Robust Modeling of Sparse and Overdispersed COVID-19 Data: A Case Study on the Impact of Air Traffic

Mintodê Nicodème Atchadé[1,2]
(joint work with Yves Morel Sokadjo)

[1]Benin National University of Sciences, Technologies, Engineering, and Mathematics
[2]Carl von Ossietzky Universität Oldenburg (Visiting)

Symposium on *Recent Advances in Meta-Analysis*
Dortmund, Germany – June 30th–July 1st, 2025

Overview

1. Introduction
2. Data and methods
3. Results and discussion
4. Meta-analysis implications and conclusion

Introduction

- COVID-19, first detected in Wuhan, China, in December 2019, has caused a global pandemic [Ayittey et al., 2020].
- Person-to-person transmission is a key factor; control measures include masks, social distancing, quarantine, and travel restrictions [Chan et al., 2020].
- Early COVID-19 cases worldwide were imported via international travel, particularly air traffic [Wu et al., 2020; Bogoch et al., 2020b].

Introduction

- COVID-19, first detected in Wuhan, China, in December 2019, has caused a global pandemic [Ayittey et al., 2020].
- Person-to-person transmission is a key factor; control measures include masks, social distancing, quarantine, and travel restrictions [Chan et al., 2020].
- Early COVID-19 cases worldwide were imported via international travel, particularly air traffic [Wu et al., 2020; Bogoch et al., 2020b].
- Several studies showed that air transport plays a major role in spreading infectious diseases such as influenza, Ebola, Zika, and Dengue [Tatem et al., 2006; Brownstein et al., 2006; Bogoch et al., 2015].
- This study investigates the hypothesis that passenger air traffic significantly influences COVID-19 spread worldwide using robust modeling and cross-validation techniques.

Data

- COVID-19 case data were obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [CSSE, 2020].
- Passenger air traffic data (number of passengers carried) were sourced from the World Bank database [World Bank, 2018].
- The COVID-19 data cover the period from January 23, 2020, to July 13, 2020, and include multiple countries worldwide.

Data

- COVID-19 case data were obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [CSSE, 2020].
- Passenger air traffic data (number of passengers carried) were sourced from the World Bank database [World Bank, 2018].
- The COVID-19 data cover the period from January 23, 2020, to July 13, 2020, and include multiple countries worldwide.
- Variables:
  - **GC**: daily count of total confirmed COVID-19 cases per country.
  - **PAT**: total passenger air traffic in 2018 per country (most recent available data).
- Data exhibit sparse and overdispersed count characteristics typical of COVID-19 case data.
- Preprocessing involved aligning dates and aggregating counts to ensure consistency.

Modeling approach: count models overview

Context

To model global COVID-19 case counts (GC), we consider count data models that can account for overdispersion and excess zeros:

- Poisson Model (PM)
- Quasi-Poisson Model (QPM)
- Negative Binomial Model (NBM)
- Zero-Inflated Models (ZIM)

Note

Hurdle models are not considered as they assume that all zeros arise from a single source - an assumption not valid for our data.

Modeling approach: Poisson and Quasi-Poisson regressions

Let i index the countries. We define:

- y_i: Daily confirmed COVID-19 cases (GC)
- x_i: Passenger air traffic (PAT) in 2018
- $\mu_i = \mathbb{E}[y_i]$: Expected number of cases

We model:

$$\log(\mu_i) = \alpha + \beta x_i \tag{1}$$

Variance assumptions:

$$\begin{aligned} \text{PM:} \quad & \mathrm{Var}(y_i) = \mu_i \\ \text{QPM:} \quad & \mathrm{Var}(y_i) = \theta \mu_i, \quad \theta > 1 \end{aligned}$$

Modeling approach: Negative Binomial regression

Used when data exhibit overdispersion relative to Poisson. It assumes a Poisson-Gamma mixture:

- $y_i \sim \text{NB}(\mu_i, \rho)$, where $\mu_i = \mathbb{E}[y_i]$ as before
- ρ: dispersion parameter

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + \rho)}{\Gamma(y_i + 1)\Gamma(\rho)} \left(\frac{\rho}{\rho + \mu_i}\right)^{\rho} \left(\frac{\mu_i}{\rho + \mu_i}\right)^{y_i} \tag{2}$$

Variance:

$$\text{Var}(y_i) = \mu_i + \frac{\mu_i^2}{\rho}$$

⋆ QPM captures overdispersion by scaling variance linearly: $\text{Var}(y_i) = \theta\mu_i$.

