Estimating Excess Deaths due to Covid-19 in Brazil using the *Cartórios* Data

Thomas Fujiwara

Princeton University

Updated: Aug 20, 2020
First Version: June 4, 2020

Motivation and background

- ▶ Possible under-reporting of deaths due to covid-19 in Brazil lead to focus on **excess deaths relative to historical averages**.

- ▶ A lot of the analysis here is superseded by more recent analysis at:
  https://www.conass.org.br/indicadores-de-obitos-por-causas-naturais/
    - ▶ Estimates in link make additional data corrections not discussed here.
    - ▶ See the technical notes in link for a more thorough discussion.

- ▶ Goal of this not is to help others do their own analysis with the *Registro Civil* or "*cartórios* data."
    - ▶ Provided by Arpen's *Portal da Transparência do Registro Civil*.
    - ▶ There are some pitfalls an issues in using this data that I highlight here.

- ▶ I provide some graphs and excess death estimates for Brazilian municipalities with over a million inhabitants.
    - ▶ See link above for analysis by state, gender, and age. Note that the analysis there makes data adjustments that I do not do here.

- ▶ All other references and links are in the last slides.

Brazilian data on mortality I

- The "standard" or "official" source for mortality in Brazil is the Datasus' *Sistema de Informação da Mortalidade*, based on vital statistics (universe of death certificates).

- **Non-preliminary Datasus' data for 2020 is not yet available.**
  - System is not designed to provide info in "real time."
  - Reliable data only up to Dec 2019 available (2019 data still preliminary).

- Recently, the Arpen provided the "cartórios data."
  - Also based on death certificates, collected from public notaries (cartórios).
  - Arpen provides deaths for 2019-2020, **but comparison with Datasus is not straightforward - as discussed later**.

- Links to data sources are in the last slides.

Brazilian data on mortality II

- ▶ The delays in the Datasus system exist for a reason: it takes time to collect and harmonize thousands of death certificates.

- ▶ We can think of Arpen as a preliminary and very likely incomplete version of the Datasus data.
  - ▶ Dealing with the related data issues is the point of these slides.

- ▶ Note that absent the Arpen data, our only option to study the effects of covid-19 on *total* mortality is waiting until the 2020 Datasus mortality numbers will be released, which can take months or over a year.

Four important things to know about the Arpen data

1. **Data takes a long time to enter Arpen's system. Data from the last 15-20 days is very incomplete and unreliable, but sometimes deaths takes months to enter the system.**
   - e.g., if you collect the data on Aug 30, vast majority of deaths from Aug 1 onwards will not have entered the system yet.
   - It can takes weeks for most deaths to enter Arpen's system.
   - **My advice is to not analyze data from the last 20 days at least (it's safer to use data from 1-2 months earlier only)** - if you do not follow this, you will always "find" that mortality is "falling" in recent days.

2. The Arpen data is continuously updated, so even death counts from 2019 or Jan-Feb 2020 may change in, say, Jul 2020.
   - **Important to get latest version of data frequently.**

Four important things to know about the ARPEN data

3. Arpen only includes deaths by "natural" or "non-external" causes.
   - to make it comparable, external deaths must be removed from Datasus.
   - e.g., homicides and accidents are not included.

4. Total Brazil-wide deaths in the Arpen data are "too few" compared to the Datasus.
   - Arpen shows an average of 3164 deaths per day in 2019.
   - Datasus' shows an average of 3270 deaths per day in 2020.
   - **Likely due to "missing reports," especially in rural areas - data quality likely higher in urban centers (more on this later).**

Making the DATASUS and ARPEN data comparable

- ▶ Given "too few deaths" in Arpen and issues in rural areas, preferable to analyze individual municipalities and not the whole country or entire states.
  - ▶ Idea is to check if Arpen data looks "reasonable" in each municipality that is analyzed.

- ▶ **To make total deaths by municipality comparable:**
  - ▶ Exclude external deaths from Datasus (by removing all deaths that have the *Instituto Médico Legal - IML* as the "*atestante*").
  - ▶ Also exclude fetal deaths from Datasus (they are not included in Arpen).
  - ▶ Datasus data contains both "municipality of residence of the deceased" and "municipality where death occurred." Arpen only provides data by municipality of occurrence (but see discussion regarding São Paulo later).
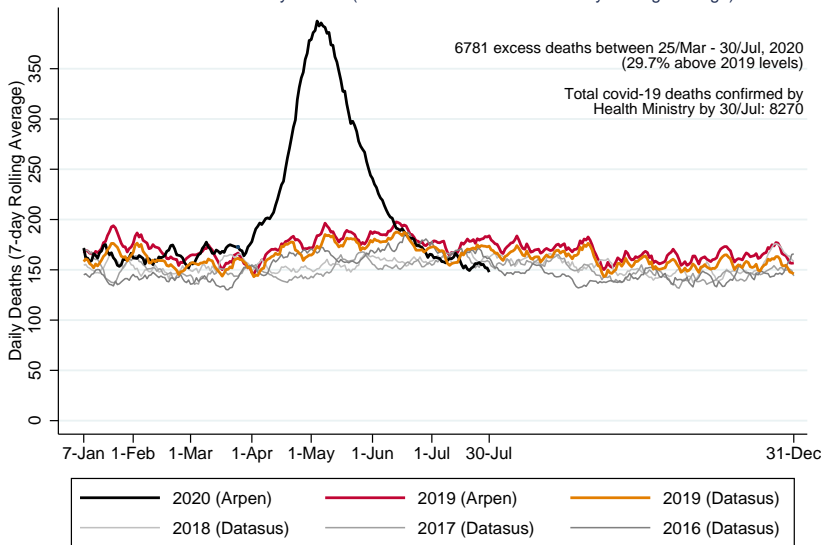
Application to 17 Brazilian municipalities

