covidregionaldata: Subnational data for COVID-19 epidemiology

Joseph Palmer[*1], **Katharine Sherratt**[*2], **Richard Martin-Nielson**[3], **Jonnie Bevan**[4], **Hamish Gibbs**[2], **CMMID COVID-19 Working Group**[2], **Sebastian Funk**[2], and **Sam Abbott**[†2]

1 Department of Biological Sciences, Royal Holloway University of London **2** Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine **3** None **4** Tessella

[*] co-first author [†] corresponding author

Summary

covidregionaldata is an R (R Core Team, 2020) package that provides an interface to subnational and national level COVID-19 data. The package provides cleaned and verified COVID-19 test-positive case counts and, where available, counts of deaths, recoveries, and hospitalisations in a consistent and fully transparent framework. The package automates common processing steps while allowing researchers to easily and transparently trace the origin of the underlying data sources. It has been designed to allow users to easily extend the package's capabilities and contribute to shared data handling. All package code is archived on Zenodo and GitHub.

Statement of need

The onset of the COVID-19 pandemic in late 2019 has placed pressure on public health and research communities to generate evidence that can help advise national and international policy in order to reduce transmission and mitigate harm. At the same time, there has been a renewed policy and public health emphasis on localised, subnational decision making and implementation (Hale et al., 2021; Liu et al., 2021). This requires reliable sources of data disaggregated to a fine spatial scale, ideally with few and/or known sources of bias.

At a national level, epidemiological COVID-19 data is available to download from official sources such as the World Health Organisation (WHO) (World Health Organisation, n.d.) or the European Centre for Disease Prevention and Control (ECDC) (European Centre for Disease Prevention and Control, n.d.). Many government bodies provide a wider range of country specific data, such as Public Health England in the United Kingdom (Public Health England, n.d.), and this is often the only way to access data at a subnational scale, for example by state, district, or province.

Sometimes collated from a range of national and subnational sources, these data come in a variety of formats, requiring users to check and standardise data before it can be combined or processed for analysis. This is a particularly time-consuming process for subnational data sets, which are often only available in the originating countries' languages and require customised methods for downloading and processing. This generates potential for errors through programming mistakes, changes to a dependency package, or unexpected changes to a data

source. This can lead to misrepresenting the data in ways which are difficult to identify. At best, an independent data processing workflow only slows down the pace of research and analysis, while at worst it can lead to misleading and erroneous results.

Because of these issues, it is important to develop robust tools that provide cleaned, checked and standardised data from multiple sources in a transparent manner. `covidregionaldata` provides easy access to clean data using a single-argument function, ready for analysing the epidemiology of COVID-19 from local to global scales, and in a framework that is easy to trace from raw data to the final standardised data set. Additional arguments to this function support users to, amongst other options, specify the spatial level of subnational data, return data with either standardised or country-specific variable names, or to access the full pipeline from raw to clean data. By default, cleaned and processed data is returned, however, the raw data from a source can also be returned. All data sources are checked daily via GitHub workflows and their status reported in the documentation section 'Data Status.' `covidregionaldata` largely depends on popular packages that many researchers are familiar with (such as the `tidyverse` suite (Wickham et al., 2019)) and can therefore be easily adopted by researchers working in R. In addition to code coverage tests, we test and report the status of all data sets daily.

Currently, `covidregionaldata` provides subnational data collated by official government bodies or by credible non-governmental efforts for 15 countries, including the UK, India, USA, and Brazil. It also provides an interface to subnational data curated by Johns Hopkins University (Dong et al., 2020), and the Google COVID-19 open data project (Wahltinez & others, 2020). National-level data is provided from the World Health Organisation (WHO) (World Health Organisation, n.d.), European Centre for Disease Prevention and Control (ECDC) (European Centre for Disease Prevention and Control, n.d.), Johns Hopkins University (JHU) (Dong et al., 2020), and the Google COVID-19 open data project (Wahltinez & others, 2020).

