B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2024, Vol. 0(0) 1–12

DOI: 10.1177/23998083241267331
journals.sagepub.com/home/epb
Sage

Packaging code and data for reproducible research: A case study of journey time statistics

Federico Botta
Department of Computer Science, University of Exeter, Exeter, UK
Fellow, The Alan Turing Institute, UK

Robin Lovelace
Fellow, The Alan Turing Institute, UK
Institute of Transport Studies, University of Leeds, Leeds, UK
Active Travel England, UK

Laura Gilbert
10DS, London, UK

Arthur Turrell
Data Science Campus, Office for National Statistics, UK

Abstract

The effective and ethical use of data to inform decision-making offers huge value to the public sector, especially when delivered by transparent, reproducible, and robust data processing workflows. One way that governments are unlocking this value is through making their data publicly available, allowing more people and organisations to derive insights. However, open data is not enough in many cases: publicly available datasets need to be accessible in an analysis-ready form from popular data science tools, such as R and Python, for them to realise their full potential. This paper explores ways to maximise the impact of open data with reference to a case study of packaging code to facilitate reproducible analysis. We present the `jtstats` project, which consists of a main Python package, and a smaller R version, for importing, processing, and visualising large and complex datasets representing journey times, for many transport modes and trip purposes at multiple geographic levels, released by the UK Department for Transport (DfT). `jtstats` shows how domain specific packages can enable reproducible research within the public sector and beyond, saving duplicated effort and reducing the risks of errors from repeated analyses. We hope that the `jtstats` project inspires others, particularly those in the public sector, to add value to their data sets by making them more accessible.

Corresponding author:
Federico Botta, Department of Computer Science, University of Exeter, North Park Road, Exeter EX44QF, UK.
Email: f.botta@exeter.ac.uk

Data Availability Statement included at the end of the article

Keywords
Data science for public good, government data, open source

Introduction

Recent years have seen an increasing volume of data being collected and generated, in a phenomenon that has been labelled a 'data revolution' (Kitchin, 2014). Our interactions with socio-technological systems create a vast amount of data on our behaviour, actions, society, and other aspects of our lives (Bannister and Botta, 2021; Botta and Gutiérrez-Roig, 2021; Conte et al., 2012; Lazer et al., 2009; Vespignani, 2009). Governments, statistical agencies, international organisations, and non-profit entities also collect, produce and release large volumes of data on the state of society, nations, economies, and a whole suite of useful indicators. To analyse these data sets and deliver the maximum amount of insight from them increasingly requires knowledge of data science techniques, such as predictive modelling, geographic data analysis and machine learning (Bengfort et al., 2018; Kuhn and Johnson, 2013; Lovelace et al., 2019b).

However, merely making datasets open access is not enough to unlock their potential value. Open access is a necessary but not sufficient condition to ensuring that the value invested in datasets is returned to society. Datasets need to be shared in formats that are easy to work with in the popular analytical tools used in data science, documented, and packaged in 'analysis-ready' formats to meet their potential. We loosely define 'analysis-ready' formats those data formats that enable the application of analytical tools in the corresponding programming language, such as `pandas` dataframes in Python, or data frames or tibbles in R. Examples of this can be found in 'data packages' such as 'tidycensus' R package (Walker, 2023), which has been cited by more than 100 papers, and the 'census' Python package (Carbaugh et al., 2022). Both packages provide programmatic interfaces to the United States Census Bureau's API and have been used to support a range of studies, including a data-driven investigation of urban trees throughout US states (McCoy et al., 2022) and – in a completely different field – an exploration of the correlates and possible causal factors associated with rates of child mistreatment across the US (Mayer, 2023). The 'stats19' R package (Lovelace et al., 2019a) is another example, that provides programmatic access to analysis-ready data on road traffic casualty records collected by police forces across Great Britain and has been used to support methodological (Francis et al., 2023; Gilardi et al., 2022) and empirical (Kokayi et al., 2023) work. It should be noted here that these papers demonstrate good practice by citing the data (and other software/technical) packages underlying the analysis, with Kokayi et al. (2023) stating that the underlying data was accessed 'using stats19 [R] package created by Lovelace et al. (2019a)', for example. However, not all academic papers will cite data packages and not all data-driven work is written-up as academic papers, meaning that citations to data packages capture only a fraction of their impact. Such wider impacts are can also be illustrated by 'stats19'. The package was used as the basis of visualisations of road traffic casualty statistics for the parliamentary library's data dashboard on 'Constituency data: road traffic collisions and casualties'' hosted at commonslibrary.parliament.uk to support evidence-based decision-making by Members of Parliament and others. The package was also the basis of a blog post 'Have UK roads become safer in the past 40 years?' in which the author states that 'I have made use of a very handy R package that saved plenty of time in preparing the data' by a researcher at the Apteco marketing software development and customer analytics company, highlighting the fact that open data and packages enabling analysis-ready analysis help collaboration between academia, government, and industry.[1]