- ▶ I analyze here the data from 17 municipalities with more than a million inhabitants.
    - ▶ Focus on total deaths.
    - ▶ No breakdown by gender, age, different causes of death (I have not yet explored how to make Arpen and Datasus comparable for this - hopefully others can follow up from here and do that).

- ▶ Arpen data was scraped on Aug 19, so I only use data up to Jul 30, as discussed previously.

- ▶ I calculate excess deaths by difference between 2020 and 2019.
    - ▶ **So always comparing Arpen data (Datasus is only used to validate it visually).**
    - ▶ I pick different periods by municipality as indicated by visual inspection of the graphs.
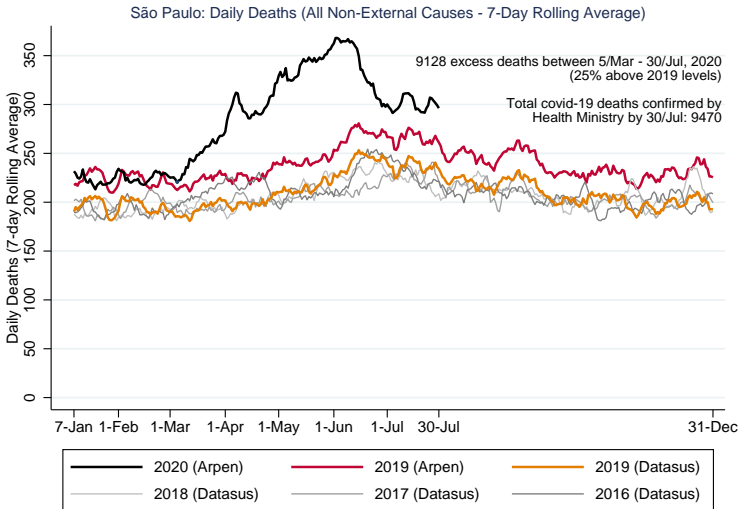
Visualizing the data

- ▶ I take the 7-day moving average (it smoothes noisy data but does not "throw variation away" like aggregating to weekly data).

- ▶ The following graphs plot the 2016-2019 Datasus data against the 2019-2020 Arpen data.
  - ▶ Besides the 7-day aggregation, this is plotting the raw data.
  - ▶ Checking if the 2019 Arpen overlaps with 2019 Datasus (as well as 2016-18) data allows to check quality of Arpen's data relative to Datasus.
  - ▶ Same for 2020 data in Jan and Feb (before covid-19 was present).
  - ▶ **In some cases, rise in excess deaths strikingly visible: deaths much higher above any other date in last 4 years.**

- ▶ I make some comments for each municipality below the graph.

Rio de Janeiro: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

6781 excess deaths between 25/Mar - 30/Jul, 2020
(29.7% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 8270

Daily Deaths (7-day Rolling Average)

| 2020 (Arpen) | 2019 (Arpen) | 2019 (Datasus) |
| 2018 (Datasus) | 2017 (Datasus) | 2016 (Datasus) |

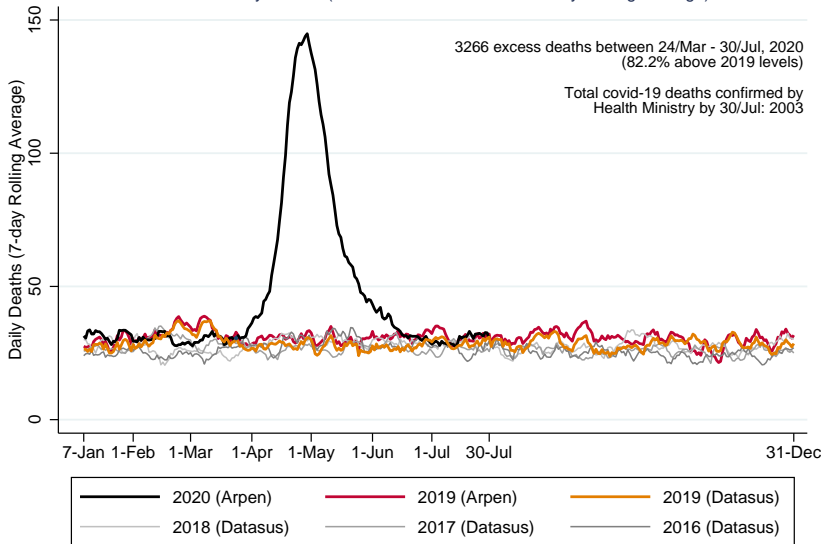The 2020 deaths in late July is likely an issue with reporting delays and not that deaths are actually below historical levels.

