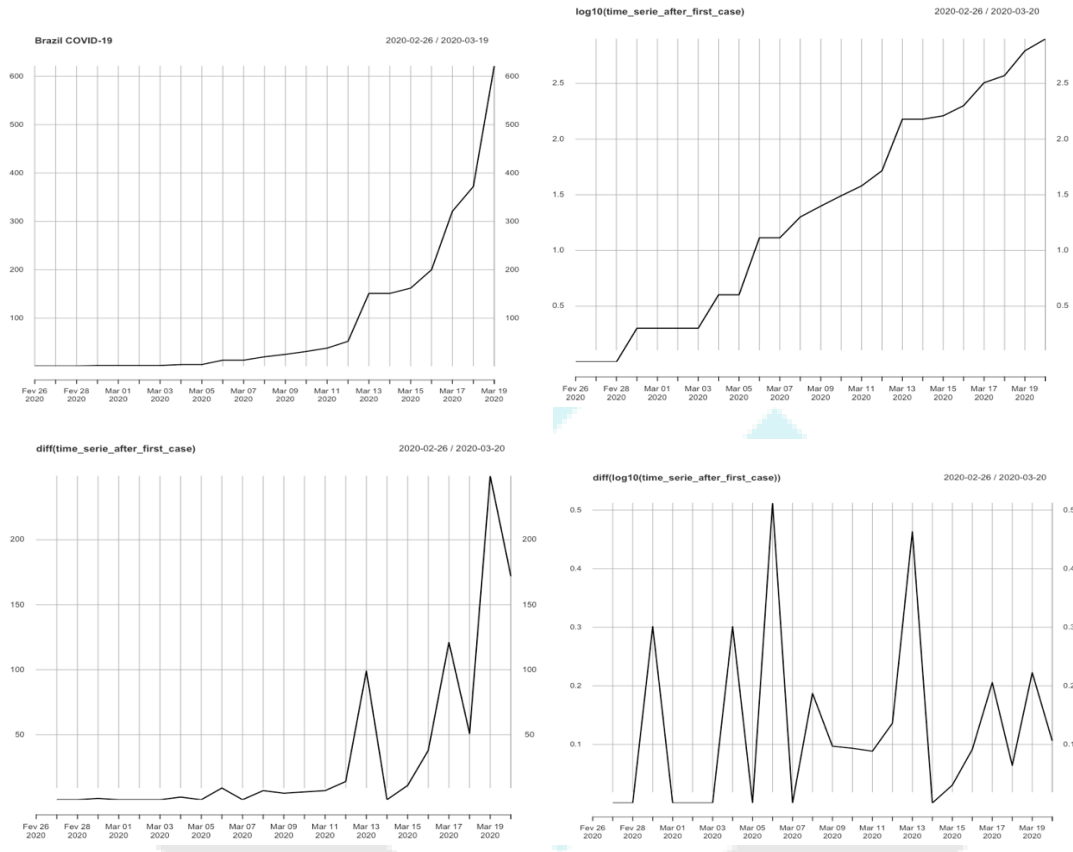
Conforme mencionado anteriormente, esta abordagem utiliza apenas a série histórica de casos confirmados no Brasil e no Rio de Janeiro para estimar a quantidade de casos esperados em no horizonte de 7 dias.

Para resolver este problema utilizamos modelos univariados consagrados da literatura e regressões apresentados abaixo:

- 1) Modelos de suavização exponencial (ETS);
- 2) Modelos auto regressivos integrados de média móvel (ARIMA);
- 3) Regressão Linear;

Nenhum dos 3 modelos apresentados acima foi concebido para tratar de séries temporais de comportamento explosivo (como é o nosso caso). Entretanto, podemos realizar algumas transformações na série original apresentada para que a série transformada atenda às condições de “contorno” das técnicas apresentadas acima. Neste

sentido, as principais formas comumente utilizadas em séries temporais são: diferenciação e transformação logarítmica (elas também podem ser combinadas). Nas figuras abaixo são apresentadas a série original e cada série transformada (logarítmica, diferenciada e logarítmica-diferenciada) de casos confirmados no Brasil.