The wider point is that datasets on one topic (journey times in this case) may be relevant to research outside that topic. Additionally, lowering entry barriers to data sets to allow researchers to

more easily explore what information is in the data can enable a wide range of people to explore the data, providing more opportunities for usage of the data.

Here, we use an England-based data set that is regularly released by the UK Department for Transport (DfT) as an example of how publicly available data can be coupled with publicly available code to maximise value – including to the public sector bodies who may be most interested in using it. It is important to note that there is nothing special about these data or this example – it merely serves to demonstrate a point, and we could have chosen any number of other examples of data released by other public sector bodies. Here, though, our case study is data released by DfT on journey time statistics (Department for Transport, 2019). To show the additional insights that can be gained when data are more accessible in a format that allows easy data linkage, we show a visual comparison of journey time statistics to data on relative deprivation, and so provide a way of assessing the accessibility of places and services with socio-economic information of the population. We emphasise that it is the combination of code, data and geometries with a clear linking method which allows for such analyses to be performed, rather than any individual part. Data on relative deprivation is retrieved directly from the UK Department for Levelling Up, Housing and Communities website (Department for Levelling Up, Housing & Communities, 2019). We also emphasise that we use data on deprivation levels as an interesting example of how packaging code and data can allow for easy and efficient data linkage. Whilst there are a number of questions which could be studied in the intersection between accessibility of services and deprivation of different areas, here this is only an example and we could have chosen other openly available data sets instead.

In our particular case study of journey time statistics, the data are crucial for understanding a broad range of questions in both academic research and policy. For example, the interplay between travel times and socio-economic inequalities is particularly acute in some types of area where inequalities may be exacerbated by poor access to education or services. Therefore, the availability and accessibility of data (and tools) to study these issues is crucial for improving our cities and rural areas. The open source software presented here aims to provide simple tools to perform this analysis and to make these valuable data more easily accessible for research. Starting from a relatively complex data structure, we provide open tools that allow most data scientists to work with these data using standard data science pipelines.

The data sets that we attempt to make more accessible in this paper, and the open source software that facilitates this within a reproducible research pipeline, are relevant to a wide range of academic research and policy questions. Variable travel times underlie many social phenomena, ranging from spatial variability in energy use (Breheny, 1995) to inequalities in educational opportunities (Moreno-Monroy et al., 2018). An easier and wider availability of data on travel times and accessibility of destinations can allow urban geographers and data scientists to study the interplay between more compact cities and pollution, inequalities and segregation. This makes open data on journey times valuable, particularly when provided at high spatial resolution, and when provided with reference to a wide range of transport modes and purposes. We envisage our work to support this, and many other, applications.

A further added value of making data sets openly available is that they can be combined and enriched by other existing open data sets. To demonstrate this, the main version of the package presented here, developed in Python, is also able to retrieve data on relative deprivation level directly from the government website. The English Indices of Multiple Deprivation (IMD) is published on the Department for Levelling Up, Housing and Communities' (formerly the Ministry of Housing, Communities & Local Government) website and provides a measure of relative deprivation of areas in England across seven different domains: Income; Employment; Education; Health; Crime; Barriers to housing & services; and Living environment.

