Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

cubble: An R Package for Organizing and Wrangling Multivariate Spatio-temporal Data

H. Sherry Zhang
Monash University

Dianne Cook
Monash University

Ursula Laa
University of Natural Resources and Life Sciences

Nicolas Langrené
BNU-HKBU United International College

Patricia Menéndez
Monash University

Abstract

Multivariate spatio-temporal data refers to multiple measurements taken across space and time. For many analyses, spatial and time components can be separately studied: for example, to explore the temporal trend of one variable for a single spatial location, or to model the spatial distribution of one variable at a given time. However for some studies, it is important to analyze different aspects of the spatio-temporal data simultaneously, like for instance, temporal trends of multiple variables across locations. In order to facilitate the study of different portions or combinations of spatio-temporal data, we introduce a new class, `cubble`, with a suite of functions enabling easy slicing and dicing on the different components spatio-temporal components. The proposed `cubble` class ensures that all the components of the data are easy to access and manipulate while providing flexibility for data analysis. In addition, the **cubble** package facilitates visual and numerical explorations of the data while easing data wrangling and modelling. The `cubble` class and the functions provided in the **cubble** R package equip users with the capability to handle hierarchical spatial and temporal structures. The `cubble` class and the tools implemented in the package are illustrated with different examples of Australian climate data.

Keywords: spatial, temporal, spatio temporal, R, environmental data, exploratory data analysis.

1. Introduction

Spatio-temporal data has a spatial component referring to the location of each observation and a temporal component that is recorded at regular or irregular time intervals. It may also include multiple variables measured at each spatial and temporal values. With spatio-temporal data, one can fix the time to explore the spatial features of the data, fix the spatial location/s to explore temporal aspects, or dynamically explore the space and time simultaneously.

In order to computationally explore the spatial, temporal and spatio-temporal faces of such data, the data needs to be stored and represented under a specific data object that allows the user to query, group and dissect all the data faces.

The Comprehensive R Archive Network (CRAN) task view SpatioTemporal (Pebesma and Bivand 2022) gathers information about R packages designed for spatio-temporal data and it has a section on *Representing data* that lists existing spatio-temporal data representations used in R. Among them, Pebesma (2012) summarises spatio-temporal data into three forms: time-wide, space-wide, and long formats. The associated package **spacetime** (Pebesma 2012) implements four spatio-temporal layouts (full grid, sparse grid, irregular, and trajectory) to handle different space and time combinations. The package **stars** (Pebesma 2021) has a new implementation to use dense arrays to represent spatio-temporal cubes. It also interfaces with the package **sf** (Pebesma 2018), commonly used for wrangling spatial data, and the **tidyverse** (Wickham *et al.* 2019) suite for general data wrangling and visualization in R.

Still, the data representation for spatio-temporal data can be further extended and there are two reasons for this. Firstly, the raw data sourced in the wild is less often presented in any one of the layouts above, and fitting the raw data into a data object can sometimes be difficult. More often, spatio-temporal data are collected in separate 2D tables and analysts need to assemble them into a whole piece before exploring the data. Examples of components of spatio-temporal data can be 1) areal data recording the shape of a collection of areas of interest; 2) geostatistical data storing the longitude and latitude coordinates of locations, typically also with other metadata related to the location, and; 3) temporal data of each location across time.

The other reason is about tidy data concepts (Wickham 2014) and how they should be applied to spatio-temporal data. According to the tidy data principles, data should be structured into 1) one row per observation, 2) one column per variable, and 3) one type of data per table. The long form data is preferred over wide data form given the downstream packages such as **dplyr** (Wickham *et al.* 2022) and **ggplot2** (Wickham 2016) for data wrangling and visualization. However, the long form can be inefficient to store feature geometries, especially for large multipolygons for hourly, daily or sub-daily periods over years, which are extensively collected and handled, for example in time series analysis. This poses the question of how to arrange spatial and temporal variables in a way that would make data wrangling, visualizing and analyzing spatio-temporal data easier.

This paper presents a new R package, **cubble**, which addresses the two issues mentioned above. In the package, a new class, also called cubble, is proposed to organize spatial and temporal variables as two forms of a single data object so that they can be wrangled separately or combined while being kept synchronized. Among the four spacetime layouts in Pebesma (2012), the `cubble` class can be applied to full grid, sparse grid, or irregular, but not trajectory, which is outside the scope of this work. The software is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=cubble.

The rest of the paper is organized as follows: Section 2 introduces the proposed cube structure as a way to conceptualize multivariate spatio-temporal data. Section 3 presents the main design and functionality of the **cubble** package. Section 4 explains how the **cubble** package deals with more advanced considerations, including data with hierarchical structure, data matching and how the package fits with existing static and interactive visualization tools. Moreover we also illustrate how the **cubble** package deals with spatio-temporal data transformations. Section 5 uses Australian weather station data and river level data as examples to demonstrate the use of the package. An example of how the **cubble** package handles Network Common Data Form (NetCDF) data is also provided. Section 6 discuss the paper contributions and future directions.

2. Conceptual framework: spatio-temporal cube

Spatio-temporal data can be conceptualized using a cubical data model with three axes which typically are, time, latitude and longitude. This abstraction can be useful for generalizing operations and visualization purposes: Lu *et al.* (2018) shows how array operations (select, scale, reduce, rearrange, and compute) can be mapped onto the cube; Bach *et al.* (2014) reviews the temporal data visualization based on space-time cube operations. Notice that the term space-time cube in their article "does not need to involve spatial data", but refers to "an abstract 2D substrate that is used to visualize data at a specific time". Despite its main focus being on temporal data, the mindset of abstracting out data representation to construct visualizations, still applies to our spatio-temporal data manipulation and visualization approaches.

The most common space-time cube uses the three axes, time, latitude, longitude, and can be considered stacking space across time. Ours is a multivariate spatio-temporal cube with the three axes defined to be time, site and variables, as illustrated in the leftmost column of Figure 1. The time axis is the same in both versions, while the site axis now captures both latitude and longitude. Finally, variables are stacked on this space-time canvas, with one observation per site and time point. This notion is adopted to avoid using hyper-cubes when describing multivariate spatio-temporal data and is the conceptual framework behind the `cubble` class. With this conceptual model, operations on spatio-temporal data can be mapped to operations on the cube and the rest of Figure 1 show examples of slicing on site, time, and variable.

While the data cube model is conceptually convenient for spatio-temporal data, a 3D data array is not sufficiently rich for data wrangling, for several reasons. Although arrays can be efficient for the computation on numerical values, spatio-temporal data typically includes various types of variables. For example, character strings and specific datetime classes are common. In addition, it will be generally useful to be able to create new variables which is trickier to manage in an array. Thus for convenient wrangling, we have opted to create a special `cubble` class.

3. The cubble package

The `cubble` class is an S3 class (Wickham 2019) built on the `tibble` class, specifically to organize spatio-temporal data. The `cubble` class uses an attribute "form", to arrange the

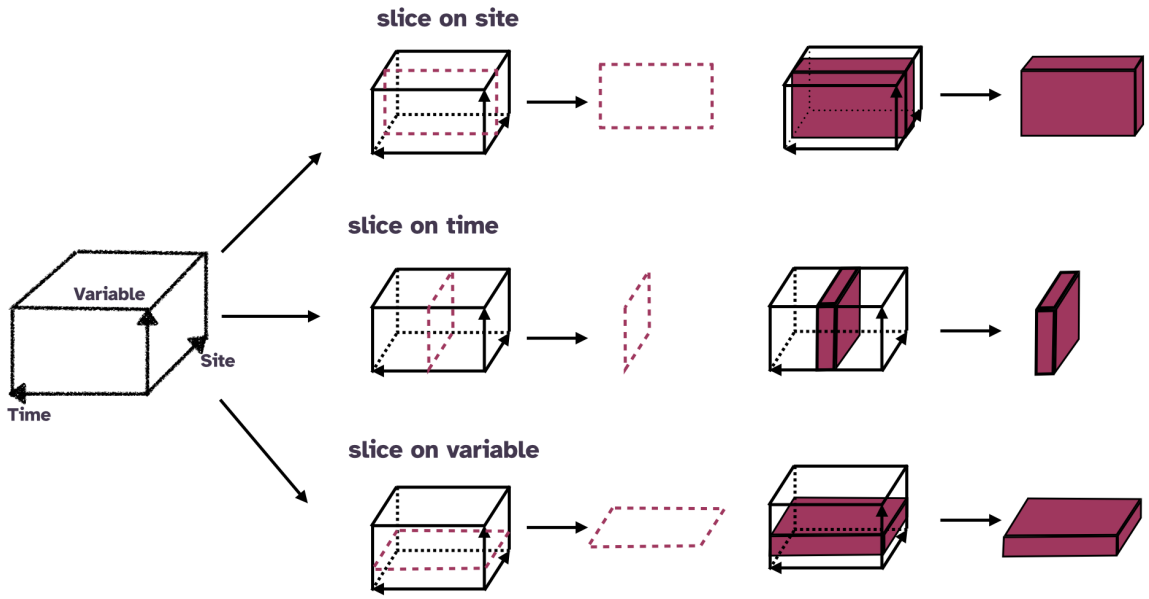

Figure 1: An illustration of the conceptual spatio-temporal cube with different slicing on time, site, and variable. For each axis, the slicing can be on a single value or a set of values.