São Paulo: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

9128 excess deaths between 5/Mar - 30/Jul, 2020
(25% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 9470

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)
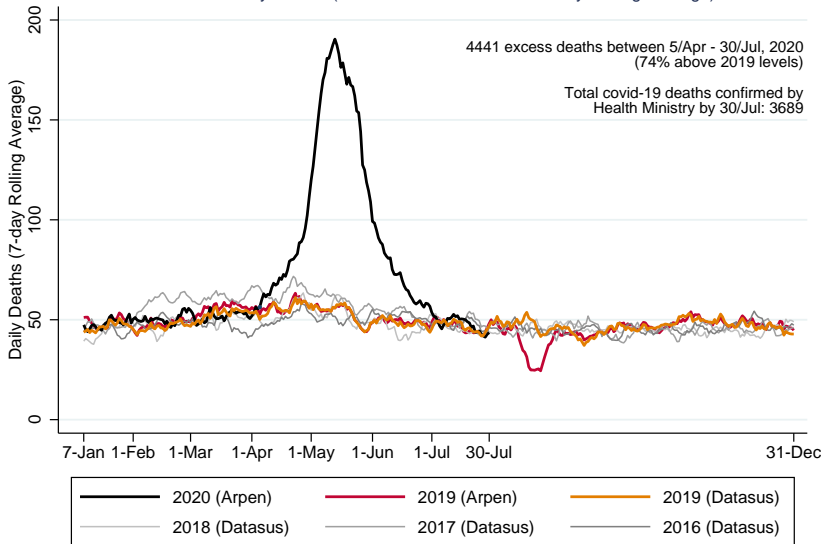
This is the only case where Arpen shows more deaths than Datatsus for 2019. I am not sure what explains this. My best guess is a distinction between how "municipality of occurrence" was coded in each dataset. Note that deaths start rising in 5/Mar (before first confirmed covid-19 death in 17/Mar).

Manaus: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

3266 excess deaths between 24/Mar - 30/Jul, 2020
(82.2% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 2003

Legend:
- 2020 (Arpen)
- 2019 (Arpen)
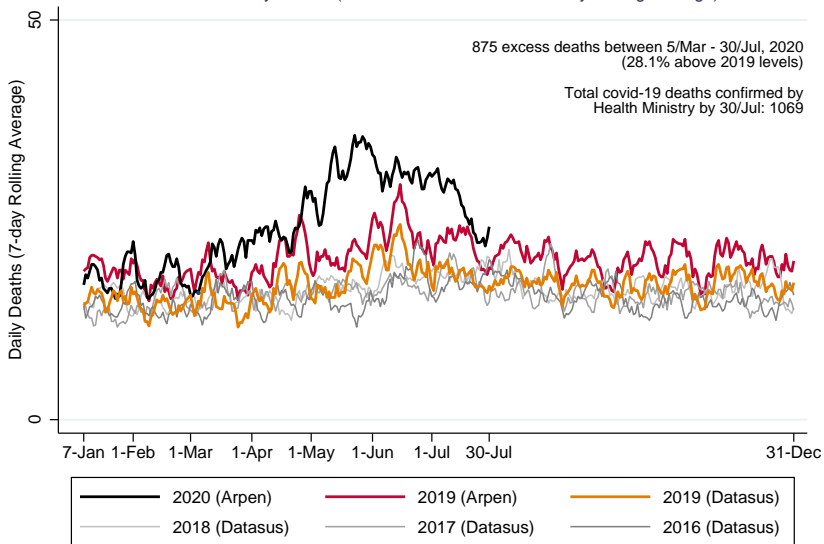- 2019 (Datasus)
- 2018 (Datasus)
- 2017 (Datasus)
- 2016 (Datasus)

Note that y-scale axis changes from graph to graph.

