The Journal of Open Source Software

TransBigData: A Python package for transportation spatio-temporal big data processing, analysis and visualization

Qing Yu [1] **and Jian Yuan** [1¶]

1 Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao'an Road, Shanghai 201804, People's Republic of China ¶ Corresponding author

DOI: 10.21105/joss.04021

Software

- Review
- Repository
- Archive

Editor: Martin Fleischmann

Reviewers:

- @jgaboardi
- @anitagraser

Submitted: 09 December 2021
Published: 02 March 2022

Summary

In recent years, data generated in the field of transportation has begun to explode. Individual continuous tracking data, such as mobile phone data, IC smart card data, taxi GPS data, bus GPS data and bicycle sharing order data, also known as "spatio-temporal big data" or "Track &Trace data" (Harrison et al., 2020), has great potential for applications in data-driven transportation research. These spatio-temporal big data typically require three aspects of information (Zhang et al., 2021): Who? When? Where? They are characterized by high data quality, large collection scope, and fine-grained spatio-temporal information, which can fully capture the daily activities of individuals and their travel behavior in the city in both temporal and spatial dimensions. The emergence of these data provides new ways and opportunities for potential transportation demand analysis and travel mechanism understanding in supporting urban transportation planning and management (Chen et al., 2021; Zhang et al., 2020). However, processing with these multi-source spatio-temporal big data usually requires a series of similar processing procedure (e.g., data quality assessment, data preprocessing, data cleaning, data gridding, data aggregation, and data visualization). There is an urgent need for a one-size-fits-all tool that can adapt to the various processing demands of different transportation data in this field.

State of the art

Typical processing for spatiotemporal data analysis involves multiple procedures, including data acquisition, data preprocessing, data analysis, data visualization, etc. Currently, there exists several open source packages in these domains:

- Data acquisition: In this aspect, existing software mainly provides data acquisition of basic geospatial data. `osmnx` (Boeing, 2017) is a Python package to download geospatial data from OpenStreetMap. Some packages provide tools to generate synthetic mobility data using standard mathematical models. `scikit-mobility` (Pappalardo et al., 2019) is a library that allows for managing and generation of mobility data of various formats.
- Data preprocessing and data analysis: Spatio-temporal data in the field of transportation involves multiple source of data. Most existing tools are developed for specific type of data, such as trajectory data, air traffic data, etc. `MovingPandas` (Graser, 2019) provides trajectory data structures and functions for data exploration and visualization. `PTRAIL` (Haidri et al., 2021) is a library for mobility data preprocessing especially in feature generation and trajectory interpolation. `traffic` (Olive, 2019) is a toolbox for processing and analyzing air traffic data. `trackintel` (Axhausen, 2007) is a framework for spatio-temporal analysis of tracking data focusing on human mobility. `PySAL` (S. J.

Rey et al., 2021; S. Rey & Anselin, 2007) is a family of packages that allows for advanced geospatial data science, which supports the development of high-level applications.
- Data visualization: Apart from data processing, `MovingPandas` also support trajectory visualizations. `moveVis` (Schwalb-Willmann et al., 2020) provides tools to visualize movement data and temporal changes of environmental data by creating video animations. `TraViA` (Siebinga, 2021) provides tools to visualize and annotate movement data in an interactive approach.

The above-mentioned libraries provide preprocessing and geometric analysis functions from sepcific aspects. However, much remains to be done in dealing with the task of transforming raw spatio-temporal data into valuable insights, and these libraries provide no single solution. Thus, a library compatible with existing tools that provides and analysis framework for transportation spatio-temporal big data can effectively facilitate the research progress in this field.

Statement of need

`TransBigData` is a Python package developed for transportation spatio-temporal big data processing, analysis, and visualization that provides fast and concise methods for processing taxi GPS data, bicycle sharing data, and bus GPS data, for example. Further, a variety of processing methods for each stage of transportation spatio-temporal big data analysis are included in the code base. `TransBigData` provides clean, efficient, flexible, and easy to use API, allowing complex data tasks to be achieved with concise code, and it has already been used in a number of scientific publications (Li et al., 2021; Yu, Zhang, Li, Sui, et al., 2020; Yu, Zhang, Li, Song, et al., 2020; Yu et al., 2021).

For some types of data, `TransBigData` also provides targeted tools for specific needs, such as extraction of origins and destinations (OD) of taxi trips from taxi GPS data and identification of arrival and departure information from bus GPS data.