spatial or temporal data components tidily. The form attribute can take a value of either "nested" or "long". The nested `cubble` is a subclass of rowwise `tibble` (`rowwise_df`). It arranges each spatial site in a row, and uses list columns to store the feature geometry and the temporal information. The long `cubble` is a subclass of grouped `tibble` (`grouped_df`), which expands the temporal information into the long form and stores the spatial information in a "spatial" attribute.

The main functions in the package are `as_cubble()`, `face_spatial()`, `face_temporal()`, and `unfold()`. The following sections explain their roles, why the new `cubble` class is needed and how the package relates to existing packages for spatial and temporal data analysis.

The data `climate_flat` is used to illustrate functionality in the **cubble** package. This is a subset from National Oceanic and Atmospheric Administration (NOAA) (NOAA 2022) Global Historical Climatology Network (GHCN) Daily data. It contains spatial variables, station id, latitude, longitude, elevation, station name, world meteorology organisation id, in additional to daily temporal information, maximum and minimum temperature values and precipitation records for year 2020. The first five rows of the data are shown below:

```
# A tibble: 1,830 x 10
  id          lat  long  elev name  wmo_id date        prcp  tmax  tmin
  <chr>     <dbl> <dbl> <dbl> <chr>  <dbl> <date>     <dbl> <dbl> <dbl>
1 ASN0000~ -31.9  116.  15.4 pert~  94610 2020-01-01     0  31.9  15.3
2 ASN0000~ -31.9  116.  15.4 pert~  94610 2020-01-02     0  24.9  16.4
3 ASN0000~ -31.9  116.  15.4 pert~  94610 2020-01-03     6  23.2  13
4 ASN0000~ -31.9  116.  15.4 pert~  94610 2020-01-04     0  28.4  12.4
5 ASN0000~ -31.9  116.  15.4 pert~  94610 2020-01-05     0  35.3  11.6
```

```
# ... with 1,825 more rows
```

3.1. Create a cubble

The function `as_cubble()` is used to create a `cubble` object with three arguments: `key` as the spatial identifier; `index` as the temporal identifier; and a vector of `coords` in the order (longitude, latitude). The arguments `key` and `index` follow the wording in the **tsibble** package to describe the temporal order and multiple series while `coords` specifies the spatial location of each site. The code below creates a `cubble` object out of `climate_flat` (a single `tibble`) with `id` as the key, `date` as the index, and `c(long, lat)` as the coordinates:

```
R> cubble_nested <- climate_flat |>
+   as_cubble(key = id, index = date, coords = c(long, lat))
R> cubble_nested

# cubble:   id [5]: nested form
# bbox:     [115.97, -32.94, 133.55, -12.42]
# temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl]
  id            lat  long  elev name           wmo_id ts
  <chr>       <dbl> <dbl> <dbl> <chr>           <dbl> <list>
1 ASN00009021 -31.9  116.  15.4 perth airport   94610 <tibble>
2 ASN00010311 -31.9  117. 179   york            94623 <tibble>
3 ASN00010614 -32.9  117. 338   narrogin        94627 <tibble>
4 ASN00014015 -12.4  131.  30.4 darwin airport  94120 <tibble>
5 ASN00015131 -17.6  134. 220   elliott         94236 <tibble>
```

Printing a `cubble` object provides some information about the data. Here `id` is the variable name to identify each location and there are five unique locations. The bounding box is `[115.97, -32.94, 133.55, -12.42]` and provides information about the coordinates in the data. The third row shows the name and type of all variables nested in the `ts` column. In this example, it includes `date [date], prcp [dbl], tmax [dbl], tmin [dbl]`.

The created `cubble` is a subclass of the `rowwise_df` class where each row forms a group. All the temporal variables are nested in a list column, hence it is also called the nested `cubble`. The rowwise structure makes it simpler to operate on the list using the `mutate()` syntax, which is simpler than the `purr::map()` when working with a list column. For example, calculating the number of rainy days can be done by:

```
R> cubble_nested |>
+   mutate(rain_day = sum(ts$prcp != 0))

# cubble:   id [5]: nested form
# bbox:     [115.97, -32.94, 133.55, -12.42]
# temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl]
  id            lat  long  elev name           wmo_id ts       rain_day
  <chr>       <dbl> <dbl> <dbl> <chr>           <dbl> <list>      <int>
1 ASN00009021 -31.9  116.  15.4 perth airport   94610 <tibble>      104
```

```
2 ASN00010311 -31.9  117. 179   york           94623 <tibble>       89
3 ASN00010614 -32.9  117. 338   narrogin       94627 <tibble>       90
4 ASN00014015 -12.4  131.  30.4 darwin airpo~  94120 <tibble>      106
5 ASN00015131 -17.6  134. 220   elliott        94236 <tibble>       63
```

A `cubble` class can be created from various common spatio-temporal data formats, including `tibble`, `tsibble`, and `sf` when both spatial and tmeporal information are available. Section 5.1 describes converting multiple tables into a `cubble` object and Section 3.6.4 illustrates how to convert a NetCDF object.

3.2. Change focus by facing the time-variables

The nested form can be used for those operations where the output is only indexed by the spatial identifier (`key`), but becomes inadequate when outputs need both a spatial and a temporal identifier (`key` and `index`). The `cubble` class also provides a long form, which expands the `ts` column and temporarily "hides" the spatial variables. The function `face_temporal()` is used to switch from the nested `cubble` into the long one. The first row in Figure 2 illustrates this operation where the focus of the cube now changes from the site-variable face to the time-variable face. This code switches the `cubble` object just created into its long form:

```
R> cubble_long <- cubble_nested |> face_temporal()
R> cubble_long

# cubble:  date, id [5]: long form
# bbox:    [115.97, -32.94, 133.55, -12.42]
# spatial: lat [dbl], long [dbl], elev [dbl], name [chr], wmo_id
#   [dbl]
  id          date        prcp  tmax  tmin
  <chr>       <date>     <dbl> <dbl> <dbl>
1 ASN00009021 2020-01-01     0  31.9  15.3
2 ASN00009021 2020-01-02     0  24.9  16.4
3 ASN00009021 2020-01-03     6  23.2  13
4 ASN00009021 2020-01-04     0  28.4  12.4
5 ASN00009021 2020-01-05     0  35.3  11.6
# ... with 1,825 more rows
```

The first line in the header now shows it in the long form and the third line has been changed to display the name and type of spatial variables: `lat [dbl], long [dbl], elev [dbl], name [chr], wmo_id [dbl]`. Unlike the nested form, the long `cubble` is built from a `grouped_df` class where all the observations from the same site form a group.