Data – journey time statistics (JTS)

This data set is released by DfT and contains statistics on modelled journeys to key services in England, such as employment centres, schools and hospitals. The data is broadly divided into three key statistics: average minimum journey time, which is the shortest travel time to a specific service by mode of transport (car, walking, cycling, or public transport); origin indicators, measuring the number of different services in an area that can be reached in a given time (such as 30 min); and destination indicators, measuring the proportion of users that can access a service in a given time. The data is provided at different levels of spatial aggregation, from Lower Layer Super Output Areas (LSOA) to national, regional, and local authority level.

It is important to highlight that the JTS data does not represent actual journeys, but rather idealised trips generated by DfT using commercial software. In particular, the journeys are modelled based on the average journey completed on a Tuesday in the second week of October of the corresponding year during the morning peak travel time between 7am and 10am. More information on the JTS data and how the journeys are calculated and aggregated can be found on the DfT website (Department for Transport, 2019).

The data are released by DfT in the *Open Document Format*, more specifically *Open Document Spreadsheet*, which is an open file format for sharing documents, texts, and spreadsheets. However, the complex nature of this data set with a large number of files and tables, formatting issues such as the inconsistent presence of footnotes and superscripts and large number of observations, make it difficult for data scientists to quickly explore the data and efficiently perform any analysis; additionally, whilst the *Open Document Spreadsheet* format could be directly read into both R and Python, the size and number of the files, coupled with the inefficiency of the format, mean that this is a very slow and memory-intensive process, which we found too resource intensive on a regular laptop. We believe that this inefficiency can act as a barrier to all interested researchers, broadly intended, who want to explore the data set to better understand what information is included in it. Having a lower entry barrier with a much more efficient way of exploring the data enables a broader range of people to experiment with the data, thus maximising the potential impact of the data set itself. The open software that we released alongside this paper first converts the files to the simpler `.csv` format, and then processes it to make it available in other standard formats used in data science analysis. As discussed in the Conclusion section below, future work should consider using state-of-the-art open formats, such as Apache's Parquet format, to improve performance and interoperability.

Code

The open source software presented in this paper can be integrated into a reproducible research pipeline that can then be used by academics and policymakers alike. The main version of this package has been developed in Python, with the R version replicating some of the key functionalities, reflecting the increasing popularity of Python as the main programming language in data science pipelines. The software, both in its main Python version as well as that in R, is built in a simple, modular structure to allow ease of use as well as to enable data scientists to easily add features over time.

Both implementations of the software rely on a version of the JTS data which has been converted from the original `.ods` file format to a more standard `.csv` format via a command line script (also openly available on the GitHub repository hosting the Python and R packages). Figure 1 depicts an overview of the workflow of the packages presented here.

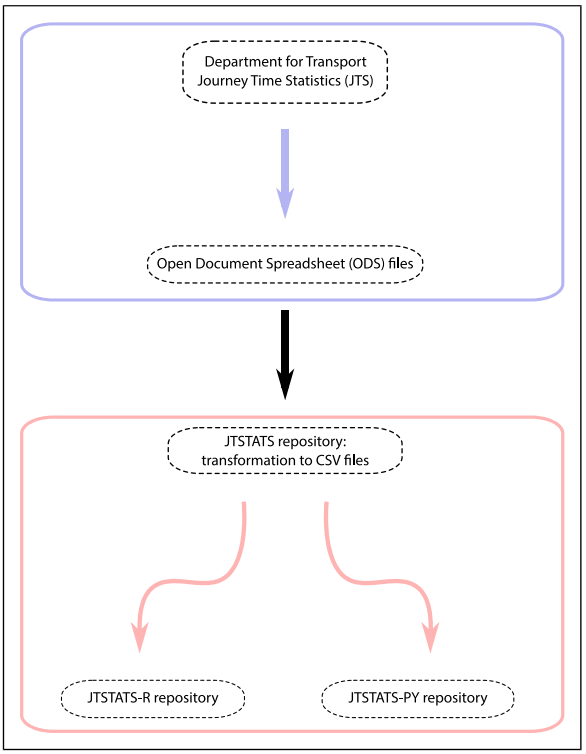

Figure 1. Code and data workflow | Packaging code and data to enable easy access to and analysis of government data is becoming increasingly important to maximise the value to public sector bodies of making their data openly available. Here, we use data on journey time statistics (JTS) from the department for transport (DfT) as a case study. The data is generated by DfT and released publicly in *open document spreadsheet* format (green box at top). We then convert the data to a more user friendly *csv* format and use that as the basis of our *R* and *Python* packages (red box at bottom).

