A TEMPLATE FOR THE *ARXIV* STYLE

A PREPRINT

H.Sherry Zhang
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia
huize.zhang@monash.edu

Dianne Cook
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia
dicook@monash.edu

Ursula Laa
Institute of Statistics
University of Natural Resources and Life Sciences
Vienna, Austria
ursula.laa@boku.ac.at

Nicolas Langrené
34 Village Street, Docklands VIC 3008 Australia
CSIRO Data61
Melbourne, Australia
nicolas.langrene@csiro.au

Patricia Menéndez
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia
patricia.menendez@monash.edu

August 9, 2021

Abstract

Enter the text of your abstract here.

***K*eywords** blah · blee · bloo · these are optional and can be removed

1 Introduction

Spatio-temporal data record changes of variables [for a reasonable length of period and spread across geographical region]. In this article, we consider spatio-temporal vector data, which are recorded in a fixed interval and are point based, characterised by longitude and latitude, in the spatial aspect. Examples of this type of data include the house price of a city or county, climate measures from weather stations in a country, and river level data from electronic gauges.

Analysing this type of data requires less considerations on the geographical geometry type and map projection but more on how measures in these fixed locations changes across the time domain and whether these changes are related for adjacent locations. For example, when nearby areas show patterns that are regular enough, visualising spatio-temporal data can 1) discover regional time series features, i.e. trend and seasonality, 2) find the Waldo sites from the crowd, and 3) see how correlation of nearby sites changes across time.

The main difficulty in visualising this type of data is to show information in both space and time dimension with the proper level of details without information overflow. This would sometimes require aggregating the time dimension into the proper level or slicing the data into a reasonable number of subset for display. In this sense, a data structure that regulates the manipulation spatio-temporal data will benefit the analysis workflow. While many implementations focus on manipulating and visualising pure spatial or temporal data, there are not sufficient tools to deal with spatio-temporal data. The purpose of this paper is to introduce a spatio-temporal vector data structure for data analysis in R.

The rest of the paper will be divided as follows: Section 2 reviews the existing data structure for spatio, temporal, and spatio-temporal data. Section 3 presents a new data structure for spatio-temporal data: cubble. Then the paper introduces the workflow of data manipulation and visualisation with the cubble structure in Section 4. Section 5 gives some examples on how common spatial and temporal manipulations are performed with cubble and how static and interactive visualisation help to understand climate and [...] data.

2 Existing data structure for spatio and temporal data

There has been a large class of implementations dedicated to processing the spatial data. This includes `sf` (E. J. Pebesma 2018) and its precedent `sp` (E. Pebesma and Bivand 2005) for ... and `raster` (Hijmans 2020) and `terra` (Hijmans 2021) for raster data. While these implementations specialised in geographic manipulations with different type of simple features, it doesn't incorporate a temporal dimension in the data structure. Project like `stars` (E. Pebesma 2021) and `spacetime` (Bivand, Pebesma, and Gomez-Rubio 2013) by R-Spatial allows for both space and time dimension for raster and vector data, but the underlying data structure is a multi-dimensional array, which could be difficult to operate for R users who are more familiar with the operation in 2D dataframe/ tibble.

In the temporal aspect, the `tsibble` (Wang, Cook, and Hyndman 2020) structure and its tidyverts ecosystem have provided a [...] workflow to work with temporal data. In a tsibble structure, temporal data is characterised by `index` and `key` where `index` is the temporal identifier and `key` is the identifier for multiple series, which could be used as a spatio identifier. However, a tsibble object, by construction, always requires the `index` in its structure. This makes it less appealing for spatio-temporal data since the output of calculated spatio-specific variables (i.e. features of each series) don't have the time dimension. Analysts will either need to have an additional step to join this output to the original tsibble or operate with variables stored in two separate objects. In addition, the long form structure of a tsibble object means spatio variables (i.e. longitude, latitude, and features of each series if joined back to the tsibble) of each spatio identifier will be repetitively recorded at each timestamp. This repetition is unnecessary and would inflate the object size for long series.

3 A new data structure for spatio-temporal data

Intro to cubble:

- list-column: rowwise_df with temporal variables, including the time index, nested.
  - Focus on spatio: those output per station
- long form: grouped_df
  - Focus on temporal

Compatible with tidyverse manipulation and tsibble

4 Manipulation and visualisation with `cubble`

Mention different types of manipulation with cubble:

- `dplyr` support for cubble:
  - basic 5s: mutate, filter, summarise, select, arrange
  - group and ungroup: group_by, ungroup
  - slice family
- summarise missing stats

5 Examples

Daily climate data (prcp, tmax, and tmin) from RNOAA - lots of stations across Australia

An exploratory data analysis questions: What's the climate profile look like in Australia

- General features: Any general trend/ fluctuation in prcp, tmax, and tmin?
- Local features: Any station stands out from the crowd?

5.1 Manipulation

- data quality check: filter out stations have variables not properly recorded
- data summary:
  - daily -> monthly/ weekly,
  - summarise by mean for tmax/ tmin, sum for prcp
-

5.2 Graphics

Static + interactive -> tooltip to show additional information upon hovering

- Where are those stations on the map?
  - Mention mostly aero, airport, and lighthouse

Summary

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. https://asdar-book.org/.

Hijmans, Robert J. 2020. *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster.

———. 2021. *Terra: Spatial Data Analysis*. https://CRAN.R-project.org/package=terra.

Pebesma, Edzer. 2021. *Stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*. https://CRAN.R-project.org/package=stars.

Pebesma, Edzer J. 2018. "Simple Features for r: Standardized Support for Spatial Vector Data." *R J.* 10 (1): 439.

Pebesma, Edzer, and Roger S Bivand. 2005. "S Classes and Methods for Spatial Data: The Sp Package." *R News* 5 (2): 9–13.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics* 29 (3): 466–78. https://doi.org/10.1080/10618600.2019.1695624.