3.3. Change focus back to the site-variable face

Wrangling spatio-temporal data can be seen as an iterative process in the spatial and temporal dimensions. Switching the focus back to the site-variable face can be accomplished by the function `face_spatial()`, which is the inverse of `face_temporal()`. The second row of Figure 2 illustrates the function, which is used as follows:

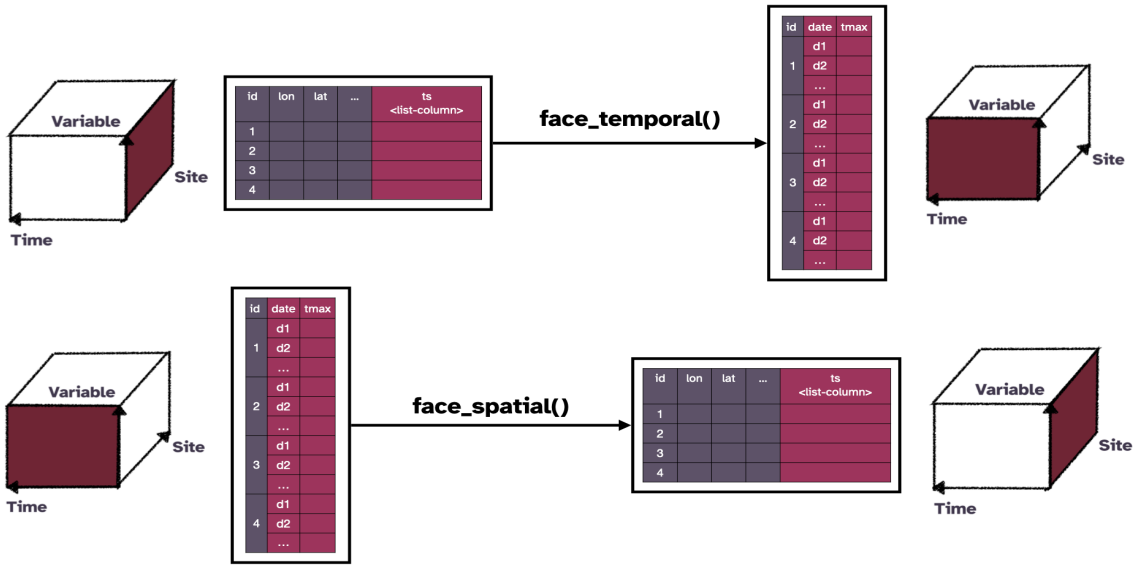

Figure 2: An illustration of function `face_temporal` and `face_spatial`. In the first row, `face_temporal` switches a `cubble` object from the nested form into the long form and the focus has switched from the spatial aspect (the side face) to the temporal aspect (the front face). In the second row, `face_spatial` switches a `cubble` object back to the nested form from the long form and shifts focus back to the spatial aspect.

```
R> cubble_back <- cubble_long |> face_spatial()
R> cubble_back

# cubble:   id [5]: nested form
# bbox:     [115.97, -32.94, 133.55, -12.42]
# temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl]
  id            lat  long  elev name           wmo_id ts
  <chr>       <dbl> <dbl> <dbl> <chr>           <dbl> <list>
1 ASN00009021 -31.9  116.  15.4 perth airport   94610 <tibble>
2 ASN00010311 -31.9  117. 179   york            94623 <tibble>
3 ASN00010614 -32.9  117. 338   narrogin        94627 <tibble>
4 ASN00014015 -12.4  131.  30.4 darwin airport  94120 <tibble>
5 ASN00015131 -17.6  134. 220   elliott         94236 <tibble>

R> identical(cubble_nested, cubble_back)

[1] TRUE
```

3.4. Unfold spatial variables into the long form

Sometimes, analysts may need to apply some variable transformation that involves both the spatial and temporal variables. An example of this is the transformation of temporal variables

into the spatial dimension in glyph maps (Wickham *et al.* 2012). (How to make glyph maps will be explained in Section 4.4, and are illustrated in the second example.) This type of operation can be seen as flattening, or *unfolding*, the cube into a 2D data frame. Here the function `unfold()` moves the spatial variables `long` and `lat` into the long `cubble`:

```
R> cubble_unfold <- cubble_long |> unfold(long, lat)
R> cubble_unfold

# cubble:  date, id [5]: long form
# bbox:    [115.97, -32.94, 133.55, -12.42]
# spatial: lat [dbl], long [dbl], elev [dbl], name [chr], wmo_id
#   [dbl]
  id          date        prcp  tmax  tmin  long   lat
  <chr>       <date>     <dbl> <dbl> <dbl> <dbl> <dbl>
1 ASN00009021 2020-01-01     0  31.9  15.3  116. -31.9
2 ASN00009021 2020-01-02     0  24.9  16.4  116. -31.9
3 ASN00009021 2020-01-03     6  23.2  13    116. -31.9
4 ASN00009021 2020-01-04     0  28.4  12.4  116. -31.9
5 ASN00009021 2020-01-05     0  35.3  11.6  116. -31.9
# ... with 1,825 more rows
```

3.5. Why not just use the existing tidyverse functions

Some readers may question why a new data structure is needed rather than directly creating a list-column on the combined data using `dplyr::nest_by()`. The reason is that the `cubble` object is specifically designed to utilize the spatio-temporal structure when arranging observations in a single object. Moreover, it enables easy pivoting between purely spatial, purely temporal, or unfolded into a combined form.

3.6. Compatibility with existing packages

The **cubble** package leverages tools available in existing packages used for spatial and temporal analysis, specifically, **dplyr**, **tsibble**, **sf** (**s2**), and **ncdf4**, as explained here.

`dplyr`

The **dplyr** package has many tools for wrangling tidy data, many of which are useful in the spatio-temporal analysis. The **cubble** package provides methods that support the use of the following operations in the **dplyr** package on both the nested and long forms: `mutate`, `filter`, `summarise`, `select`, `arrange`, `rename`, `left_join`, and the slice family (`slice_*`).

`tsibble`

The `tsibble` class is a subclass of `tibble` where the `index` and `key` components are used to store temporal and strata information, that makes working with temporal data cognitively efficient. A `cubble` object can use the `tsibble` class to store the temporal information, and effectively utilize the specialist time series operations in the **tsibble** package. A `tsibble`

object can also be casted into a `cubble` object through supplying the coordinate information in the argument `coords`:

```
R> climate_flat_ts <- climate_flat |>
+   tsibble::as_tsibble(key = id, index = date)
R> climate_flat_cb <-  climate_flat_ts |>
+   cubble::as_cubble(coords = c(long, lat))
R> climate_flat_cb

# cubble:   id [5]: nested form
# bbox:     [115.97, -32.94, 133.55, -12.42]
# temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl]
  id            lat  long  elev name           wmo_id ts
  <chr>       <dbl> <dbl> <dbl> <chr>           <dbl> <list>
1 ASN00009021 -31.9  116.  15.4 perth airport   94610 <tbl_ts>
2 ASN00010311 -31.9  117. 179   york            94623 <tbl_ts>
3 ASN00010614 -32.9  117. 338   narrogin        94627 <tbl_ts>
4 ASN00014015 -12.4  131.  30.4 darwin airport  94120 <tbl_ts>
5 ASN00015131 -17.6  134. 220   elliott         94236 <tbl_ts>
```

When a nested `cubble` is created, each element in the list-column `ts` is in the `tsibble` class (labelled `tbl_ts`) and operations available to the `tsibble` class are still valid on these elements. For example, the code below calculates two time series features (mean and variance) of maximum temperature, utilizing the `tsibble` syntax in the `cubble` object:

```
R> climate_flat_cb |>
+   mutate(fabletools::features(
+     ts, tmax, list(tmax_mean = mean, tmax_var = var)
+     ))

# cubble:   id [5]: nested form
# bbox:     [115.97, -32.94, 133.55, -12.42]
# temporal: date [date], prcp [dbl], tmax [dbl], tmin [dbl]
  id          lat  long  elev name  wmo_id ts       tmax_mean tmax_var
  <chr>     <dbl> <dbl> <dbl> <chr>  <dbl> <list>       <dbl>    <dbl>
1 ASN00009~ -31.9  116.  15.4 pert~  94610 <tbl_ts>      25.7    38.6
2 ASN00010~ -31.9  117. 179   york   94623 <tbl_ts>      26.2    51.1
3 ASN00010~ -32.9  117. 338   narr~  94627 <tbl_ts>      23.7    45.4
4 ASN00014~ -12.4  131.  30.4 darw~  94120 <tbl_ts>      33.1     3.02
5 ASN00015~ -17.6  134. 220   elli~  94236 <tbl_ts>      34.6    24.7
```

`sf` *(*`s2`*)*

The `sf` class is also a subclass of `tibble` with a specialized feature geometry list-column (`sfc`) to store different geometry types (`POINT`, `LINESTRING`, `POLYGON`, `MULTIPOLYGON`, etc). The package **sf** provides functions that operate efficiently on this spatial information. A

`cubble` object can store spatial information in the `sf` class. Methods for the `sfc` class can be applied in the nested form of the cubble object. An illustration is in Section 5.1. The spatial information can also be stored as an `s2` vector in a `cubble` object.