Fortaleza: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

4441 excess deaths between 5/Apr - 30/Jul, 2020
(74% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 3689

Legend:
- 2020 (Arpen)
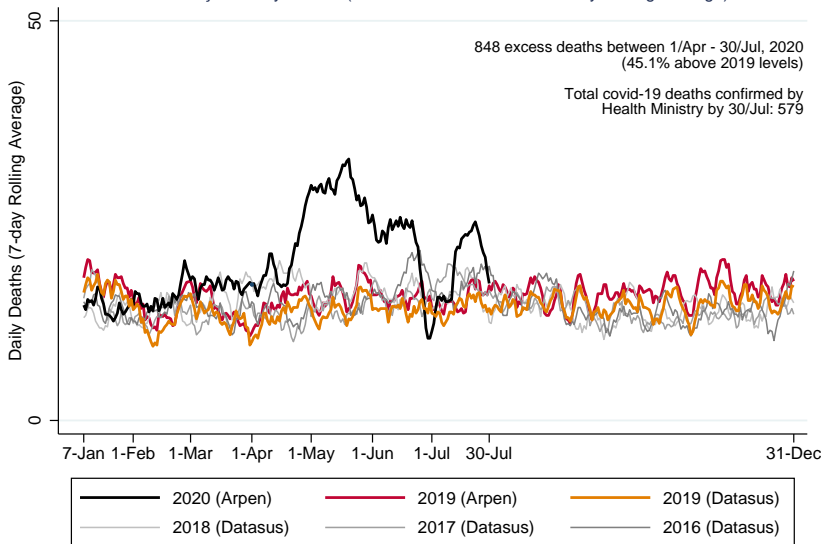- 2019 (Arpen)
- 2019 (Datasus)
- 2018 (Datasus)
- 2017 (Datasus)
- 2016 (Datasus)

Note that y-scale axis changes from graph to graph.

Guarulhos: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

875 excess deaths between 5/Mar - 30/Jul, 2020
(28.1% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 1069

Legend:
- 2020 (Arpen)
- 2019 (Arpen)
- 2019 (Datasus)
- 2018 (Datasus)
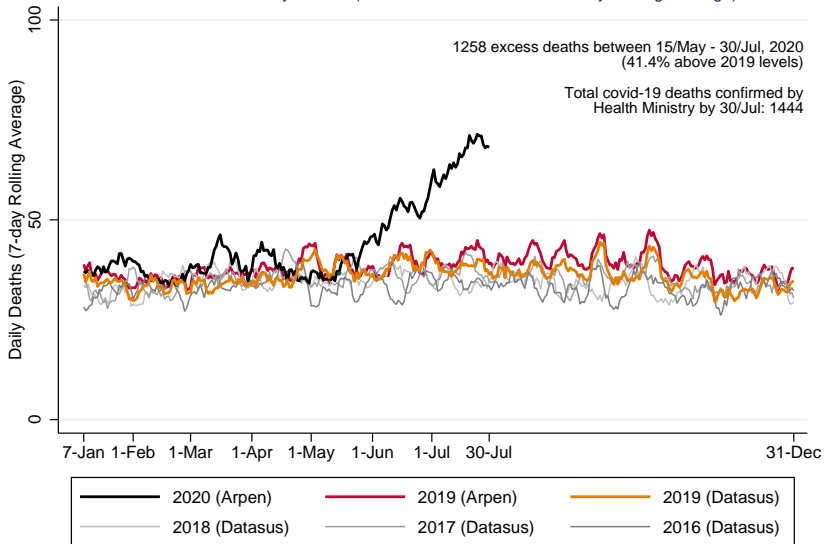- 2017 (Datasus)
- 2016 (Datasus)

Note that data is noisier in a relatively smaller municipality (also y-axis smaller).

São Gonçalo: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

848 excess deaths between 1/Apr - 30/Jul, 2020
(45.1% above 2019 levels)

Total covid-19 deaths confirmed by
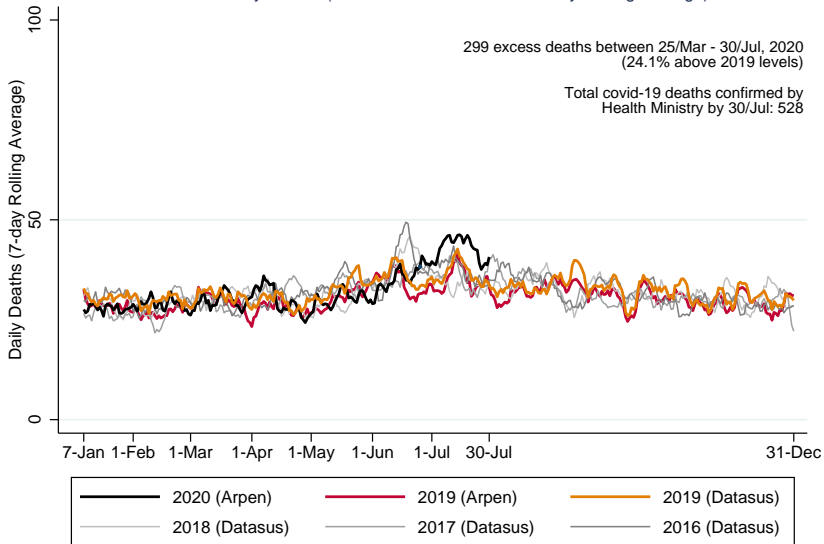Health Ministry by 30/Jul: 579

Note that data is noisier in a relatively smaller municipality (also y-axis smaller). That "dip" in early July could possibly be a reporting issue (which may be corrected in the future).

