The Journal of Open Source Software

epimargin: A Toolkit for Epidemiological Estimation, Prediction, and Policy Evaluation

Satej Soman[1], **Caitlin Loftus**[1], **Steven Buschbach**[1], **Manasi Phadnis**[1], **and Luís M. A. Bettencourt**[1, 2, 3, 4]

1 Mansueto Institute for Urban Innovation, University of Chicago **2** Department of Ecology & Evolution, University of Chicago **3** Department of Sociology, University of Chicago **4** Santa Fe Institute

DOI: 10.21105/joss.03464

Software

- Review
- Repository
- Archive

Editor: Nikoleta Glynatsi

Reviewers:

- @wxwx1993
- @dilawar

Submitted: 18 June 2021
Published: 09 September 2021

Summary

As pandemics (including the COVID-19 crisis) pose threats to societies, public health officials, epidemiologists, and policymakers need improved tools to assess the impact of disease, as well as a framework for understanding the effects and tradeoffs of health policy decisions. The `epimargin` package provides functionality to answer these questions in a way that incorporates and quantifies irreducible uncertainty in both the input data and complex dynamics of disease propagation.

The `epimargin` software package primarily consists of:

1. a set of Bayesian estimation procedures for epidemiological metrics such as the reproductive rate (R_t), which is the average number of secondary infections caused by an active infection
2. a flexible, stochastic epidemiological model informed by estimated metrics and reflecting real-world epidemic and geographic structure, and
3. a set of tools to evaluate different public health policy choices simulated by the model.

The software is implemented in the Python 3 programming language and is built using commonly-used elements of the Python data science ecosystem, including NumPy (Harris et al., 2020), Scipy (Virtanen et al., 2020), and Pandas (McKinney & others, 2011).

Statement of need

The `epimargin` software package is designed for the data-driven analysis of policy choices related to the spread of disease. It consists primarily of a set of estimators for key epidemiological metrics, a stochastic model for predicting near-future disease dynamics, and evaluation tools for various policy scenarios.

Included with the package are connectors and download utilities for common sources of disease data for the COVID-19 pandemic (the pressing concern at the time of writing), as well as a set of tools to prepare and clean data in a format amenable to analysis. It is widely understood that preprocessing epidemiological data is necessary to make inferences about disease progression (Gostic et al., 2020). To that end, `epimargin` provides commonly-used preprocessing routines to encourage explicit documentation of data preparation, but is agnostic to which procedures

are used due to the fact that all metadata required for certain preparations may not be uniformly available across geographies.

This same modularity extends to both the estimation procedures and epidemiological models provided by `epimargin`. While the package includes a novel Bayesian estimator for key metrics, classical approaches based on rolling linear regressions and Markov chain Monte Carlo sampling are also included. The core model class in `epimargin` in which these estimates are used is known as a compartmental model: a modeled population is split into a number of mutually-exclusive compartments (uninfected, infected, recovered, vaccinated, etc) and flows between these compartments are estimated from empirical data. The exact choice of compartments and interactions is left to the modeler, but the package includes several commonly-used models, as well as variations customized for specific policy questions (such as large-scale migration during pandemics, or the effects of various vaccine distribution policies).

For similar data downloading tools, see `covidregionaldata` (Palmer et al., 2021); for similar estimation tools, see `EpiEstim` (Cori et al., 2013) and `EpiNow2` (Abbott et al., 2020). While many of these tools are used in conjunction with each other, `epimargin` aims to offer tools for an end-to-end epidemiological workflow in one package, while offering the flexibility in estimator choice and data preparation methods.

Attempts to use a compartmental model to drive policy decisions often treat the systems under study as deterministic and vary parameters such as the reproductive rate across a range deemed appropriate by the study authors (Bubar et al., 2021). This methodology complicates incorporation of recent disease data and the development of theories for why the reproductive rate changes due to socioeconomic factors external to the model. The incorporation of stochasticity into the models from the outset allows for the quantification of uncertainty and the illustration of a range of outcomes for a given public health policy under consideration.