Python

The Python implementation of this package has been developed in a Poetry environment running Python 3.9.1 and consists of four main parts:

- `jts`: this implements the main functionalities to access the JTS data. It retrieves the required data based on the user input (see Table 1 for an example), and returns it in a simple `pandas` data frame format, ready for analysis, and without any of the unnecessary formatting features included in the original *Open Document Spreadsheet* files, such as superscripts and footnotes;
- `imd`: this retrieves the additional IMD data set based on the specified user input for the year and specific domain of deprivation;
- `geo`: this retrieves the spatial (GeoJSON) files for the LSOAs and local authorities;
- `plot`: this implements a few plotting functionalities, such as those used to generate the maps in Figure 2. This allows a simple, initial exploration of the data.

A brief tutorial script is also provided on the GitHub page to demonstrate basic functionalities and usage of this package (https://github.com/datasciencecampus/jtstats-py).

Table 1. Retrieval of LSOA-level JTS data using the Python module The Python version of the module allows easy retrieval of the JTS data using the get_jts() function. The key parameters that can be specified in order to specify what data to retrieve are: type_code, spec, sheet and table_code (even though this last parameter is only rarely needed). The table here reports the values to be used for these parameters in order to retrieve the JTS data at the LSOA level, which are likely to be the most commonly used tables. The sheet parameter should be simply set to the year for which data is needed, for example, sheet = '2019'. The get_jts() function returns a pandas dataframe. The size of the returned dataframe is given in the last column (in the format # *rows* × # *columns*) of the table, for the specific case of 2019 data.

Table title	Description	type_code	Spec, sheet	Table size (# rows × # columns)
JTS0501	Journey times for employment centres by mode of travel (LSOA)	jts05	employment	32844 × 113
JTS0502	Journey times for primary schools by mode of travel (LSOA)	jts05	primary	32844 × 41
JTS0503	Journey times for secondary schools by mode of travel (LSOA)	jts05	secondary	32844 × 41
JTS0504	Journey times for further education by mode of travel (LSOA)	jts05	further	32844 × 41
JTS0505	Journey times for GPs by mode of travel (LSOA)	jts05	gp	32844 × 41
JTS0506	Journey times for hospitals by mode of travel (LSOA)	jts05	hospital	32844 × 41
JTS0507	Journey times for food stores by mode of travel (LSOA)	jts05	food	32844 × 41
JTS0508	Journey times for town centres by mode of travel (LSOA)	jts05	town	32844 × 41
JTS0509	Journey times to pharmacy by cycle and car (LSOA)	jts05	pharmacy	32844 × 23

Thanks to its modular structure, retrieving different data sets, or different variables within a data set, can be easily done within this software. For instance, the data needed for Figure 2 can be retrieved by simply calling `jts.get_jts(type_code = 'jts05', purpose = 'employment', sheet = 2019)`, whereas the data for Figure 3 can be retrieved by `jts.get_jts(table_code = 'jts0101', sheet = 'JTS0101')`. Table 1 provides a simple description of how the JTS05 tables can be retrieved using the Python module, along with some information on size of the retrieved data. Full information on how to retrieve the remaining JTS tables can be found at https://github.com/datasciencecampus/jtstats.

To support further future development of the package, we also implemented, and provided alongside the package, a series of tests written using `pytest`. These test the retrieval of each JTS table by checking a pre-specified set of known values in the tables. This enables any interested developer to implement changes to the package and test that they result in the same output data as what is contained in the raw data set. Additionally, these tests also act as a partial validation step of the package because they confirm that, at least for some predefined entries, the output data of the Python package contains the same values as the raw underlying data. This validation is crucial when developing tools that are aimed at facilitating the use and accessibility of data sets, since any mistake in the data processing done by such tools will affect all analyses based on them. However, it is important to highlight the challenge in ensuring that all data points, rather than predefined ones, match value with the raw data. This would be potentially intractable and not feasible, as this validation requires an element of manual verification to develop the tests released here.