Distrito Federal: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

1258 excess deaths between 15/May - 30/Jul, 2020
(41.4% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 1444

Legend:
- 2020 (Arpen)
- 2019 (Arpen)
- 2019 (Datasus)
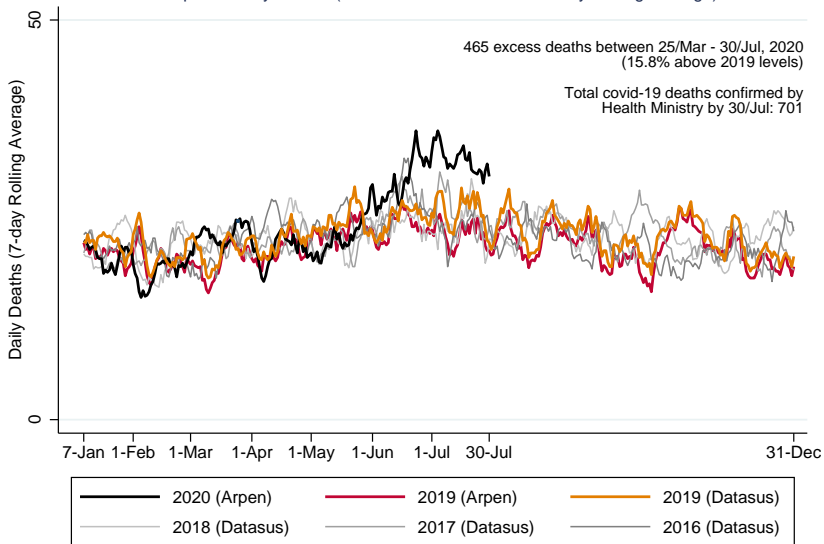- 2018 (Datasus)
- 2017 (Datasus)
- 2016 (Datasus)

Note that excess deaths start occurring relatively later (late May to early June).

Curitiba: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

299 excess deaths between 25/Mar - 30/Jul, 2020
(24.1% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 528

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)

Note that excess deaths start occurring relatively later (mid-June)

Campinas: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

465 excess deaths between 25/Mar - 30/Jul, 2020
(15.8% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 701

Daily Deaths (7-day Rolling Average)

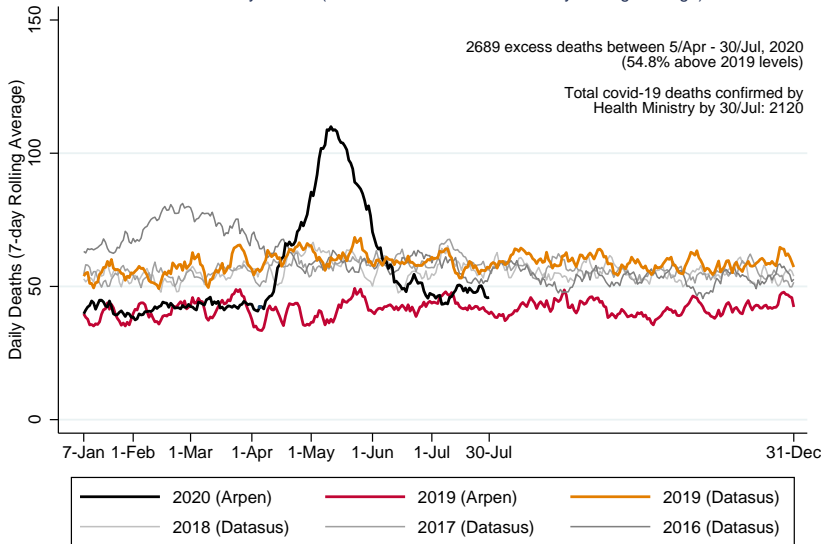| 2020 (Arpen) | 2019 (Arpen) | 2019 (Datasus) |
| 2018 (Datasus) | 2017 (Datasus) | 2016 (Datasus) |

This is an interesting case: Campinas is only 100km away from São Paulo.

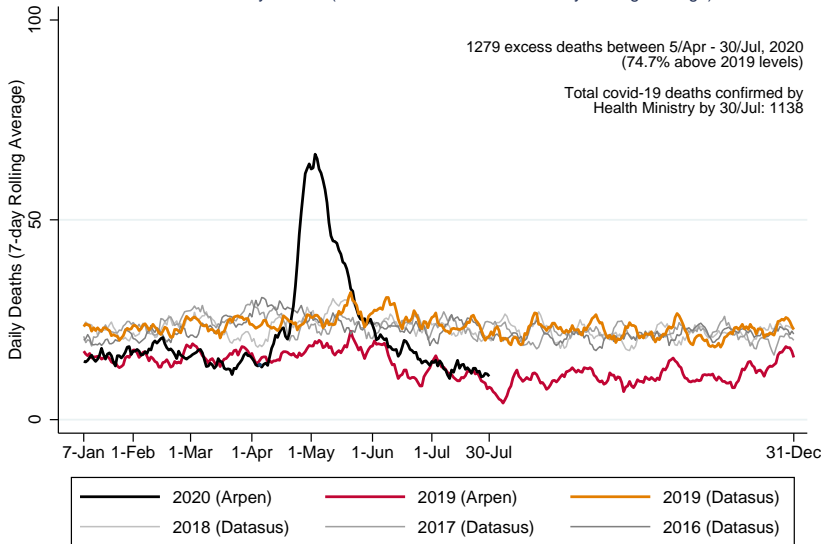# Cases where Arpen does not match Datasus data is less clear

- ► In the remaining municipalities, there is the issue that the Arpen data seems to differ from the Datasus levels.

