![JOSS - The Journal of Open Source Software]

stats19: A package for working with open road crash data

R Lovelace[1], M Morgan[1], L Hama[1], and M Padgham[1]

1 Institute for Transport Studies (ITS) and Leeds Institute for Data Analytics (LIDA), University of Leeds **2** ATFutures GmbH.

Summary

stats19 provides functions for downloading and formatting road crash data. Specifically, it enables access to the UK's official road traffic casualty database, STATS19 (the name comes from the form used by the police to record car crashes and other incidents resulting in casualties on the roads). Finding, reading-in and formatting the data for research can be a time consuming process subject to human error, leading to previous (incomplete) attempts to facilitate the processes with open source software (Lovelace & Ellison, In press). **stats19** speeds-up these data access and cleaning stages by streamlining the work into 3 stages:

1. **Download** the data, by `year`, `type` and/or `filename`. An interactive menu of options is provided if there are multiple matches for a particular year.
2. **Read** the data in and with appropriate formatting of columns.
3. **Format** the data so that labels are added to the raw integer values for each column.

Functions for each stage are named `dl_stats19()`, `read_*()` and `format_*()`, with `*` representing the type of data to be read-in: STATS19 data consists of `accidents`, `casualties` and `vehicles` tables, which correspond to incident records, people injured or killed, and vehicles involved, respectively.

The package is needed because currently downloading and formatting STATS19 data is a time-consuming and error-prone process. By abstracting the process to its fundamental steps (download, read, format), the package makes it easy to get the data into appropriate formats (of classes `tbl`, `data.frame` and `sf`), ready for for further processing and analysis steps. We developed the package for road safety research, building on a clear need for reproducibility in the field (Lovelace, Roberts, & Kellar, 2016) and the importance of the geo-location in STATS19 data for assessing the effectiveness of interventions aimed to make roads safer and save lives (Sarkar, Webster, & Kumari, 2018). A useful feature of the package is that it enables creation of geographic representations of the data, geo-referenced to the correct coordinate reference system, in a single function call (`format_sf()`). The package will be of use and interest to road safety data analysts working at local authority and national levels in the UK. The datasets generated will also be of interest to academics and educators as an open, reproducible basis for analysing large point pattern data on an underlying route network, and for teaching on geography, transport and road safety courses.

References

Lovelace, R., & Ellison, R. (In press). Stplanr: A Package for Transport Planning. *The R Journal*.

Lovelace, R., Roberts, H., & Kellar, I. (2016). Who, where, when: The demographic and geographic distribution of bicycle crashes in West Yorkshire. *Transportation Research Part F: Traffic Psychology and Behaviour*, Bicycling and bicycle safety, *41, Part B*. doi:10.1016/j.trf.2015.02.010

Sarkar, C., Webster, C., & Kumari, S. (2018). Street morphology and severity of road casualties: A 5-year study of Greater London. *International Journal of Sustainable Transportation*, *12*(7), 510–525. doi:10.1080/15568318.2017.1402972