Currently, `TransBigData` mainly provides the following methods:

- *Data Quality*: Provides methods to quickly obtain the general information of the dataset, including the data amount, the time period, and the sampling interval.
- *Data Preprocess*: Provides methods to fix multiple types of data error.
- *Data Gridding*: Provides methods to generate multiple types of geographic grids (rectangular and hexagonal) in the research area. Provides fast algorithms to map GPS data to the generated grids (Figure 1).
- *Data Aggregating*: Provides methods to aggregate GPS and OD data into geographic polygons.
- *Trajectory Processing*: Provides quick methods to re-organize the data structure and implement data augmentation from various data formats, including generating trajectory linestrings from GPS points, and trajectory densification, etc.
- *Data Visualization*: Built-in visualization capabilities leverage the visualization package `keplergl` to interactively visualize data in Jupyter notebooks with simple code.
- *Basemap Loading*: Provides methods to display Mapbox basemaps in `matplotlib` figures (Figure 2).

The target audience of `TransBigData` includes: 1) Data science researchers and data engineers in the field of transportation big data, smart transportation systems, and urban computing, particularly those who want to integrate innovative algorithms into intelligent trasnportation systems; 2) Government, enterprises, or other entities who expect efficient and reliable management decision support through transportation spatio-temporal data analysis.

The latest stable release of the software can be installed via `pip` and full documentation can be found at https://transbigdata.readthedocs.io/en/latest/.

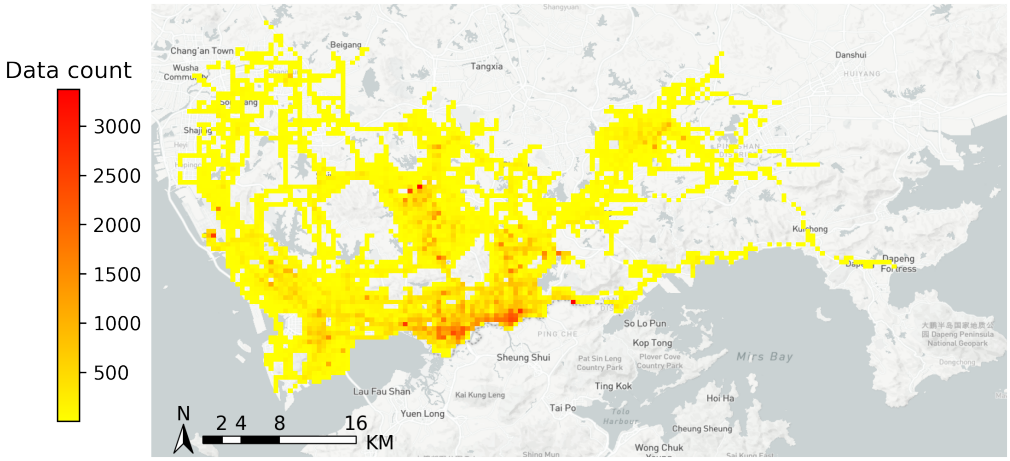

Figure 1: TransBigData generates rectangular grids and aggregates GPS data to the grids.

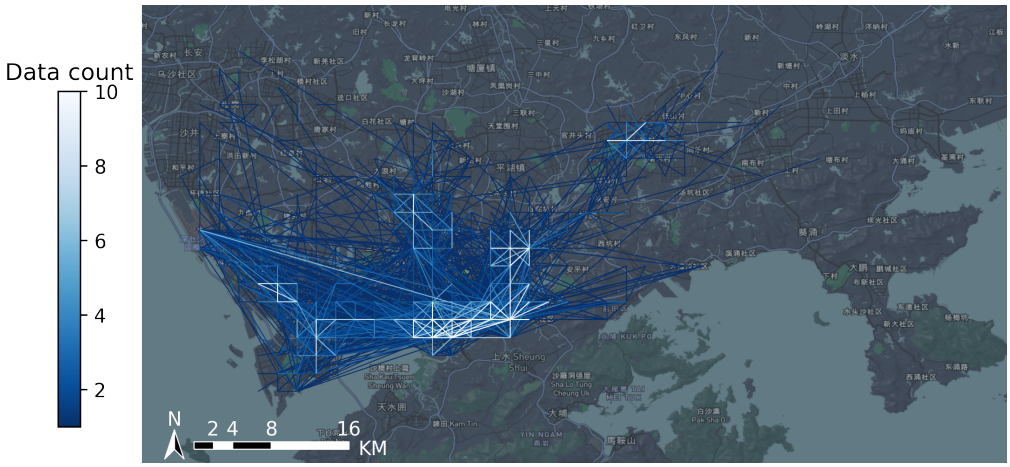

Figure 2: TransBigData visualizes taxi trip ODs and displays basemaps with matplotlib.