Aplicando o mesmo teste ADF agora em cada série transformada chegamos aos seguintes resultados e conclusões:

- 1) Transformação logarítmica: Dickey-Fuller = -2,6247 e p-valor = 0,6656. Não rejeitamos a hipótese nula de estacionaridade (com ou sem tendência).
- 2) Transformação por diferenciação: Dickey-Fuller = 0,40658 e p-valor < 0,01. Rejeitamos a hipótese nula de estacionaridade (com ou sem tendência) e assumimos que é uma série de comportamento explosivo.
- 3) Transformação logarítmica diferenciada: Dickey-Fuller = -3,4024 e p-valor = 0,9226. Não rejeitamos a hipótese nula de estacionaridade (com ou sem tendência).

Para obter os resultados que aparecerão a seguir utilizamos a série original (SO) e/ou a série "logaritimizada" (SL) em cada modelo apresentado.

- 1) ETS na série original (ETS SO): não capturou o padrão de crescimento e se manteve constante ao longo do tempo;
- 2) ETS na série logaritimizada (ETS SL): capturou o padrão de crescimento explosivo;
- 3) ARIMA na série original (ARIMA SO): capturou o padrão de crescimento linear;

- 4) ARIMA na série logarimitizada (ARIMA SL): capturou o padrão de crescimento explosivo;
- 5) Regressão linear na série logarimitizada (RL SO): capturou o padrão de crescimento linear;
- 6) Regressão linear na série logarimitizada (RL SL): capturou o padrão de crescimento explosivo;

As previsões otimista, esperada e pessimista de acordo com cada modelo são apresentados nas tabelas abaixo.

CENÁRIO OTIMISTA	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	808	2104	525	1618	1641	1614
24/03/2020	345	2601	560	2184	2172	2205
25/03/2020	-138	3081	596	2947	2881	3011
26/03/2020	-687	3547	632	3979	3827	4113
27/03/2020	-1334	4000	667	5373	5088	5617
28/03/2020	-2113	4443	703	7257	6772	7670
29/03/2020	-3059	4875	738	9802	9019	10472

CENÁRIO ESPERADO	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	1592	2165	721	1997	2014	1992
24/03/2020	1592	2737	758	2714	2723	2726
25/03/2020	1592	3309	796	3689	3681	3731
26/03/2020	1592	3881	833	5014	4976	5105
27/03/2020	1592	4453	871	6815	6727	6986
28/03/2020	1592	5025	908	9262	9093	9559
29/03/2020	1592	5597	946	12589	12292	13081