State of the field

Multiple organisations have built private COVID-19 data curation pipelines similar to that provided in `covidregionaldata`, including Johns Hopkins University (JHU) (Dong et al., 2020), Google (Wahltinez & others, 2020), and the COVID-19 Data Hub (Guidotti & Ardia, 2020). However, most of these efforts aggregate the data they collate into a separate data stream, breaking the linkage with the raw data, and often do not fully surface their data processing pipeline for others to inspect. In contrast `covidregionaldata` provides a clear set of open and fully documented tools that directly operate on raw data where possible in order to make the full data cleaning process transparent to end users.

Other interfaces to COVID-19 data are available in R, though there are fewer that provide tools for downloading subnational data for multiple countries and none that are known to the authors provide a consistent cleaning pipeline of the data sources they support. COVID-19 Data Hub (Guidotti & Ardia, 2020) provides cleaning functions, a wrapper to a custom database hosted by COVID-19 Data Hub, and access to snapshots of data reported historically. `Covdata` (Healy, 2020) provides weekly COVID-19 data updates as well as mobility and activity data from Apple (Apple, n.d.) and Google (Google, n.d.). `Sars2pack` (Davis & Carey, 2021) provides interfaces to a large number of data sets curated by external organisations. To our knowledge, none of these packages provide an interface to individual country data sources or a consistent set of data handling tools for both raw and processed data.

`covidregionaldata` has been used by researchers to source standardised data for estimating the effective reproductive number of COVID-19 in real-time both nationally and subnationally (Abbott et al., 2020). It has also been used in analyses comparing effective reproduction numbers from different subnational data sources in the United Kingdom (Sherratt et al., 2020), and estimating the increase in transmission related to the B.1.1.7 variant (Davies et

al., 2021). As well as its use in research it has also been used to visualise and explore current trends in COVID-19 case, deaths, and hospitalisations.

Acknowledgements

Funding statement

References

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

Apple. (n.d.). *COVID-19 – Mobility Trends Reports – Apple*. Retrieved May 11, 2021, from https://covid19.apple.com/mobility

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., Zandvoort, K. van, Silverman, J. D., Diaz-Ordaz, K., Keogh, R., Eggo, R. M., … Edmunds, W. J. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, *372*(6538), eabg3055. https://doi.org/10.1126/science.abg3055

Davis, S., & Carey, V. (2021). *sars2pack: COVID-19 data resources and analysis tools*. https://github.com/seandavi/sars2pack

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

European Centre for Disease Prevention and Control. (n.d.). *Download historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide*. Retrieved May 11, 2021, from https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Google. (n.d.). *COVID-19 Community Mobility Reports*. Retrieved May 11, 2021, from https://www.google.com/covid19/mobility/?hl=en

Guidotti, E., & Ardia, D. (2020). COVID-19 data hub. *Journal of Open Source Software*, *5*(51), 2376. https://doi.org/10.21105/joss.02376

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*(4), 529–538. https://doi.org/10.1038/s41562-021-01079-8

Healy, K. (2020). *Covdata: COVID-19 case and mortality time series*. http://kjhealy.github.io/covdata

Liu, Z., Guo, J., Zhong, W., & Gui, T. (2021). Multi-Level Governance, Policy Coordination and Subnational Responses to COVID-19: Comparing China and the US. *Journal of Comparative Policy Analysis: Research and Practice*, *23*(2), 204–218. https://doi.org/10.1080/13876988.2021.1873703

Public Health England. (n.d.). *About the data | Coronavirus in the UK*. Retrieved May 11, 2021, from https://coronavirus.data.gov.uk/details/about-data

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sherratt, K., Abbott, S., Meakin, S. R., Hellewell, J., Munday, J. D., Bosse, N., Jit, M., & Funk, S. (2020). *Evaluating the use of the reproduction number as an epidemiological tool, using spatio-temporal trends of the Covid-19 outbreak in England* (p. 2020.10.18.20214585). medRxiv. https://doi.org/10.1101/2020.10.18.20214585

Wahltinez, O., & others. (2020). *COVID-19 open-data: Curating a fine-grained, global-scale data repository for SARS-CoV-2*. https://goo.gle/covid-19-open-data

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

World Health Organisation. (n.d.). *WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data*. Retrieved May 11, 2021, from https://covid19.who.int/