`ncdf4`

The NetCDF data is another format commonly used for storing spatio-temporal data. It has two main components: *dimension* for defining the spatio-temporal grid (longitude, latitude, and time) and *variable* that populates the defined grid. Attributes can be associated with dimensions or variables. Because there can be many different styles of representing this information there is a metadata convention (Hassell *et al.* 2017) to standardize the format of the attributes. A few packages in R exist for manipulating NetCDF data and these include a high-level R interface: **ncdf4** (Pierce 2019), a low-level interface that calls a C-interface: **RNetCDF** (Michna and Woods 2021), and a tidyverse implementation: **tidync** (Sumner 2020).

The **cubble** package provides an `as_cubble()` method to coerce the `ncdf4` class from the **ncdf4** package into a `cubble` object. It maps each combination of longitude and latitude into an `id` as the `key`:

```
R> raw <- ncdf4::nc_open(here::here("data/era5-pressure.nc"))
R> dt <- as_cubble(raw, vars = c("q", "z"))
```

Sometimes NetCDF data can be quite large, and it is best to subset the data when converting to a `cubble` object. We would recommend reducing to about 300 × 300 grid points for three daily variables in one year. A 300 by 300 spatial grid can be a bounding box of [100, -80, 180, 0] at 0.25 degree resolution or a global bounding box [-180, -90, 180, -90] at 1 degree resolution. The size of spatial grid can be reduced if longer time periods or more variables are needed, through the arguments `long_range` and `lat_range`:

```
R> dt <- as_cubble(
+   my_ncdf, vars = c("q", "z"),
+   long_range = seq(-180, 180, 1), lat_range = seq(-90, 90, 1)
+   )
```

4. Other features and considerations

4.1. Hierarchical structure

Spatial locations can have grouping structures either inherent to the data e.g., state within country or obtained during the analysis e.g., cluster id. In this case, it can be useful to summarize variables at various levels of the hierarchy. The function `switch_key()` can be used to change the grouping level of spatial locations. The diagram in Figure 3 shows how this function can be used to switch the grouping from station ids to cluster ids. The result can also be stretched into long form. By specifying `cluster_nested <- station_nested %>% switch_key(key = cluster)`, the `cubble` object redefines its `key` from the `id` column in `station_nested` to the `cluster` column in `cluster_nested`. All the spatial variables

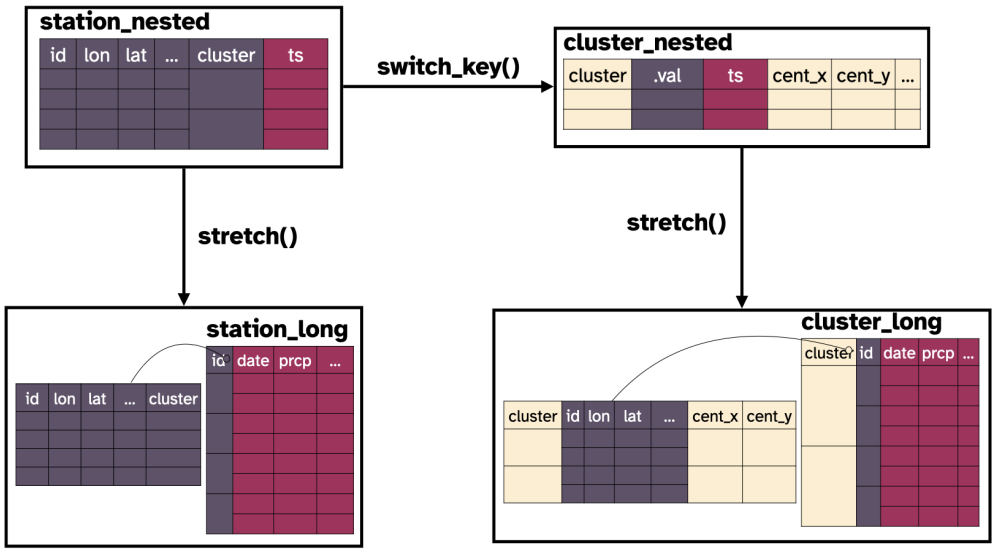

Figure 3: Hierarchical spatial structure can be handled using `switch_key()`, to create summaries based on any level. Here the switch is between the station id and a cluster id. Once the change is made the data can be stretched into the long form.

belonging to the `cluster` column are now nested into a `.val` column which allows for summarizing based on cluster.

4.2. Data fusion and matching

One task that may interest spatio-temporal analysts is combining data collected at nearby but not exactly the same sites, for example, weather station measured rainfall and river levels. This can be considered to be a matching problem (Stuart 2010; McIntosh *et al.* 2018) to pair similar time series from nearby locations, or even a data fusion exercise that merges data collected from different sources (Cocchi 2019). The function `match_sites()` in the **cubble** package provides a simple algorithm for this task. The algorithm first matches spatially by computing the pairwise distance on latitude and longitude. Then it matches temporally by computing the number of matched peaks within a fixed length moving window. Figure 4 illustrates this temporal matching. In the two series, A and a, three peaks have been identified in each. An interval, of fixed length, is constructed for each peak in series A, while the peaks in series a are tested against whether they fall into any of the intervals. Here two out of three peaks match. Options for `match_sites()` are:

- `spatial_n_keep`: the number of spatial match for each site to keep;
- `spatial_dist_max`: the maximum distance allowed for a matched pair;
- `temporal_n_highest`: the number of peaks used - 3 in the example above;
- `temporal_window`: the length of the interval - 5 in the example above; and
- `temporal_min_match`: the minimum number of matched peaks for a valid matched pair.

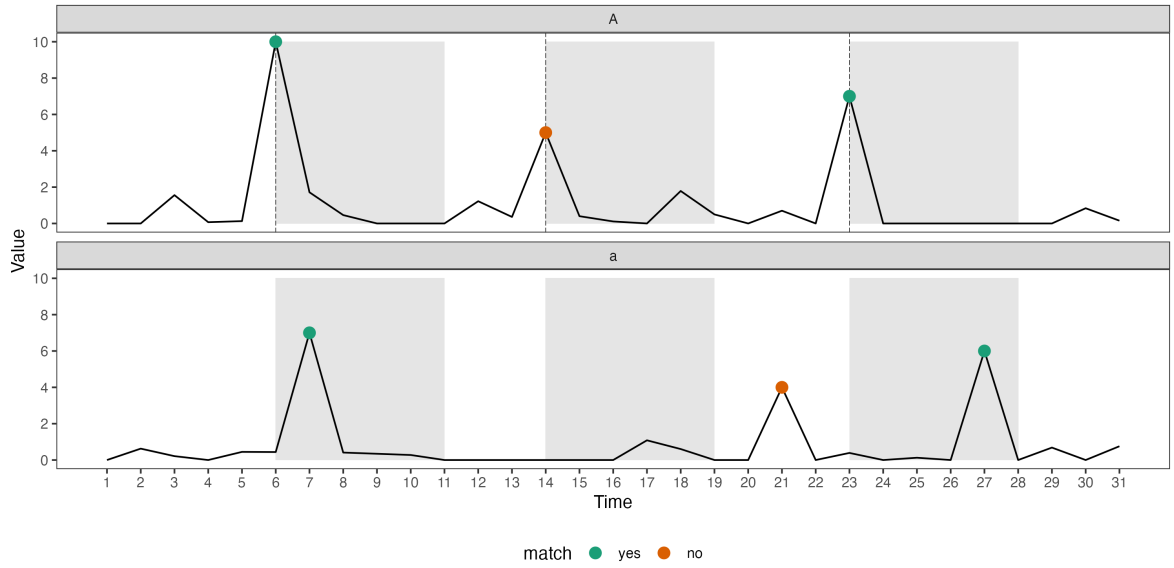

Figure 4: An illustration of temporal matching in the **cubble** package. Three highest peaks are identified in each series and intervals are constructed on series `A`. Two peaks in series `a` fall into the intervals and hence the two series are considered to have two matches.

4.3. Interactive graphics

The workflow with the `cubble` class fits works well with an interactive graphics pipeline (e.g., Buja *et al.* (1988), Buja *et al.* (1996), Sutherland *et al.* (2000), Xie *et al.* (2014), Cheng *et al.* (2016)) that is available in R with the package **crosstalk** (Cheng and Sievert 2021). Figure 5 illustrates how linking can be achieved between a map and multiple time series in a `cubble` object. The map (produced from the nested form) and time series (produced from the long form) are both shared `crosstalk` objects. When a user makes a selection on the map, the site is highlighted (a). This activates a row in the nested `cubble`, which is then communicated to the long `cubble` – all the observations with the same id (b) will be selected. The long `cubble` will then highlight the corresponding series in the time series plot (c).

