

---

# 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 10, 2021

## Abstract

Enter the text of your abstract here.

**Keywords** 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

In spatio-temporal data, a variable is usually either group-related or time-related. While time-related variables record variables that change in time, group-related variables are invariant to the time and have only one value per group. A **cubble** simplifies the workflow in spatio-temporal data through managing group and time-related variables separately in two forms: list-column and long form. The list column form:

- defines each group in a row,
- displays the group-related variables in columns, and
- nests all the time-related variables into a column called **ts**.

In the long form,

- each combination of group and timestamp occupies a row
- time-related variables are displayed, and
- group-related variables are not explicitly displayed but can be accessed through the **meta** attribute

Below are the how the list-column and long form look like for Australia climate data, which records daily precipitation, maximum and minimum temperature for 55 stations across Australia from 2015- 2020.

```
## # Cubble: station-wise: list-column
## # Group: station [55]
## # Meta: station [fct], lat [dbl], long [dbl], elevation [dbl], name [fct]
## station lat long elevation name ts
## <fct> <dbl> <dbl> <dbl> <fct> <list>
## 1 ASN00001019 -14.3 127. 23 kalumburu <tbl_ts [2,147 x 4]>
## 2 ASN00002012 -18.2 128. 422 halls creek airport <tbl_ts [2,191 x 4]>
## 3 ASN00003003 -17.9 122. 7.4 broome airport <tbl_ts [2,192 x 4]>
## 4 ASN00006011 -24.9 114. 4 carnarvon airport <tbl_ts [2,192 x 4]>
## 5 ASN00009021 -31.9 116. 15.4 perth airport <tbl_ts [2,192 x 4]>
## 6 ASN00009193 -32.0 116. 43.1 rotnnest island <tbl_ts [2,192 x 4]>
## 7 ASN00009518 -34.4 115. 13 cape leeuwin <tbl_ts [2,184 x 4]>
## 8 ASN00009789 -33.8 122. 25 esperance <tbl_ts [2,187 x 4]>
## 9 ASN00010286 -31.6 117. 217. cunderdin airfield <tbl_ts [2,190 x 4]>
## 10 ASN00010917 -32.7 117. 275 wandering <tbl_ts [2,192 x 4]>
## # ... with 45 more rows

## # Cubble: time-wise: long form [tsibble]
## # Group: station [55]
## # Meta: station [fct], lat [dbl], long [dbl], elevation [dbl], name [fct]
## station date prcp tmax tmin
## <fct> <date> <dbl> <dbl> <dbl>
## 1 ASN00001019 2015-01-01 164 31.5 25
## 2 ASN00001019 2015-01-02 124 31.3 25
## 3 ASN00001019 2015-01-03 50 29.2 24.7
## 4 ASN00001019 2015-01-04 204 32.1 24.3
## 5 ASN00001019 2015-01-05 412 31.4 24.6
## 6 ASN00001019 2015-01-06 200 28.7 25.1
## 7 ASN00001019 2015-01-07 822 30.2 24.5
## 8 ASN00001019 2015-01-08 22 31 26.6
## 9 ASN00001019 2015-01-09 62 32.7 26.2
## 10 ASN00001019 2015-01-10 274 32.9 22.6
## # ... with 120,311 more rows
```

## 4 Manipulation and visualisation with cubble

### 4.1 Cubble verbs

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>.