- ► While this raises some concerns, there is still useful information in these comparisons.

- ► There are many cases that suggest substantial excess deaths and cases (Belo Horizonte, in particular) where the data issues make it difficult to draw any conclusions.

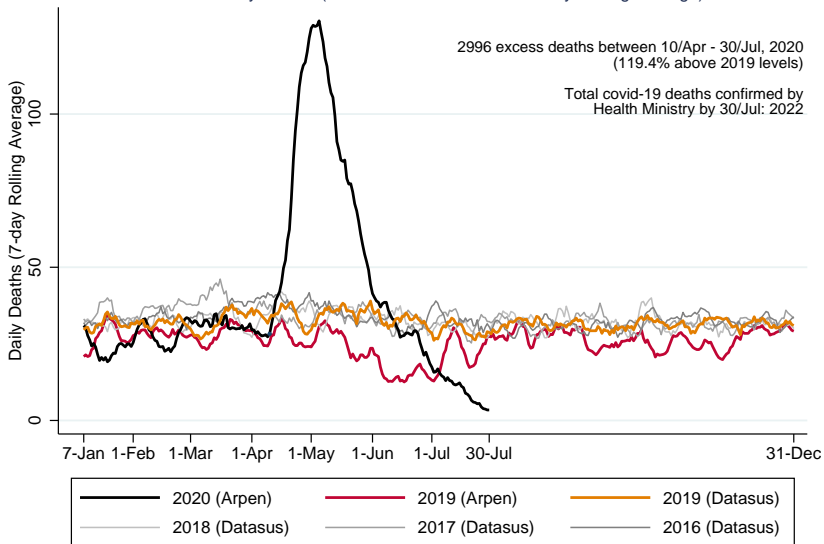Recife: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

2689 excess deaths between 5/Apr - 30/Jul, 2020
(54.8% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 2120

Daily Deaths (7-day Rolling Average)

| 2020 (Arpen) | 2019 (Arpen) | 2019 (Datasus) |
| 2018 (Datasus) | 2017 (Datasus) | 2016 (Datasus) |

Data from Arpen tends to be below Datasus' levels (apart from spike during the pandemic)

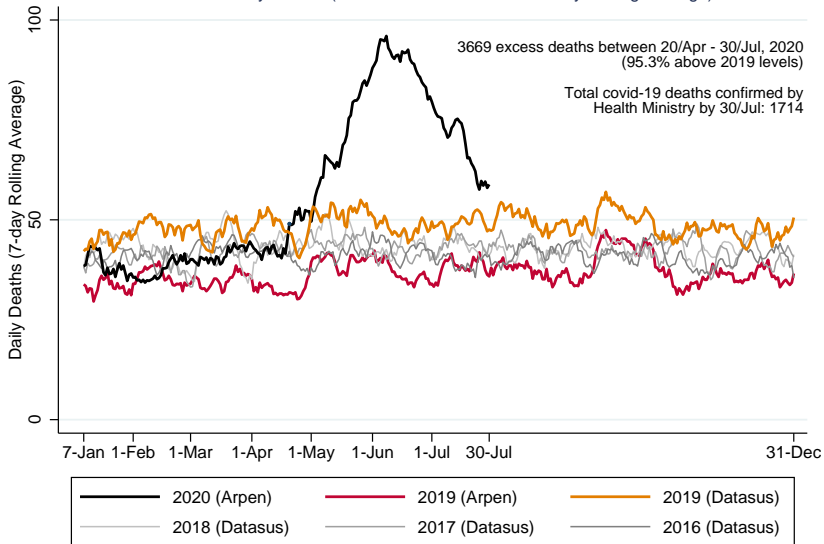São Luís: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

1279 excess deaths between 5/Apr - 30/Jul, 2020
(74.7% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 1138

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)

Data from Arpen tends to be below Datasus' levels (apart from spike during the pandemic)

Belém: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

2996 excess deaths between 10/Apr - 30/Jul, 2020
(119.4% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 2022