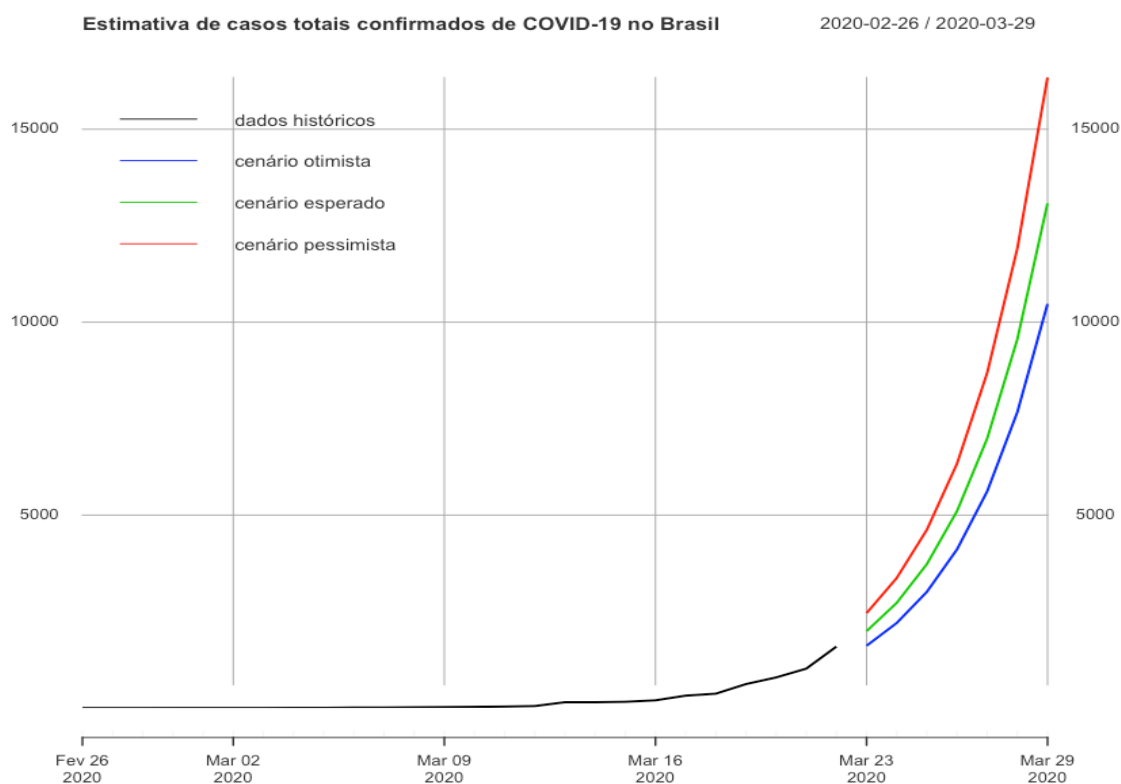
CENÁRIO PESSIMISTA	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	2376	2226	917	2464	2473	2459
24/03/2020	2838	2873	956	3373	3414	3371
25/03/2020	3321	3537	996	4616	4703	4621
26/03/2020	3871	4215	1035	6317	6471	6336
27/03/2020	4518	4906	1074	8643	8893	8688
28/03/2020	5296	5607	1114	11823	12211	11915
29/03/2020	6242	6319	1153	16170	16753	16341

Conforme comentado anteriormente, podemos descartar a priori todos os modelos que utilizaram a séries original (SO) porque não foram capazes de capturar o crescimento explosivo que devemos prever. Assim, analisaremos apenas os resultado dos modelos ETS SL, ARIMA SL e RL SL. Para selecionarmos qual modelo deveríamos usar para estimar a quantidade total de casos esperados para os próximos 7 dias utilizamos os seguintes critérios:

- 1) Teste de normalidade dos resíduos.
 - a. ETS SL: p-valor = 0,538. Não podemos rejeitar a hipótese nula de normalidade.
 - b. ARIMA SL: p-valor = 0,022. Para um nível de significância de 5% devemos rejeitar a hipótese nula de normalidade.
 - c. RL SL: p-valor = 0,426. Não podemos rejeitar a hipótese nula de normalidade.
- 2) Teste de autocorrelação dos resíduos (ACF). Apenas RL SO apresentou auto correlação dos resíduos, mas ela já foi descartada previamente.
- 3) Menor RMSE dentre os modelos aprovados pelos critérios anteriores.

MODELOS	ME	RMSE	MAE	MPE	MAPE
ETS SO	-61,32	138,78	61,32	-42,38	42,38
ARIMA SO	-8,83	41,41	20,68	-4,3	22,28
RL SO	-22	86,95	42	-22,61	36,41
ETS SL	-9,3	46,1	22,64	-3,23	22,20
ARIMA SL	0	255,5	193,42	59,14	76,84
RL SL	-7,3	34,95	17,84	-3,79	20,39

Desta forma, escolhemos o modelo de RL SL para prever os próximos valores esperados de quantidade de casos confirmados de COVID-19 no Brasil. O comportamento explosivo que a nossa previsão apresentou pode ser observado na figura abaixo.