Linking is also available starting from the time series plot, by selecting points. This will be activate rows having the same id in the long `cubble` The corresponding rows in the nested `cubble` are activated, and highlighted the map. (An illustration can be found in the appendix.) Note that this type of linking, both from the map or the time series, is what Cook and Swayne (2007) would call categorical variable linking, where station id is the categorical variable.

4.4. Spatio-temporal transformations

Spatio-temporal data lends itself to a range of transformations. Glyph maps (Section 3.4) transform the measured variable and time coordinates into microplots at the spatial locations. Calendar plots (Wang *et al.* 2020a) deconstruct time to produce plots of variables in a calendar format. Summarizing multiple variables is commonly done using projections, or linear combinations. Here we elaborate on the transformations made to produce a glyph map.

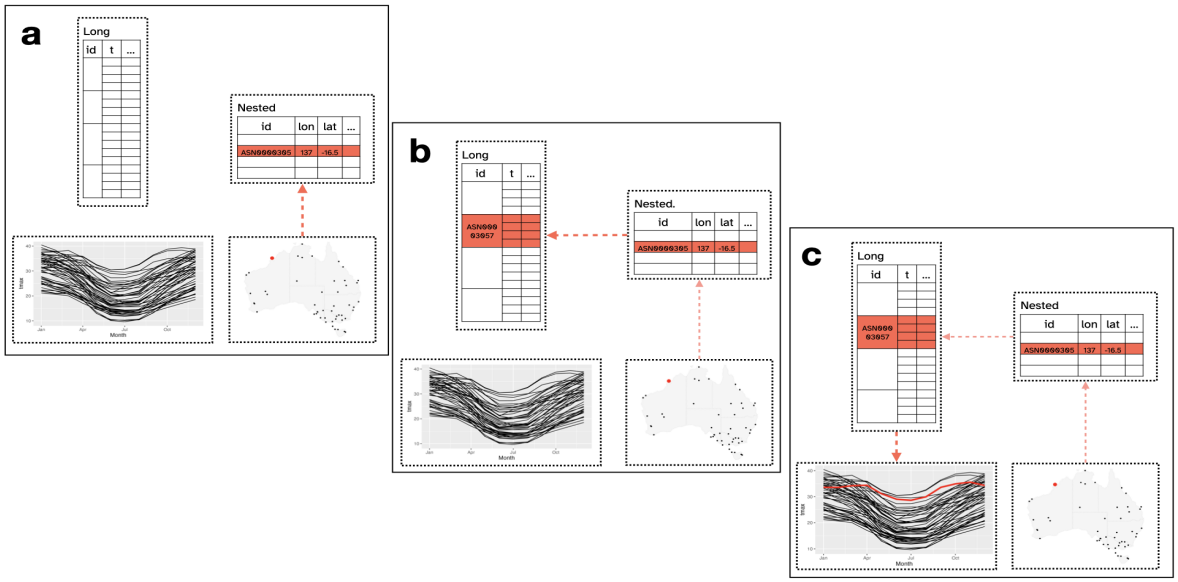

Figure 5: Linking between multiple plots. The line plots and the map are constructed from shared `crosstalk` objects (long and nested `cubble`). When a station is selected on the map (a), the corresponding row in the nested `cubble` will be activated. This will link to all the rows with the same id in the long `cubble` (b) and update the line plot (c).

The package **GGally** (Schloerke *et al.* 2021) has implemented glyph maps through the `glyphs()` function. The function constructs a `data.frame` with calculated position (`gx`, `gy`, `gid`) of each point on the time series using linear algebra (Equations 1 and 2 in Wickham *et al.* (2012)). The data can then be piped into ggplot to create the glyph map as:

```
R> library("ggplot2")
R> gly <- glyphs(data,
+                x_major = ..., x_minor = ...,
+                y_major = ..., y_minor = ..., ...)
R>
R> ggplot(gly, aes(gx, gy, group = gid)) +
+   geom_path()
```

A new implementation of the glyph map as a ggproto, `GeomGlyph`, has been made in the **cubble** package so that a glyph map can be created with `geom_glyph()`:

```
R> ggplot(data = data) +
+   geom_glyph(aes(x_major = ..., x_minor = ...,
+                  y_major = ..., y_minor = ...))
```

An example using a glyph map is shown in Section 5.2.

Some useful controls over the glyph map are also available in the `geom_glyph()` implementation. Polar coordinate glyph maps are specified using `polar = TRUE`, and arguments `width` and `height` can be specified in either absolute or relative value. Global and local scale is

specified with `global_rescale`, which defaults to `TRUE`. Reference boxes and lines can be added with separate `geom_glyph_box()` and `geom_glyph_line()` lines.

5. Examples

The five examples here are chosen to illustrate these aspects of the **cubble** package: creating a `cubble` object from two Coronavirus (COVID) data tables with the complication of differing location names, using spatial transformations to make a glyph map of seasonal temperature changes over years, aggregating information spatially to explore precipitation patterns, matching river level data and weather station records for analysis of water supply, reading NetCDF format data to reproduce a climate reanalysis plot, and the workflow to create complex interactive linked plots. (There is an additional example and figures in the Appendix, and more examples in the package vignettes.)

5.1. Victoria COVID spatio-temporal incidence and spread

Since the start of the pandemic, the Victoria State Government in Australia has provided daily COVID counts by local government area (LGA). This data can be used to visualize COVID incidence and spread spatially, when combined with map polygon data available from the Australian Bureau of Statistics. These different sources need to be combined for the analysis, by matching the LGA names. Here is how to do this with the **cubble** package, including how to handle mismatches arising from different names of the same LGAs in the two tables. The COVID data is store in a csv file and looks like:

```
R> covid |> head(5)
```

```
# A tsibble: 5 x 5 [1D]
# Key:       lga [1]
# Groups:    lga, source [1]
  date       lga        source                            n roll_mean
  <date>     <chr>      <chr>                         <int>     <dbl>
1 2022-01-01 Alpine (S) Contact with a confirmed case     1        NA
2 2022-01-02 Alpine (S) Contact with a confirmed case     2        NA
3 2022-01-03 Alpine (S) Contact with a confirmed case     4        NA
4 2022-01-04 Alpine (S) Contact with a confirmed case     4        NA
5 2022-01-05 Alpine (S) Contact with a confirmed case     2        NA
```

and the spatial polygons are an ESRI shapefile as follows:

```
R> lga |> head(5)
```

```
Simple feature collection with 5 features and 7 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 142.3535 ymin: -38.67876 xmax: 147.3909 ymax: -36.39269
Geodetic CRS:  WGS 84
```

```
    lga_code_2018            lga state_code_2016 state_name_2016
132         20110     Alpine (S)               2        Victoria
133         20260    Ararat (RC)               2        Victoria
134         20570   Ballarat (C)               2        Victoria
135         20660    Banyule (C)               2        Victoria
136         20740 Bass Coast (S)               2        Victoria
    areasqkm_2018 cent_long  cent_lat                       geometry
132     4788.1568  146.9742 -36.85357 MULTIPOLYGON (((146.7258 -3...
133     4211.1171  142.8432 -37.47271 MULTIPOLYGON (((143.1807 -3...
134      739.0321  143.7815 -37.49286 MULTIPOLYGON (((143.6622 -3...
135       62.5402  145.0851 -37.73043 MULTIPOLYGON (((145.1357 -3...
136      865.8095  145.5581 -38.50730 MULTIPOLYGON (((145.5207 -3...
```

The function `as_cubble()` is used to create a `cubble` object from the two spatial and temporal tables, and requires specifying the arguments `key`, `index`, and `coords` (as described in Section 3.1). It will automatically try to match the sites in both tables using the location names and will show a warning message when there are mismatches, as shown below:

```
R> cb <- as_cubble(
+   list(spatial = lga, temporal = covid),
+   key = lga, index = date, coords = c(cent_long, cent_lat)
+   )

! Some sites in the temporal table don't have corresponding spatial information

! Some sites in the spatial table don't have corresponding temporal information

! Use argument `output = "unmatch"` to check on the unmatched key
```