⋆ NBM captures overdispersion with quadratic variance: $\text{Var}(y_i) = \mu_i + \mu_i^2/\rho$.

Modeling approach: Zero-Inflated regressions (ZIM)

ZIM assumes two latent processes:

- A binary process generating excess zeros (modeled via logistic regression)
- A count process generating actual cases (modeled via PM or NBM)

$$\Pr(y_i \mid x_i) = \begin{cases} \pi_i + (1 - \pi_i) g(y_i \mid x_i) & \text{if } y_i = 0 \\ (1 - \pi_i) g(y_i \mid x_i) & \text{if } y_i > 0 \end{cases} \tag{3}$$

Where:

- π_i: probability that country i belongs to the structural zero group (i.e., always-zero process)
- $g(y_i \mid x_i)$: count model (e.g., PM or NBM)

Modeling approach: analysis workflow

Daily regression framework

We evaluate the impact of passenger air traffic (PAT) on daily confirmed cases (GC) across 174 daily snapshots (from 23/01/2020 to 13/07/2020), for each count model.

Model selection criterion

To compare model performance, we use the Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

where y_i is the observed count and $\hat{y}_i$ is the predicted count.
Lower RMSE indicates a better model fit.

Results

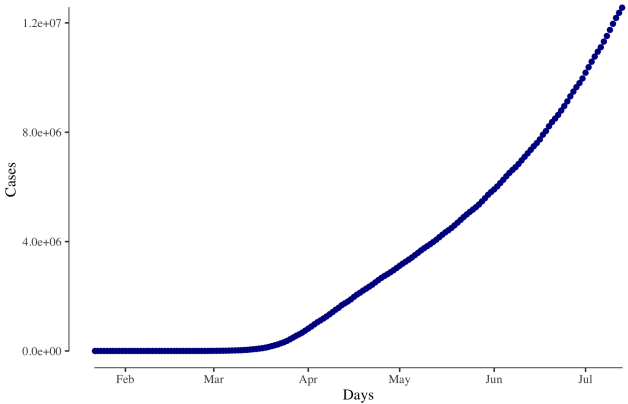

Figure 1: Daily confirmed COVID-19 cases worldwide (23/01/2020 - 13/07/2020)

Results: Zero-case period and model behavior

Key observations

- From 23/01/2020 to 09/04/2020, several countries reported zero confirmed cases, motivating the use of zero-inflated models (ZIM).
- ZIP and ZINB models mostly failed to converge, except for ZIP between 24/03/2020 and 09/04/2020, likely due to relatively small daily sample sizes, fixed predictor (PAT), and difficulty distinguishing structural from sampling zeros.
- As a result, ZIP and ZINB estimates were not used for further statistical interpretation.

Results: Zero-case period and model behavior

Key observations

- From 23/01/2020 to 09/04/2020, several countries reported zero confirmed cases, motivating the use of zero-inflated models (ZIM).
- ZIP and ZINB models mostly failed to converge, except for ZIP between 24/03/2020 and 09/04/2020, likely due to relatively small daily sample sizes, fixed predictor (PAT), and difficulty distinguishing structural from sampling zeros.
- As a result, ZIP and ZINB estimates were not used for further statistical interpretation.

Robustness of other models

- PM, QPM, and NBM all showed significant associations between passenger air traffic (PAT) and COVID-19 cases, with p-values < 0.05 for most days.
- The QPM produced insignificant results only between 22/02/2020 and 17/03/2020.
- Coefficient estimates were near zero, indicating that a one-unit increase in PAT leads to a proportional increase in reported cases.