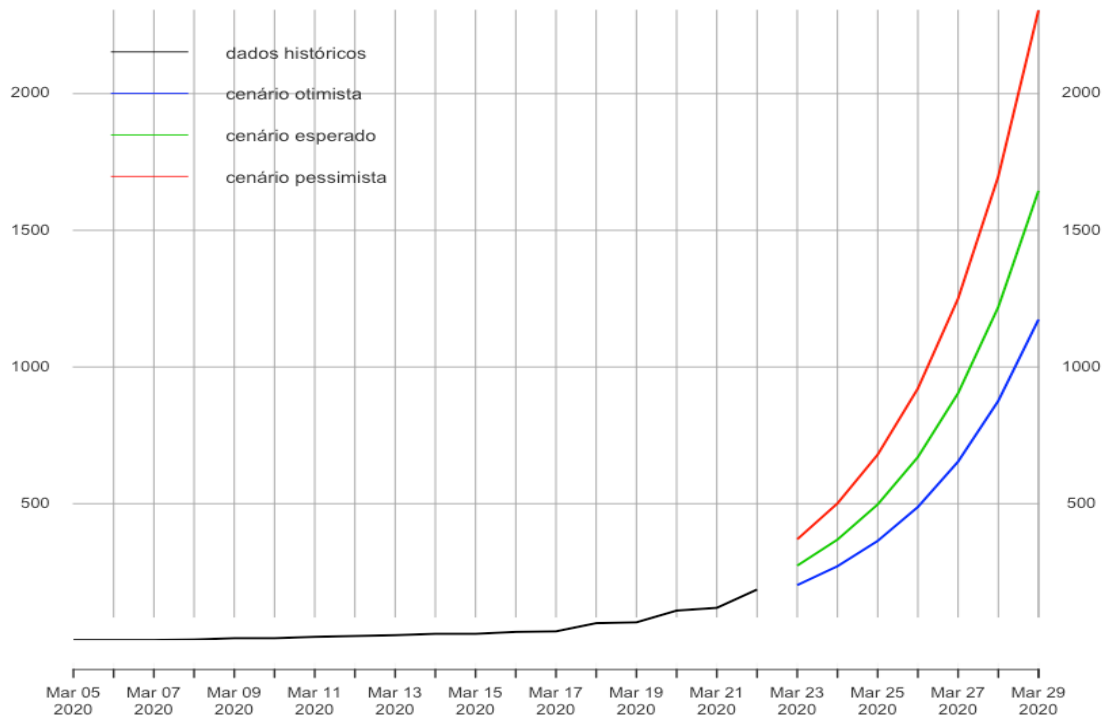
Seguindo os mesmos passos estimar a quantidade de casos esperados para o Estado do Rio de Janeiro obtivemos os resultados nos cenários otimista, esperado e pessimista para cada modelos nas tabelas abaixo. Foi escolhido o modelo que apresentou o menor erro (RL SL). O comportamento explosivo que a nossa previsão de casos totais para o Estado do Rio de Janeiro apresentou pode ser observado na figura abaixo.

CENÁRIO OTIMISTA	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	91	203	95	196	201	202
24/03/2020	34	276	103	245	248	271
25/03/2020	-27	301	111	312	314	364
26/03/2020	-97	378	118	401	401	488
27/03/2020	-181	410	126	519	516	654
28/03/2020	-284	490	133	675	668	876
29/03/2020	-410	527	141	881	869	1174

CENÁRIO ESPERADO	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	186	209	117	251	253	273
24/03/2020	186	286	125	341	344	369
25/03/2020	186	319	133	465	468	497
26/03/2020	186	405	141	634	636	671
27/03/2020	186	449	149	863	865	904
28/03/2020	186	542	157	1176	1176	1220
29/03/2020	186	595	165	1602	1600	1645