The `epimargin` package has been used to drive a number of research projects and inform policy decisions in a number of countries:

1. lockdown, quarantine planning, migrant return policies, and vaccine distribution in India and Indonesia (at the behest of national governments, regional authorities, and various NGOs)
2. an illustration of a novel Bayesian estimator for the reproductive rate as well as general architectural principles for real-time epidemiological systems (Bettencourt & Soman, 2020)
3. a trigger-based algorithmic policy for determining when administrative units of a country should exit or return to a pandemic lockdown based on projected reproductive rates and case counts (Malani et al., 2020)
4. a World Bank study of vaccination policies in South Asia ("South Asia Vaccinates," 2021)
5. a general framework for quantifying the health and economic benefits to guide vaccine prioritization and distribution (Malani et al., 2021)

Figures

Sample output for common workflows are illustrated in the following figures:

downloaded and cleaned time series

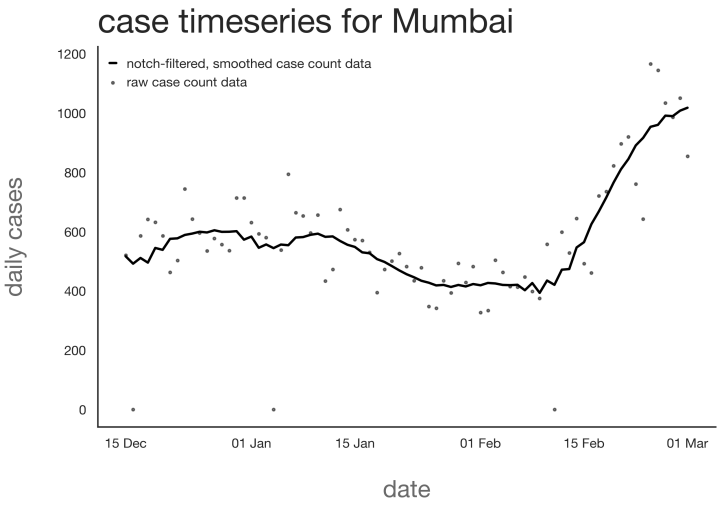

Figure 1: Raw and cleaned case count timeseries for Mumbai downloaded from COVID19India.org.