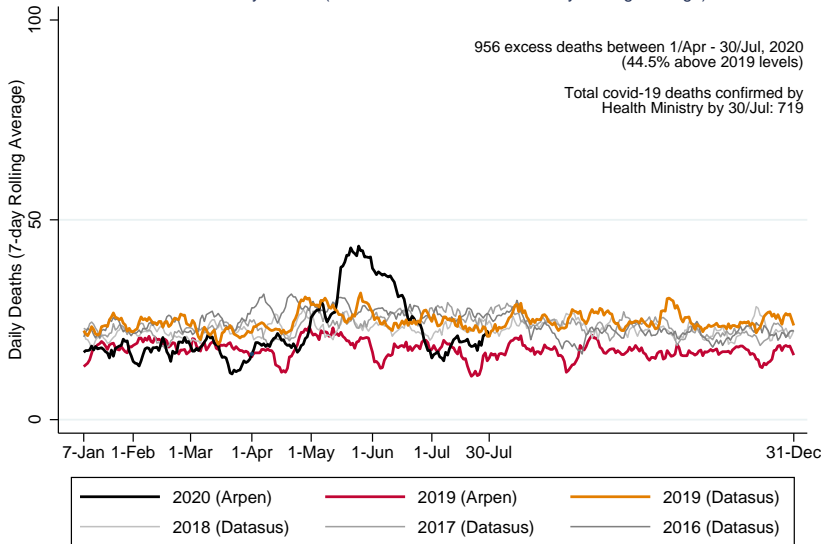Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)

Data from Arpen tends to be below Datasus' levels (apart from spike during the pandemic).
The sharp decline in July is likely a reporting issue too.

Salvador: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

3669 excess deaths between 20/Apr - 30/Jul, 2020
(95.3% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 1714

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)

Arpen tends to be below Datasus' levels (apart from spike during the pandemic)

Maceió: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

956 excess deaths between 1/Apr - 30/Jul, 2020
(44.5% above 2019 levels)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 719

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)

Arpen tends to be below Datasus' levels (apart from spike during the pandemic)

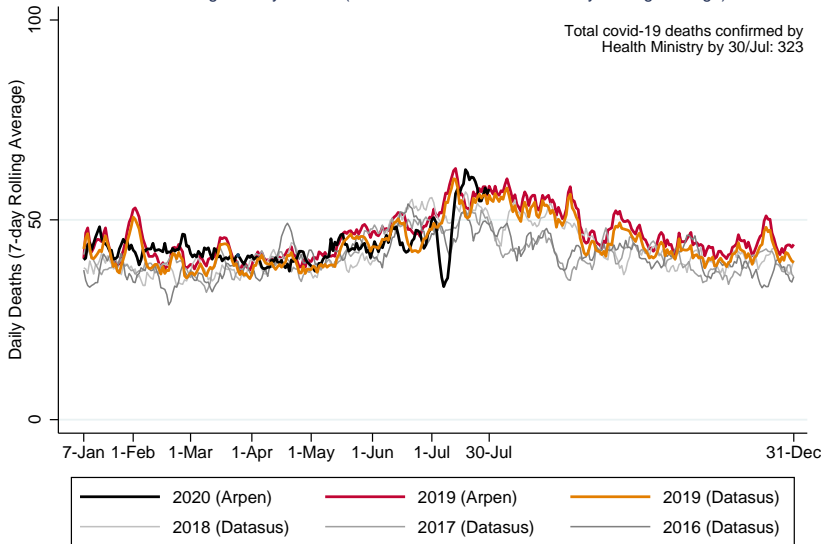Goiânia: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 441

Legend: 2020 (Arpen), 2019 (Arpen), 2019 (Datasus), 2018 (Datasus), 2017 (Datasus), 2016 (Datasus)
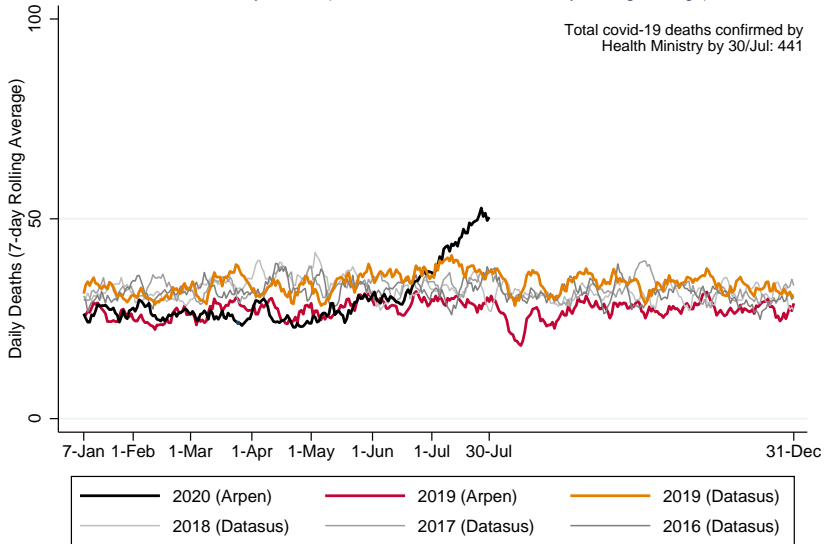
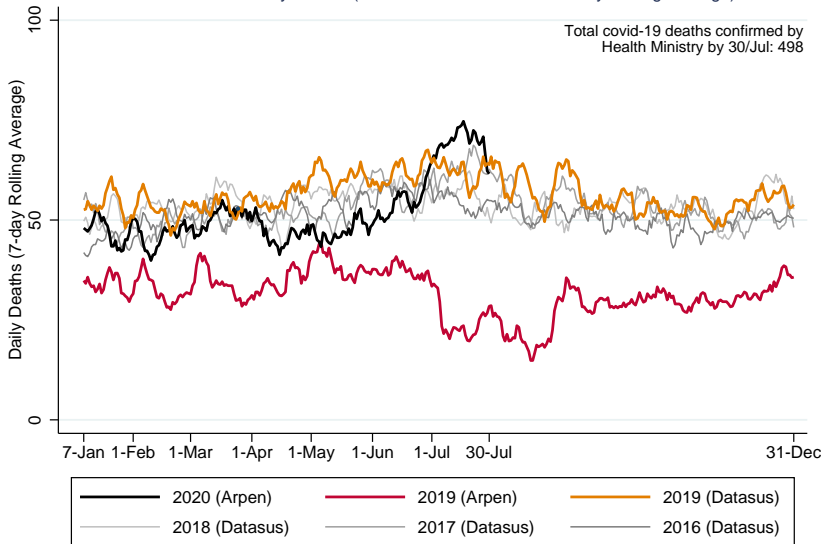This is closer to a "no excess deaths" case, but Arpen levels are slight below Datasus.