CENÁRIO PESSIMISTA	ETS SO	ARIMA SO	RL SO	ETS SL	ARIMA SL	RL SL
23/03/2020	281	215	138	320	319	370
24/03/2020	338	295	147	476	477	501
25/03/2020	399	338	155	693	698	679
26/03/2020	469	431	164	1001	1009	922
27/03/2020	553	488	172	1435	1449	1250
28/03/2020	656	593	181	2048	2071	1697
29/03/2020	782	662	189	2913	2946	2305

Estimativa do Estado do Rio de Janeiro para casos confirmados de COVID-19 2020-03-05 / 2020-03-29



Esta abordagem univariada com poucos dados não é capaz de estimar quando o comportamento explosivo parará e, por este motivo, na Abordagem 2 avaliamos o perfil de outros países, capturamos estes comportamentos e avaliamos o que aconteceria se eles acontecessem dentro da série brasileira. Este tipo de abordagem se mostra de grande valor porque captura informação de países com mais casos acumulados e que começaram a apresentar casos antes do Brasil. Assim, entendemos ter uma forma mais robusta de estimar a quantidade total de casos confirmados no Brasil quando comparamos os resultados das duas abordagens.

Abordagem 2

Foi coletado o total populacional de cada país no site: <https://www.worldometers.info/world-population/population-by-country/>. Neste estudo, foram considerados apenas os países com mais de 10 milhões de habitantes. Posteriormente, as séries de casos confirmados de covid-19 de cada país foram divididas pelo respectivo total populacional de modo a obter séries históricas com o total de casos em cada país por 100 milhões de habitantes. Desse modo, este estudo optou por trabalhar com estas séries.

Vale destacar que, foi considerado como o primeiro dia de cada série o momento em que foi atingido ou ultrapassado o total de 32,46 casos por 100 milhões de habitantes,

correspondente aos 68 casos confirmados em 11/03/2020 dividido pela população brasileira e multiplicado por um 100 milhões.

Em seguida, foram selecionados para o estudo apenas os países que obtiveram um número de casos confirmados igual ou superior ao número de casos confirmados no Brasil no dia 19/03/2020, totalizando 33 países. Os países pré-selecionados se encontram na tabela abaixo:

Austrália	China	Irã	Malásia	A. Saudita	Tunísia
Azerbaijão	R. Dominicana	Iraque	Holanda	África do S.	Turquia
Bélgica	Equador	Itália	Peru	Espanha	Reino Unido
Cambodia	França	Japão	Polônia	Sri Lanka	
Canadá	Alemanha	Jordânia	Portugal	Suécia	
Chile	Grécia	Coréia do S.	Romênia	Tailândia	

Dentre os países acima, foram selecionados os quatro cujas séries históricas eram as mais semelhantes à brasileira. Para isso, foram utilizadas duas medidas de distância em relação à série brasileira. A primeira medida de distância considerada foi a correlação de Pearson das séries diferenciadas em 1 lag com relação à série diferenciada do Brasil. Vale destacar que a série diferenciada consiste no crescimento absoluto ao longo do tempo de cada país. Portanto, a correlação de Pearson estará capturando se as séries de crescimento estão evoluindo no mesmo sentido, isto é, se o aumento no crescimento de uma série é acompanhado pelo aumento no crescimento da outra série ou não. Adicionalmente, foi utilizada uma segunda medida de distância. Esta segunda medida foi definida como a diferença entre o número de casos confirmados de cada país e o Brasil no último dia de dados disponíveis, neste caso 19/03/2020. A partir destas medidas, foram definidos os quatro países que seriam utilizados na previsão da série do Brasil. Como se trata de uma decisão multicritério, foi feito um ranqueamento dos países com relação as duas medidas de distância. Em seguida, somou-se a posição de cada país nos dois ranques. Este somatório final foi definido como a pontuação final de cada país. Por fim, foram escolhidos os quatro países que tiveram a menor pontuação. A lista de países selecionados pode ser visualizada abaixo:

Reino Unido	Canadá	Peru	Tunísia
-------------	--------	------	---------

Posteriormente, foi realizada uma decomposição stl na série de cada país restante, inclusive a brasileira, após uma transformação logarítmica. Em seguida, foram geradas 1000 séries bootstrap para cada país a partir dos resíduos gerados pela decomposição de cada série anteriormente. Assim, foram geradas ao todo 5000 séries que somadas às séries originais totalizam 5005 séries temporais.