estimated reproductive rate

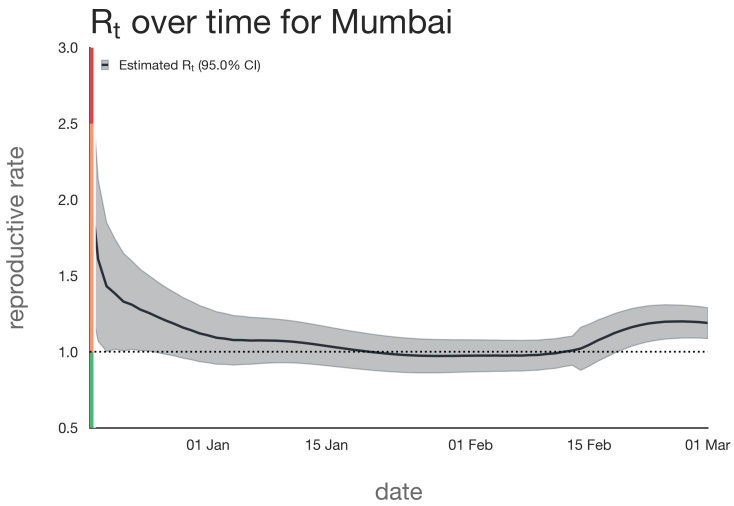

Figure 2: Estimated reproductive rate over time for Mumbai

forward projection/policy comparison

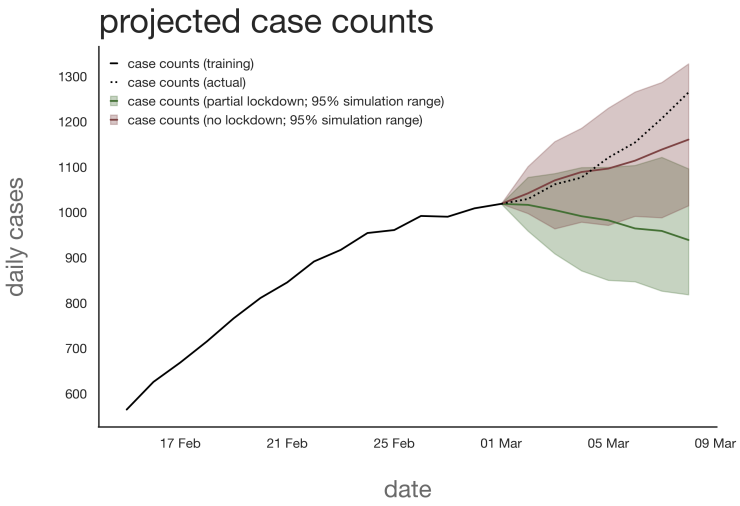

Figure 3: Projected case counts using a stochastic compartmental model and reproductive rate estimates

Acknowledgements

We acknowledge code review and comments from Gyanendra Badgaiyan (IDFC Institute), ongoing conversations with Anup Malani (University of Chicago) and Nico Marchio (Mansueto Institute) and helpful discussions with Katelyn Gostic (University of Chicago) and Sarah Cobey (University of Chicago).

References

Abbott, S., Hellewell, J., Thompson, R., Sherratt, K., Gibbs, H., Bosse, N., Munday, J., Meakin, S., Doughty, E., Chun, J., Chan, Y., Finger, F., Campbell, P., Endo, A., Pearson, C., Gimma, A., Russell, T., null, null, Flasche, S., … Funk, S. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with reservations]. *Wellcome Open Research*, *5*(112). https://doi.org/10.12688/wellcomeopenres.16006.2

Bettencourt, L., & Soman, S. (2020). Systems architecture for real time epidemiological prediction and control. *Mansueto Institute for Urban Innovation Research Paper*, *26*. https://doi.org/10.2139/ssrn.3748704

Bubar, K. M., Reinholt, K., Kissler, S. M., Lipsitch, M., Cobey, S., Grad, Y. H., & Larremore, D. B. (2021). Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science*, *371*(6532), 916–921. https://doi.org/10.1101/2020.09.08.20190629

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*, *178*(9), 1505–1512. https://doi.org/10.1093/aje/kwt133

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., & others. (2020). Practical considerations for

measuring the effective reproductive number, r t. *PLoS Computational Biology*, *16*(12), e1008409. https://doi.org/10.1371/journal.pcbi.1008409

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., & others. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Malani, A., Soman, S., Asher, S., Novosad, P., Imbert, C., Tandel, V., Agarwal, A., Alomar, A., Sarker, A., Shah, D., & others. (2020). *Adaptive control of COVID-19 outbreaks in india: Local, gradual, and trigger-based exit paths from lockdown*. National Bureau of Economic Research. https://doi.org/10.3386/w27532

Malani, A., Soman, S., Ramachandran, S., Chen, A., Lakdawalla, D., & others. (2021). *Vaccine allocation priorities using disease surveillance and economic data*. forthcoming.

McKinney, W., & others. (2011). Pandas: A foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

Palmer, J., Sherratt, K., Martin-Nielsen, R., Bevan, J., Gibbs, H., Group, C. C. W., Funk, S., & Abbott, S. (2021). Covidregionaldata: Subnational data for COVID-19 epidemiology. *Journal of Open Source Software*, *6*(63), 3290. https://doi.org/10.21105/joss.03290

South asia vaccinates. (2021). In *South Asia Economic Focus: Spring 2021*. World Bank.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., & others. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272.