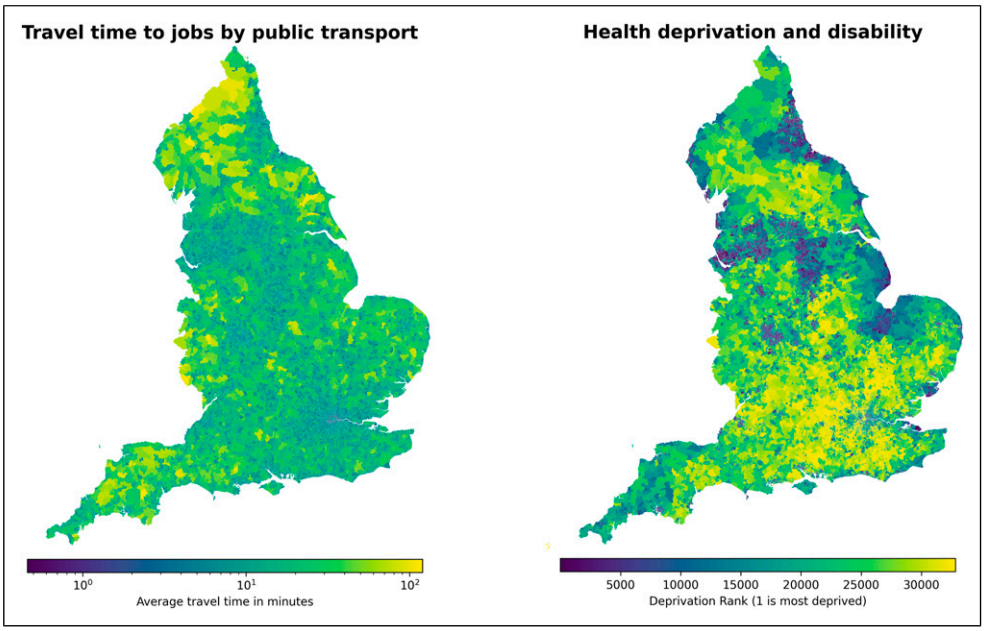

Figure 2. Journey time statistics and deprivation data we demonstrate the potential use of our package by depicting two variables from the data sets made easily available for analysis within our code. *(left)* Here, we depict the travel time to employment centres with 100 to 499 jobs by public transport as made available by the *journey time statistics* data set published by the *department for transport*. Travel time is capped to a maximum of 120 min. *(right)* We also present data on relative deprivation. In particular, we show here deprivation ranks related to health and disability, as published by the Ministry of Housing, communities local government in 2019.

A final note to mention regards missing values in some of the JTS tables. In particular, the JTS09 tables contains entries of the form '..', which have been replaced with missing values in the package (as *NumPy* `NaN` values); according to the JTS data description, such entries correspond to trips of duration 240 min or more. Similarly, entries in the JTS0930 table which are missing (represented by '–' in the original files) have also been replaced by *NumPy* missing values.

R

`jstats` has also been implemented as an R package which can be installed and loaded with the following commands from the R console (we recommend using a modern IDE such as RStudio or VSCode):

```
install.packages('remotes')
remotes::install_github('datasciencecampus/
jtstats-r')
library(jtstats)
```

The R version of the package replicates some of the key functionalities of the Python version presented above.