It can be seen that there are two-way mismatches – LGAs in the COVID data that do not match with LGAs names in the spatial polygon data, and vice versa. The mismatches can be identified by using the `output = "unmatch"` argument in the `as_cubble()` function:

```
R> pair <- as_cubble(
+   list(spatial = lga, temporal = covid),
+   key = lga, index = date, coords = c(cent_long, cent_lat),
+   output = "unmatch"
+   )
R>
R> pair

$paired
# A tibble: 2 x 2
  spatial             temporal
  <chr>               <chr>
1 Kingston (C) (Vic.) Kingston (C)
```

```
2 Latrobe (C) (Vic.)  Latrobe (C)

$others
$others$temporal
[1] "Interstate" "Overseas"   "Unknown"

$others$spatial
[1] "No usual address (Vic.)"
[2] "Migratory - Offshore - Shipping (Vic.)"
```

With this information both tables can be fixed, to create the desired `cubble` object, as follows:

```
R> lga <- lga |>
+   mutate(lga = ifelse(lga == "Kingston (C) (Vic.)", "Kingston (C)", lga),
+          lga = ifelse(lga == "Latrobe (C) (Vic.)", "Latrobe (C)", lga)) |>
+   filter(!lga %in% pair$others$spatial)
R>
R> covid <- covid |> filter(!lga %in% pair$others$temporal)
R>
R> cb <- as_cubble(data = list(spatial = lga, temporal = covid),
+                  key = lga, index = date, coords = c(cent_long, cent_lat))
```

5.2. Australian historical maximum temperature

The GHCN provides daily climate measures from stations across the world. The data used here (`historical_tmax`) is a subset extracted using the package **rnoaa** (Chamberlain 2021), containing the records of maximum temperature for 236 Australian stations from 1859 through 1969 and provides information also on the latitude, longitude and elevation of each of the stations. The goal of this example is to compare the monthly average maximum temperature between two periods, 1971-1975 and 2016-2020, for stations in Victoria and New South Wales (NSW), using a glyph map.

First, the stations need to be filtered to those in Victoria and NSW by using the station identifiers, stored within the 11 digits of the `id` variable entries. The country code is in the first 5 digits (Australia is represented by "ASN00") and the next 6 digits encode the station following the Australian Bureau of Meteorology (BOM) (Commonwealth of Australia 2022) coding protocols. The NSW stations correspond to entries in the range 46-75 and the Victorian stations to 76-90. Filtering Victoria and NSW stations is a *spatial operation* and hence uses the nested `cubble`:

```
R> tmax <- historical_tmax |>
+   filter(between(stringr::str_sub(id, 7, 8), 46, 90))
```

Next, the monthly maximum average temperature is calculated for both periods. This is a *temporal operation* requiring a switch into the long `cubble` using the `face_temporal()` function. In addition, a new indicator for the two time periods of interest is created before the calculation of monthly averages:

```
R> tmax <- tmax |>
+   face_temporal() |>
+   group_by(month = lubridate::month(date),
+          group = as.factor(
+            ifelse(lubridate::year(date) > 2015,
+            "2016 ~ 2020", "1971 ~ 1975"))) |>
+   summarise(tmax = mean(tmax, na.rm = TRUE))
```

A quick check on the number of observations for each location is made, revealing that there are several with less than 24 observations – these stations lack temperature values for some months. In this example, those stations are removed by switching to a long `cubble` to operate on the spatial component over time, and then, move back into the nested `cubble` (to make the glyph map):

```
R> tmax |>
+   face_spatial() |>
+   mutate(n = nrow(ts)) |>
+   arrange(n) |>
+   pull(n) |>
+   head(10)

 [1] 12 12 12 13 19 19 20 24 24 24

R> tmax <- tmax |>
+   face_spatial() |>
+   filter(nrow(ts) == 24) |>
+   face_temporal()
```

In order to create a glyph map displaying the monthly series (Figure 6), the spatial variables need to be unfolded with the temporal variables. The reason being that the major (`long`, `lat`) and minor (`month`, `tmax`) coordinates need to be on the same table to create the glyph map. The `geom_glyph()` function does both the transformation and the plotting.

```
R> nsw_vic <- ozmaps::abs_ste |>
+   filter(NAME %in% c("Victoria", "New South Wales"))
R>
R> ggplot() +
+   geom_sf(data = nsw_vic,
+           fill = "transparent", color = "grey",
+           linetype = "dotted") +
+   geom_glyph(data = tmax,
+              aes(x_major = long, x_minor = month,
+                  y_major = lat, y_minor = tmax,
+                  group = interaction(id, group), color = group),
+              width = 1, height = 0.5) +
+   ...
```

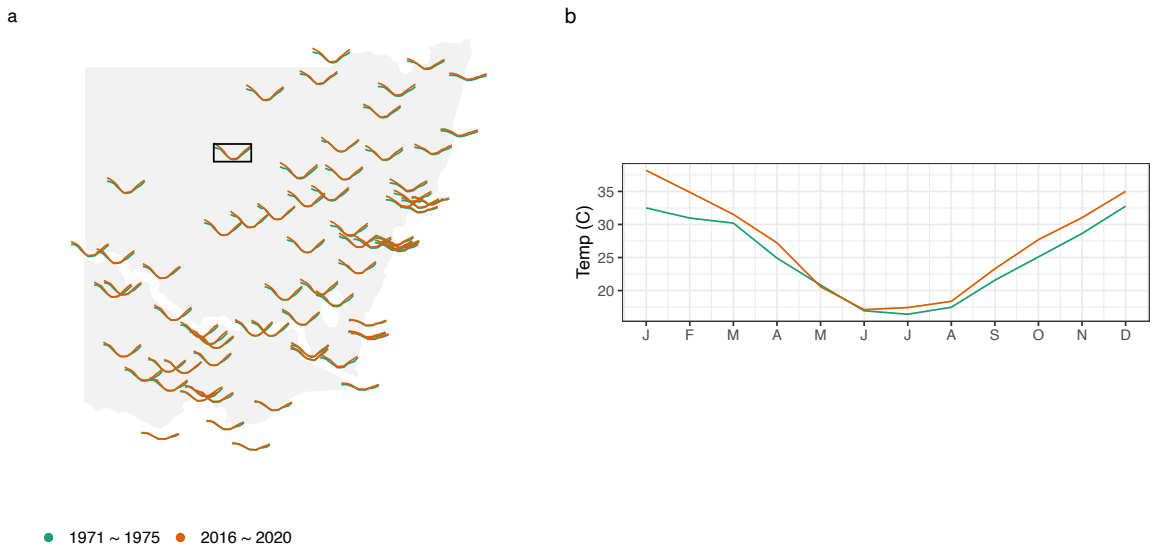

Figure 6: A glyph map of the monthly maximum average temperature for weather stations in Victoria and New South Wales (NSW) for the periods (1971-1975, 2016-2020). The corresponding average time series for the cobar station are display on the top left corner. From the glyph map we can observe that the monthly trend is similar for all locations (low in the winter, high in the summer), and small increased temperatures, particularly in late summer can be seen at most stations in NSW.

Glyph maps work well to explore temporal patterns across spatial locations, particularly when the spatial locations are gridded. In this example, they are irregularly spaced, which can result in overlapping glyphs obscuring each other. To fix this, one could aggregate data from nearby stations. An example of this using is included in the Appendix.

5.3. River levels and rainfall in Victoria

River level and rainfall data for the same areas should have some similarity. Here we examine the river gauge data (`Water_course_level`) from the Bureau of Meteorology (Commonwealth of Australia 2022) in relation to weather station rainfall from NOAA's climate data (`climate`). The goal is to match water gauges with nearby weather stations, spatially and temporally, using 2020 measurements using `match_sites()` function.