Results: Variability in PAT coefficients

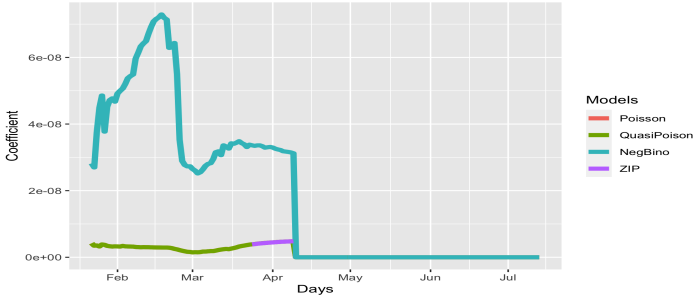

Figure 2: Estimated coefficients for PAT models

- Estimated coefficients remain close to zero (log-scale), meaning a one-unit increase in PAT leads to a proportional rise in COVID-19 cases close to 1.
- Despite the small magnitude, most coefficients are statistically significant throughout the study period.

Results: Model performance evaluation

Forecast and model comparison

- Using training data, we estimated model parameters and produced forecasts on test data.
- We compared models using root mean square error (RMSE).

Period	Best Model(s)	Period	Best Model(s)
22/01/2020 – 23/01/2020	PM, QPM	03/03/2020	NBM
24/01/2020 – 28/01/2020	PM, QPM	04/03/2020 – 23/03/2020	NBM
29/01/2020 – 21/02/2020	PM, QPM	24/03/2020 – 09/04/2020	ZIP
22/02/2020 – 24/02/2020	NBM	10/04/2020 – 13/07/2020	NBM
25/02/2020 – 02/03/2020	NBM		

- Early in the epidemic, simpler Poisson and QPM models performed best.
- Later, with higher variance and overdispersion, NBM provided better predictive accuracy.
- ZIP briefly performed best when there were many zeros.

Discussion

- Our study shows higher passenger air traffic (PAT) is associated with increased COVID-19 infections, supporting earlier evidence (Lau et al., 2020; Bogoch et al., 2020a).
- Air travel was key to seeding outbreaks globally (Ayittey et al., 2020; Wilson and Chen, 2020); local incidence rose as infected travelers arrived.
- In Africa, lower PAT partly explains fewer reported cases despite early concerns (Lone and Ahmad, 2020).
- Our multi-model approach over 174 days (with cross-validation) provides robust evidence; exception: insignificant effect during 22/02–17/03/2020, likely due to evolving travel restrictions (WHO, 2020b).
- Limitation: daily PAT data were unavailable; future studies should use detailed time series to better capture dynamic travel effects.

Meta-analysis implications

- **Multi-model approach with cross-validation** improves robustness and helps produce more reliable pooled estimates (Bogoch et al., 2020a; Grais et al., 2003).
- **Temporal variation** in air traffic effect (e.g., weaker influence during WHO pandemic declaration) suggests meta-analyses should include *dynamic moderators* such as policy changes and travel restrictions (Chinazzi et al., 2020; Kraemer et al., 2020).
- **Detailed time series data** on passenger mobility could reduce heterogeneity and enhance predictive accuracy in meta-analytic models.

Conclusion

Key findings

- Passenger air traffic had a significant effect on COVID-19 spread during the early pandemic.
- Count data models provided robust evidence of this relationship.
- Results complement meta-analytical evidence that human mobility drives epidemic spread.

Perspective

- Now that the pandemic has ended, meta-analyses can combine early and later data to assess whether these effects persisted across waves and policy changes.

References

[1] Sokadjo, Y.M., & Atchadé, M.N. (2020). The influence of passenger air traffic on the spread of COVID-19 in the world. *Transp. Res. Interdisc. Persp.*, 8, 100213. https://doi.org/10.1016/j.trip.2020.100213

[2] Chinazzi, M., Davis, J.T., Ajelli, M., et al. (2020). The effect of travel restrictions on the spread of COVID-19 outbreak. *Science*, 368(6489), 395–400.

[3] Wu, J.T., Leung, K., & Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the COVID-19 outbreak. *Lancet*, 395(10225), 689–697.

[4] Bogoch, I.I., Watts, A., Thomas-Bachli, A., et al. (2020). Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. *J. Travel Med.*, 27(2), taaa008.