Porto Alegre: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 323

Legend: 2020 (Arpen) | 2019 (Arpen) | 2019 (Datasus) | 2018 (Datasus) | 2017 (Datasus) | 2016 (Datasus)

This "dip" in deaths around early July 2020 is most likely a reporting issue: data not being uploaded into Arpen's system.

Belo Horizonte: Daily Deaths (All Non-External Causes - 7-Day Rolling Average)

Total covid-19 deaths confirmed by
Health Ministry by 30/Jul: 498

Daily Deaths (7-day Rolling Average)

Legend:
- 2020 (Arpen)
- 2019 (Arpen)
- 2019 (Datasus)
- 2018 (Datasus)
- 2017 (Datasus)
- 2016 (Datasus)

Data from ARPEN tends to be below DATASUS' levels (and more so in Jul-Dec 2019)

Data sources

- Arpen data:
  - Available (click and point) at:
    transparencia.registrocivil.org.br/especial-covid
  - Scraped municipality-level data (used here) at:
    github.com/capyvara/brazil-civil-registry-data
  - State-level data also at: brasil.io/dataset/covid19/obito_cartorio/
  - All of this involves a great public good provided by everyone involved.
    - Thanks to Arpen to providing the data during a pandemic, and to Marcelo Oliveira, Alvaro Justen and the Brasil.io team for making it easily accessible

- Datasus data: datasus.gov.br/DATASUS/
- Also individual files at: ftp://ftp.datasus.gov.br/ .

- All the code and date used here is at www.princeton.edu/~fujiwara/

References I

- A lot of the analysis here is superseded by more recent analysis at:
  https://www.conass.org.br/indicadores-de-obitos-por-causas-naturais/
  - Estimates in link make additional data corrections not discussed here.
  - See the technical notes in link for a more thorough discussion.

- Similar analysis (without using Datasus) by state: mortalidade.com.br
  - Check data state-by-state for sharp drops in recent dates.
  - Such sharp drops in are most likely reporting problems.
  - They contribute to artificially making deaths fall in recent days when we look at Brazil as a whole.
  - So a lot of caution needed to interpret this as "coming down from peak."

- Another similar analysis: https:
  //demografiaufrn.net/2020/05/11/covid-excesso-de-mortalidade/
  - There is a lot of detail here and also a careful discussion on how to interpret excess mortality.

- Links are great resources on conceptual/interpretation issues regarding excess death calculations, as well as applications to other countries.

References II

- ▶ Coverage of results here and much more on under-notification in Brazil:
  https://g1.globo.com/mundo/blog/helio-gurovitz/post/2020/06/04/
  nao-acredite-nos-numeros-oficiais.ghtml
  - ▶ Linked post covers results using data up to May 17, a lot has changed.

- ▶ An earlier (similar) analysis for 5 municipalities at G1:
  g1.globo.com/mundo/blog/helio-gurovitz/post/2020/05/04/
  mortes-subiram-30percent-em-5-cidades.ghtml
  - ▶ By H. Gurovitz, A. Justen, P. Lotufo, and M. Oliveira.

- ▶ Similar analysis for Manaus and Fortaleza at the Financial Times:
  ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441
  - ▶ Not using Datasus data though.
  - ▶ FT just added a graph for Brazil as a whole - however, as I mentioned above a lot of deaths are missing when you analyze at this leve (see next slide).

- ▶ Similar analysis (building on Gurovitz et al) at the NY Times: nytimes.
  com/interactive/2020/04/21/world/coronavirus-missing-deaths.html