This function requires passing the major and minor data sets used for matching, in this case those are `river` and `climate`. The variables used for the temporal matching are `Water_course_level` from the `river` data set and `prcp` in the climate data set. The rest of the arguments, as explained in Section 4.2, correspond to the maximum and minimum number of peaks in the time series to be matched. In this example those are set to be a maximum of 30 and a minimum of 15 (approximately 2 and 1 per month).

```
R> res <- match_sites(
+   river, climate,
+   temporal_by = c("Water_course_level" = "prcp"),
```

```
+   temporal_independent = "prcp",
+   temporal_n_highest = 30,
+   temporal_min_match = 15,
+ )
```

This function returns a `cubble` object, with additional columns: `dist` storing the distance between matched stations, `group` summarizing spatial matching, and `n_match` showing the temporal matching.

```
# cubble:   id [8]: nested form
# bbox:     [144.52, -37.73, 146.06, -36.55]
# temporal: date [date], matched_var [dbl]
  id          name        lat  long type   dist group ts       n_match
  <chr>       <chr>     <dbl> <dbl> <chr> <dbl> <int> <list>     <int>
1 405234      SEVEN CR~ -36.9  146. river  6.15     5 <tibble>      21
2 404207      HOLLAND ~ -36.6  146. river  8.54    10 <tibble>      21
3 ASN00082042 strathbo~ -36.8  146. clim~  6.15     5 <tibble>      21
4 ASN00082170 benalla ~ -36.6  146. clim~  8.54    10 <tibble>      21
5 230200      MARIBYRN~ -37.7  145. river  6.17     6 <tibble>      19
# ... with 3 more rows
```

Figure 7 shows four matched pairs on the map (a) and standardized data as time series (b). The expected concurrent increase in precipitation and water level can be seen clearly.

5.4. ERA5: climate reanalysis data

Figure 8 reproduces the ERA5 data row of Figure 19 in (Hersbach *et al.* 2020). Here we explain how this would be done using in the **cubble** package. The plots show that the southern polar vortex splits into two on 2002-09-26 and further splits into four on 2002-10-04. Further explanation of why this is interesting can be found in the figure source, and also (Simmons *et al.* 2020), (Simmons *et al.* 2005).

The ERA5 data (Hersbach *et al.* 2020) provides hourly estimates across the Earth for atmospheric, land and oceanic climate variables. The data is available in the NetCDF format from the European Centre for Medium-Range Weather Forecasts (ECMWF). It can be directly downloaded from Copernicus Climate Data Store (CDS) (Copernicus Climate Change Service 2022) website or via the **ecmwfr** package (Hufkens *et al.* 2019). For the reproduction, we focus on the `era5-pressure` data, hourly pressure levels from 1970 to present, with the *specific humidity* and *geopotential*. The downloaded NetCDF data is read into R using the **ncdf4** package, and converted to a `cubble` object:

```
R> raw <- ncdf4::nc_open(here::here("data/era5-pressure.nc"))
R> dt <- as_cubble(
+   raw, vars = c("q", "z"),
+   long_range = seq(-180, 180, 1), lat_range = seq(-88, 88, 1))
```

Creating the plot requires making transformations on time, unfolding the data for computing the statistic of interest, which is plotted directly as a contour plot with ggplot. Code is provided to accomplish this in the supplementary material.

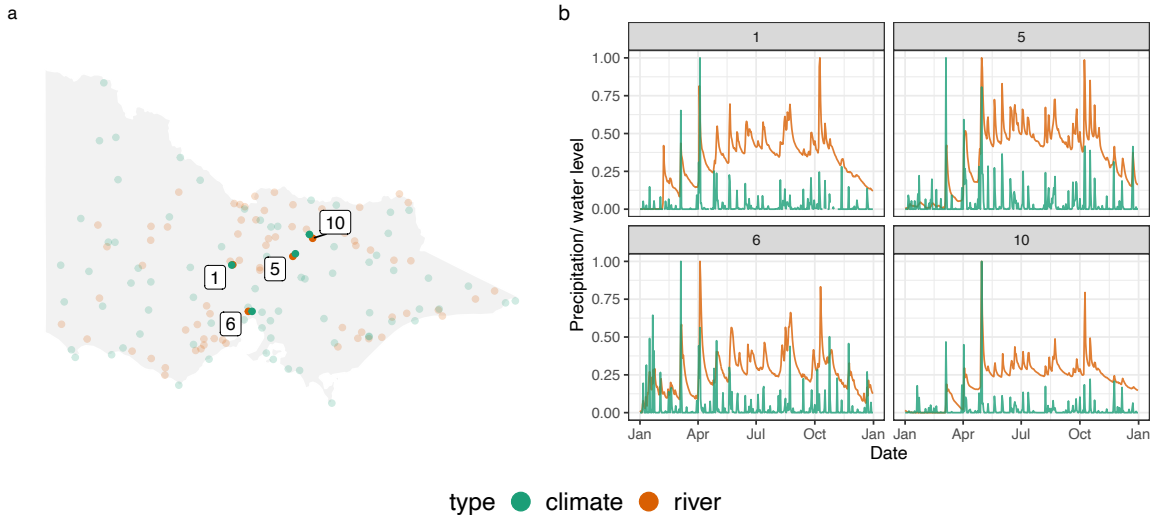

Figure 7: Weather stations and river gauges with matched pairs labelled on the map (a) and plotted across time (b). Precipitation and water level have been standardised between 0 and 1 to be displayed on the same scale. The water level reflects the increase in precipitation. The numbers (1, 5, 6, 10) indicate the group index derived from spatial matching, only those that were selectd by temporal matching are shown here.

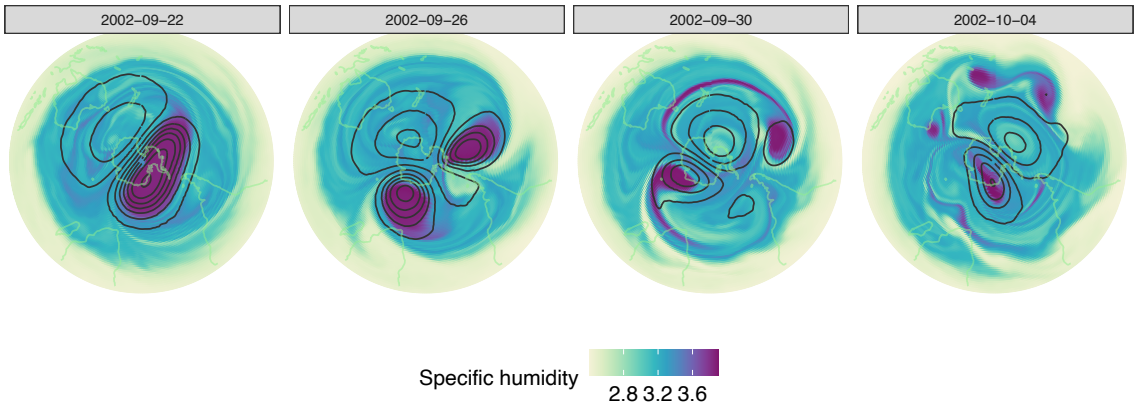

Figure 8: A reproduction of the second row (ERA5 data) of Figure 19 in Hersbach et al (2020) to illustrate the break-up of sourthern polar vortex in late September and early October 2002. The polar vortex, signalled by the high speicfic humidity, splits into two on 2002-09-26 and furthers split into four on 2002-10-04.

5.5. Interactive graphics

Interactive graphics can be useful because they make it possible to look at the data in a multiple of ways on-the-fly. This is especially important for spatio-temporal data, where we would like to interactively connect spatial and temporal displays. This example describes the process of using the **cubble** package with the **crosstalk** package to build an interactive display connecting a map of Australia, with ribbon plots of temperature range observed at the stations. The purpose is to explore the variation of monthly temperature range over the country. Figure 9 shows three snapshots of the interactivity.

The key steps are to convert both the nested and long forms of the data into shared `crosstalk` objects, and to plot these side-by-side. The two are linked by the station identifier.

```
clean <- climate_full |> ...

nested <- clean |> SharedData$new(~id, group = "cubble")
long <- face_temporal(clean) |> SharedData$new(~id, group = "cubble")

p1 <- nested |> ...
p2 <- long |> ...

crosstalk::bscols(plotly::ggplotly(p1), plotly::ggplotly(p2), ...)
```

Plot (a) shows the initial state of the interactive display: all locations are shown as dots on the map, coloured by temperature range, and the right plot shows the ribbons representing maximum to minimum for all stations. In plot (b) the "Mount Elizabeth" station, which shows a high variance colour on the initial map, is selected on the map and this produces the ribbon on the right. In plot (c) the lowest temperature in August is selected, which is "Thredbo" station on the left map. It was surprising to us that this was not a station in Tasmania, so for comparison a station in Tasmania is selected on the map to show in relation to Thredbo. We can see that Thredbo has a bigger winter dip in temperature, and although Tasmania is cold generally, it's temperatures are more consistent.

