Estimation of New COVID-19 Cases in Brazil Using Google Search Data

by Lilian C. K. Biasi

• BACKGROUND
  • People’s behavior are affected in a pandemic, and can be used to predict disease progression

• OBJECTIVE
  • To correlate Google searches for some keywords with the number of daily confirmed new cases of COVID-19 in Brazil

• METHODS
  • Data from: Google Trends and Brazilian govt. websites
  • Studied period: Jan. 1st to Oct. 4th, 2020
  • Model adopted: Random Forest Regression
  • 90% of the observation points for training and 10% for validation
DATA COLLECTION - TRAINING

Interest over time of some Keywords in Google

Number of New Covid-19 Cases

Train

Trained Model
Random Forest Regression

Searched words

<table>
<thead>
<tr>
<th>In Portuguese</th>
<th>Translation to English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covid</td>
<td>COVID-19</td>
</tr>
<tr>
<td>Coronavirus</td>
<td>Corona Virus</td>
</tr>
<tr>
<td>Corona</td>
<td>Corona</td>
</tr>
<tr>
<td>Álcool</td>
<td>Ethanol</td>
</tr>
<tr>
<td>Máscara</td>
<td>Mask</td>
</tr>
<tr>
<td>Febre</td>
<td>Fever</td>
</tr>
<tr>
<td>Desemprego</td>
<td>Unemployment</td>
</tr>
<tr>
<td>Suicídio</td>
<td>Suicide</td>
</tr>
</tbody>
</table>

Obtained from Google Trends – "2020-01-01 to 2020-10-04"


Number of New Covid-19 Cases

Validation/Use

Train

Random Forest Regression

New cases per day (14-day moving average)

New cases per day
RESULTS

Metrics

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.944834</td>
</tr>
<tr>
<td>Validation</td>
<td>0.954537</td>
</tr>
<tr>
<td>Overall</td>
<td>0.94591</td>
</tr>
</tbody>
</table>

Dendrogram

Variable importance

Suicide: 0.20
Unemployment: 0.15
Corona: 0.25
Coronavirus: 0.10
Mask: 0.05
Covid: 0.00
Ethanol: 0.30

R² = 0.94591

Predicted vs. Real

Pred $= 0.94591$ Real
CONCLUSIONS

The model can be used to estimate the number of new COVID-19 cases

Acknowledgments: lckbiasi@unicamp.br