After the package has been installed and loaded, you can start using it to get any of the 192 tables available from the JTS project. You can get the table JTS0101 with the following command:

```
jts0101_data = get_jts(code = 'JTS0101')
```

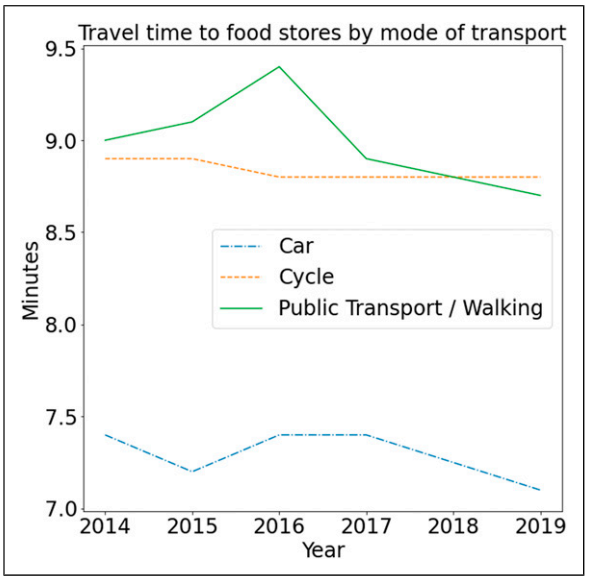

Figure 3. Travel time to food stores by mode of transport | Our package enables easy access to the data by returning a nicely formatted data frame which can be used for analysis. Here, we depict an example of the data which can be retrieved with our package: travel time to food stores over the years, disaggregated by mode of transport.

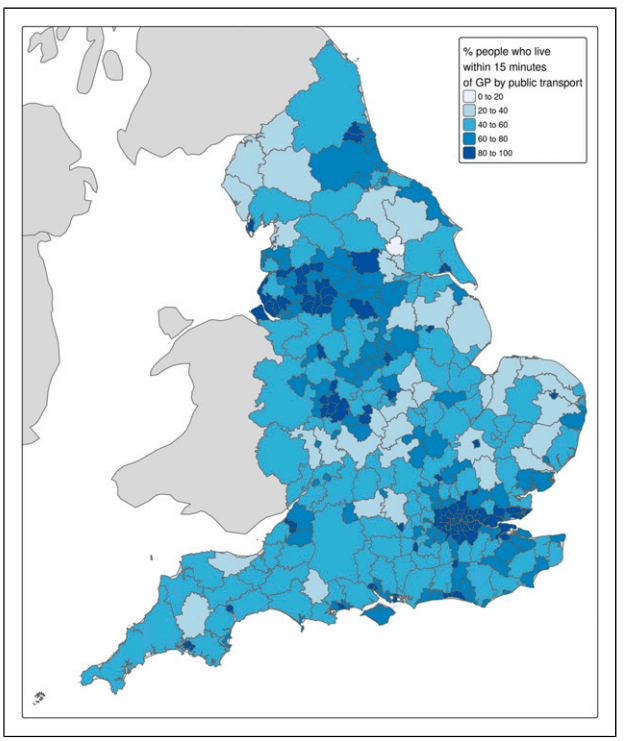

Figure 4. Travel time to healthcare by public transport: colour is proportional to the percentage of people in each LSOA who can reach general practitioner (GP) locations by rail or bus within 15 min.

You can get geographic datasets (such as LSOA boundaries) for relevant tables by setting the geo argument to TRUE, as follows:

```
lsoa_employment = get_jts(type = 'jts05',
purpose = 'employment' sheet =, 2017, geo = TRUE)
```

The `purpose` parameter refers to the trip purpose as defined in the JTS data set (analogous to the same parameter in the Python version of the package). Another example, which highlights how getting the data into R can provide access to high level data visualisation packages (in this case the tmap package), is shown below. This downloads data on access to healthcare and generates a map showing the percentage of people in each small area (LSOA) who can access General Practitioner service locations (GPs) within 15 min (Figure 4).

```
jts_geo = get_jts(
  type = "jts04",
  purpose = "GPs",
  sheet = 2017,
  geo = TRUE
  )
library(tmap)
bb = sf::st_bbox(jts_geo)
tm_shape(uk, bbox = bb) +
  tm_polygons() +
  tm_shape(ie) +
  tm_polygons() +
  tm_shape(jts_geo) +
  tm_polygons(
    "GPPT15pct",
    palette = "Blues"
    ) +
tm_layout(
  legend.position = c("right", "top"),
  outside = TRUE
  )
```

Conclusion

Our experience of developing Python and R packages to enable the import of JTS data more easily demonstrates how software development can enable reproducible data science for public bodies and beyond. Government bodies are increasingly tasked with providing large and complex data sets to multiple stakeholders. 'Packaging-up' tools to enable direct access to them can support this task while building software development capacity and ensuring transparency and reproducibility of research with official datasets. Transparent and reproducible research is also important in promoting an ethical use of data, as everyone is able to see and reproduce the process used to go from the original data set to the output of an analysis.