Posteriormente, foi ajustado um modelo de regressão linear em cada série após uma transformação logarítmica da série original diferenciada. Portanto, foram ajustados 5005 modelos. A partir de cada modelo, foi feita uma previsão in-sample para 9 dias a contar de 11/03/2020. Assim, foram obtidas 5005 curvas de previsão. Escolheu-se como previsão "final" a mediana de todas as previsões e calculou-se o RMSE (46,81).

	1º Quantil	Mediana	3º Quantil
11/mar	62	70	79
12/mar	79	86	106
13/mar	101	114	141
14/mar	132	162	188
15/mar	174	228	249
16/mar	230	307	331
17/mar	308	390	439
18/mar	408	470	578
19/mar	518	619	755

Em seguida, foram realizadas previsões out-of-sample com o modelo anterior e calculou-se o RMSE (73,79). Os resultados das previsões podem ser visualizados abaixo:

	1º Quantil	Mediana	3º Quantil
20/mar	601	801	983
21/mar	662	1107	1321
22/mar	708	1473	1846

Em uma tentativa de prever o número de casos confirmados no Rio de Janeiro, utilizou-se a média das proporções históricas de casos confirmados no Rio de Janeiro em relação ao Brasil. Desse modo, foi obtida a seguinte previsão in-sample (RMSE = 15,85):

	1º Quantil	Mediana	3º Quantil
11/mar	10	11	12
12/mar	12	13	17
13/mar	16	18	22
14/mar	21	25	29
15/mar	27	36	39
16/mar	36	48	52
17/mar	48	61	69
18/mar	64	73	90
19/mar	81	97	118

Em contrapartida, foi obtida a seguinte previsão out-of-sample (RMSE = 24,01):

	1º Quantil	Mediana	3º Quantil
20/mar	94	125	154
21/mar	103	173	206
22/mar	110	230	288

No dia 23/03/2020, os dados foram atualizados com a quantidade de casos confirmados e todo o procedimento referente à calibragem dos modelos foi refeito para, em seguida,

ser realizada uma previsão para os cinco dias seguintes. O RMSE do ajuste para a série Brasil passou a ser de 36,02. As previsões podem ser visualizadas abaixo.

	1º Quantil	Mediana	3º Quantil
23/mar	1191	1963	2179
24/mar	1863	2518	3011
25/mar	2910	3388	4107
26/mar	3782	4608	6517
27/mar	4416	5928	9794

Em contrapartida, o RMSE do ajuste para a série Rio passou a ser 15,38. As previsões podem ser visualizadas abaixo.

	1º Quantil	Mediana	3º Quantil
23/mar	181	299	331
24/mar	283	383	458
25/mar	443	515	625
26/mar	575	701	991
27/mar	672	902	1490

Conclusão

Seguindo duas abordagens diferentes foi possível estimar a quantidade de total de casos previstos para o Brasil e para o Estado do Rio de Janeiro. Nota-se que, mesmo utilizando abordagens diferentes, os valores estimados para a quantidade de casos do Rio de Janeiro convergiram para valores próximos.

Autores:

Daniel Assad

Currículo Lattes: <http://lattes.cnpq.br/0258423859812498>

Rodrigo Alvim

Currículo Lattes: <http://lattes.cnpq.br/8885133451473532>

Nota técnica divulgada em 23 de março de 2020.