References

Axhausen, K. W. (2007). Definition of movement and activity for transport modelling. In *Handbook of transport modelling*. Emerald Group Publishing Limited. https://doi.org/10.1108/9780857245670-016

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Chen, J., Li, W., Zhang, H., Cai, Z., Sui, Y., Long, Y., Song, X., & Shibasaki, R. (2021). GPS data in urban online ride-hailing: A simulation method to evaluate impact of user scale on emission performance of system. *Journal of Cleaner Production*, *287*, 125567. https://doi.org/10.1016/j.jclepro.2020.125567

Graser, A. (2019). MovingPandas: Efficient Structures for Movement Data in Python. *GI_Forum – Journal of Geographic Information Science*, *1*, 54–68. https://doi.org/10.1553/giscience2019_01_s54

Haidri, S., Haranwala, Y. J., Bogorny, V., Renso, C., Fonseca, V. P. da, & Soares, A. (2021). *PTRAIL – a python package for parallel trajectory data preprocessing*. https://arxiv.org/abs/2108.13202

Harrison, G., Grant-Muller, S. M., & Hodgson, F. C. (2020). New and emerging data forms in transportation planning and policy: Opportunities and challenges for "Track and Trace" data. *Transportation Research Part C: Emerging Technologies*, *117*, 102672. https://doi.org/10.1016/j.trc.2020.102672

Li, Y., Li, W., Yu, Q., & Yang, H. (2021). Taxi global positioning system data in urban road network: A methodology to identify key road clusters based on travel speed–traffic volume correlation. *Transportation Research Record*, 03611981211036684. https://doi.org/10.1177/03611981211036684

Olive, X. (2019). Traffic, a toolbox for processing and analysing air traffic data. *Journal of Open Source Software*, *4*(39), 1518. https://doi.org/10.21105/joss.01518

Pappalardo, L., Simini, F., Barlacchi, G., & Pellungrini, R. (2019). *Scikit-mobility* (Version 1.1.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5528110

Rey, S. J., Anselin, L., Amaral, P., Arribas-Bel, D., Cortes, R. X., Gaboardi, J. D., Kang, W., Knaap, E., Li, Z., Lumnitz, S., Oshan, T. M., Shao, H., & Wolf, L. J. (2021). The PySAL Ecosystem: Philosophy and Implementation. *Geographical Analysis*. https://doi.org/10.1111/gean.12276

Rey, S., & Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, *37*(1), 5–27. https://doi.org/10.1007/978-3-642-03647-7_11

Schwalb-Willmann, J., Remelgado, R., Safi, K., & Wegmann, M. (2020). moveVis: Animating movement trajectories in synchronicity with static or temporally dynamic environmental data in r. *Methods in Ecology and Evolution*, *11*(5), 664–669. https://doi.org/doi.org/10.1111/2041-210X.13374

Siebinga, O. (2021). TraViA: A traffic data visualization and annotation tool in python. *Journal of Open Source Software*, *6*(65), 3607. https://doi.org/10.21105/joss.03607

Yu, Q., Li, W., Yang, D., & Zhang, H. (2021). Partitioning urban road network based on travel speed correlation. *International Journal of Transportation Science and Technology*, *10*(2), 97–109. https://doi.org/10.1016/j.ijtst.2021.01.002

Yu, Q., Zhang, H., Li, W., Song, X., Yang, D., & Shibasaki, R. (2020). Mobile phone GPS data in urban customized bus: Dynamic line design and emission reduction potentials analysis. *Journal of Cleaner Production*, *272*, 122471. https://doi.org/10.1016/j.jclepro.2020.122471

Yu, Q., Zhang, H., Li, W., Sui, Y., Song, X., Yang, D., Shibasaki, R., & Jiang, W. (2020). Mobile phone data in urban bicycle-sharing: Market-oriented sub-area division and spatial analysis on emission reduction potentials. *Journal of Cleaner Production*, *254*, 119974. https://doi.org/10.1016/j.jclepro.2020.119974

Zhang, H., Chen, J., Li, W., Song, X., & Shibasaki, R. (2020). Mobile phone GPS data in urban ride-sharing: An assessment method for emission reduction potential. *Applied Energy*, *269*, 115038. https://doi.org/10.1016/j.apenergy.2020.115038

Zhang, H., Song, X., & Shibasaki, R. (2021). *Big Data and Mobility as a Service*. Elsevier. https://doi.org/10.1016/c2020-0-02866-5