The JTS data files used here showcase the challenges faced by many public sector bodies wishing to obtain the maximum value from their data: developing documentation, providing pipelines for data cleaning, providing context (e.g. provision of geometries alongside entries representing administrative zones), and releasing data in the most relevant formats, all take extra resources. Developing tools such as that presented here requires significant initial time investment in writing and testing the code, ensuring it works as expected and that it performs the right processing of the data. We suggest that adopting this approach of packaging code and data together from the onset of a project can be beneficial in the long term. Investing in this can enhance the ability of researchers and

analysts to work with the data. The case study presented here not only provides tools for accessing the data, but it also encourages a range of users by supporting two of the most popular languages for data science. Whilst not all policy makers will have the skills or background to use such tools, we anticipate these becoming more common and widespread in the future, as a larger number of people are trained in data science methods. The true potential of publicly available data sets can only be realised when a broad range of data scientists and analysts can work with them, with minimal barriers to entry. Other benefits of the approach include reducing risks of each researchers introducing errors into their work, leading to results that are not only erroneous, but hard for others to fix further down the 'data pipeline'.

The approach we have taken is not without limitations or challenges. Indeed, there are many additional improvements that could be made, including the provision of intermediate data files in a compressed and analysis-ready format (such as Apache's Parquet format), enabling future proof and high performance in memory analysis using the multi-lingual Arrow framework (currently files are provided in .csv format); further development of the R version and more rigorous tests to ensure that the outputs from R and Python implementations of `jstats` are the same; and even more importantly, performing rigorous and comprehensive validation of the processing to ensure that our tool replicates the raw data faithfully is challenging, as validating each data point would not be feasible nor practical. Hence, our proposed approach (in the Python release of the package) to test a preselected subset of data points and ensure that the raw values are the same as the corresponding processed values; and finally, better integration with upstream processes that led to the creation of the .ods files (raising the question of how to design data collection projects such as JTS around data science tools). Whilst we argue that our work allows for greater transparency and reproducibility, this does not necessarily apply to the entire analysis undertaken subsequently, where researchers and data scientists would need to commit to open up their code. Finally, we encourage the development of such tools and packages by both data scientists in academia and the public sector, but we also acknowledge that this may not always be possible when time or resources are limited.

The `jstats` project presented here is an example of how publicly available government data can be made available for reproducible analysis by developing open source tools packaged in popular and languages for data science. We urge others working with and publishing open datasets to develop packages for effective and transparent research. Publishing such packages alongside open datasets will add vast amounts of value to government digital assets that are in the public domain for the benefit of all.

Acknowledgements

We would like to thank Greg Haigh (Advanced Analytics, Department for Transport) and Stephen Reynolds (Travel and Environment Data and Statistics (TRENDS) Division, Department for Transport) for feedback on the paper.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The views expressed are those of the authors and may not reflect the views of the Office for National Statistics, Active Travel England, 10DS, or the wider UK Government.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Economic and Social Research Council (ESRC) & ADR UK as part of the ESRC-ADR UK No.10 Data Science (10DS) fellowship in collaboration with 10DS and ONS (Federico Botta, grant number ES/W003937/1; Robin Lovelace, grant number ES/W004305/1).

ORCID iDs

Federico Botta https://orcid.org/0000-0002-5681-4535
Robin Lovelace https://orcid.org/0000-0001-5679-6536

Data availability statement

The data set is publicly available at https://www.gov.uk/government/statistics/journey-time-statistics-england-2019

Note

1. See apteco.com/insights/blog/have-uk-roads-become-safer-past-40-year.

References

Bannister A and Botta F (2021) Rapid indicators of deprivation using grocery shopping data. *Royal Society Open Science* 8(12): 211069.

Bengfort B, Bilbro R and Ojeda T (2018) *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. Sebastopol, CA: O'Reilly Media, Inc.

Botta F and Gutiérrez-Roig M (2021) Modelling urban vibrancy with mobile phone and openstreetmap data. *PLoS One* 16(6): e0252015.

Breheny M (1995) The compact city and transport energy consumption. *Transactions of the Institute of British Geographers* 20(1): 81. DOI: 10.2307/622726. https://www.jstor.org/stable/622726?origin=crossref

Carbaugh J and Gregg F, Contributors (2022) Census https://github.com/datamade/census

Conte R, Gilbert N, Bonelli G, et al. (2012) Manifesto of computational social science. *The European Physical Journal - Special Topics* 214(1): 325–346.

Department for Levelling Up (2019) *Housing & Communities*. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Department for Transport (2019) https://www.gov.uk/government/collections/journey-time-statistics

Francis J, Bright J, Esnaashari S, et al. (2023) Unsupervised feature extraction of aerial images for clustering and understanding hazardous road segments 13: 10922. DOI:10.1038/s41598-023-38100-1. https://www.nature.com/articles/s41598-023-38100-1

Gilardi A, Mateu J, Borgoni R, et al. (2022) Multivariate hierarchical analysis of car crashes data considering a spatial network lattice 185(3): 1150–1177. DOI: 10.1111/rssa.12823.

Kitchin R (2014) The data revolution: big data. In: *Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks, CA: Sage.

Kokayi A, Shiode S and Shiode N (2023) Geographical exploration of the underrepresentation of ethnic minority cyclists in England 15: 5677. https://www.mdpi.com/2071-1050/15/7/5677

Kuhn M and Johnson K (2013) *Applied Predictive Modeling*. Salmon Tower Building, New York City: Springer, Vol. 26.

Lazer D, Pentland AS, Adamic L, et al. (2009) Social science. Computational social science. *Science* 323: 721–723. DOI: 10.1126/science.1167742.

Lovelace R, Morgan M, Hama L, et al. (2019a) Stats19 A package for working with open road crash data 4(33): 1181. DOI:10.21105/joss.01181.

Lovelace R, Nowosad J and Muenchow J (2019b) *Geocomputation with R*. Boca Raton, FL: Chapman and Hall/CRC.

Mayer DJ (2023) Social capital and the nonprofit infrastructure; an ecological study of child maltreatment 51(5): 1961–1976. DOI:10.1002/jcop.22984.

McCoy DE, Goulet-Scott B, Meng W, et al. (2022) Species clustering, climate effects, and introduced species in 5 million city trees across 63 us cities. *Elife* 11: e77891.

Moreno-Monroy AI, Lovelace R and Ramos FR (2018) Public transport and school location impacts on educational inequalities: insights from São Paulo. *Journal of Transport Geography* 67: 110–118. DOI: 10.1016/j.jtrangeo.2017.08.012. https://www.sciencedirect.com/science/article/pii/S0966692316303453

Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325: 425–428. DOI: 10.1126/science.1171990.

Walker K (2023) *Analyzing US Census Data: Methods, Maps, and Models in R*. Boca Raton, FL: Chapman and Hall/CRC.

Federico Botta is Senior Lecturer in Data Science at the University of Exeter.

Robin Lovelace is Associate Professor of Transport Data Science at the University of Leeds. He is also Head of Data and Digital at Active Travel England. The views expressed here are those of the author and may not reflect the views of Active Travel England and the wider UK government.

Laura Gilbert is Chief Analyst and Director of Data Science at 10 Downing Street.The views expressed here are those of the author and may not reflect the views of 10 Downing Street and the wider UK government.

Arthur Turrell is Deputy Director at the UK Office for National Statistics Data Science Campus. The views expressed here are those of the author and may not reflect the views of the Office for National Statistics and the wider UK government.