6. Conclusion

This paper presents an R package **cubble** for organizing, manipulating and visualizing spatio-temporal data. The package introduces a new data class for spatio-temporal data, `cubble`, that connects the time invariant and varying variables and that allows the user to work with a nested and long form of the data. This work adds capabilities into the spatio-temporal practitioners toolbox to integrate it with a tidy data framework. The data structure and functions introduced in this package can be used and combined with existing spatial data analysis packages such as **sf**, data wrangling packages such as **dplyr**, and visualization packages such as **ggplot2**, **plotly** and **leaflet**.

Numerous examples are provided in the main text, appendix and package vignettes. These include creating and coercing data with mismatched names, handling hierarchical data, matching time series spatially and temporally, reproducing ERA5 plots from NetCDF data. Visualization of the `cubble` objects can be done with interactive graphic pipelines using **crosstalk**, **plotly** and **leaflet**.

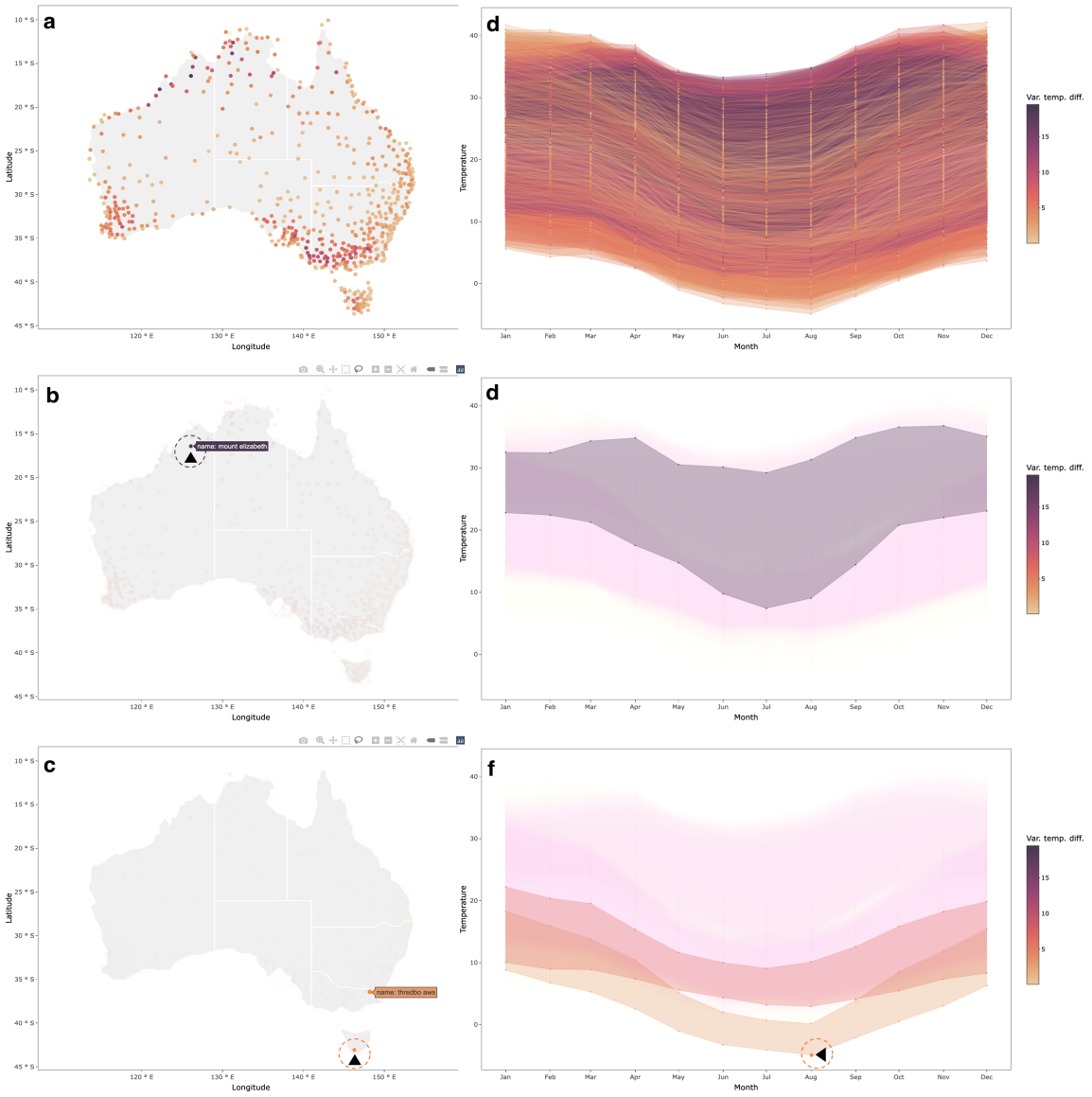

Figure 9: Exploring temperature variation using linking of a map and seasonal display. Each row is a screen dump of the process. The top row shows all locations and all temperature profiles. Selecting a particular location on the map (here Mount Elizabeth) produces the plot in the second row. The maximum and minimum temperatures are shown using a ribbon. The bottom row first selects the lowest temperature in August in the seasonal display, which highlights the corresponding station on the map (Thredbo). Another station, located in the Tasmania Island, is then selected to compare its temperature variation with the Thredbo station.

There are several possible future directions for the work. The data structure only described fixed spatial sites, and it could be useful to provide tools to accommodate moving coordinates as might be encountered in animal movement data. That could be achieved with a list-column for the location coordinates, and an additional form that these locations can be pivoted into, like the long form for temporal variables. For multiple measured variables, the **cubble** package could be integrated with dimension reduction methods, and dynamic graphics for multiple dimensions such as a tour (Wickham *et al.* 2011).

7. Acknowledgement

This work is funded by a Commonwealth Scientific and Industrial Research Organisation (CSIRO) Data61 Scholarship and started while Nicolas Langrené was affiliated with CSIRO's Data61. The article is created using the package **knitr** (Xie 2015) and **rmarkdown** (Xie *et al.* 2018) in R with the `rticles::jss_article` template. The source code for reproducing this paper can be found at: https://github.com/huizezhang-sherry/paper-cubble.

References

Bach B, Dragicevic P, Archambault D, Hurter C, Carpendale S (2014). "A Review of Temporal Data Visualizations Based on Space-Time Cube Operations." *Eurographics conference on visualization*, p. 19. URL https://hal.inria.fr/hal-01006140/.

Buja A, Asimov D, Hurley C (1988). "Elements of A Viewing Pipeline." *Dynamic Graphics Statistics*, p. 277.

Buja A, Cook D, Swayne DF (1996). "Interactive High-dimensional Data Visualization." *Journal of Computational and Graphical Statistics*, **5**(1), 78–99. URL https://doi.org/10.2307/1390754.

Castanedo F (2013). "A Review of Data Fusion Techniques." *The Scientific World Journal*, **2013**.

Chamberlain S (2021). ***rnoaa**: 'NOAA' Weather Data from R*. R package version 1.3.8, URL https://CRAN.R-project.org/package=rnoaa.

Cheng J, Karambelkar B, Xie Y (2021). ***leaflet**: Create Interactive Web Maps with the JavaScript **leaflet** Library*. R package version 2.0.4.1, URL https://CRAN.R-project.org/package=leaflet.

Cheng J, Sievert C (2021). ***crosstalk**: Inter-Widget Interactivity for HTML Widgets*. R package version 1.1.1, URL https://CRAN.R-project.org/package=crosstalk.

Cheng X, Cook D, Hofmann H (2016). "Enabling Interactivity on Displays of Multivariate Time Series and Longitudinal Data." *Journal of Computational and Graphical Statistics*, **25**(4), 1057–1076. URL https://doi.org/10.1080/10618600.2015.1105749.

Cocchi M (2019). *Data Fusion Methodology and Applications*. Elsevier.

Commonwealth of Australia (2022). "Australia's Official Weather Forecasts & Weather Radar - Bureau of Meteorology." Online; accessed 24 June 2022, URL http://www.bom.gov.au/.

Cook D, Swayne D (2007). *Interactive and Dynamic Graphics for Data Analysis with examples using R and GGobi*. Springer-Verlag New York, New York. With contributions from Buja, A., Temple Lang, D., Hofmann, H., Wickham, H. and Lawrence, M. and additional data, R code and demo movies at http://www.ggobi.org.

Copernicus Climate Change Service (2022). "Climate Data Store." Online; accessed 24 June 2022, URL https://cds.climate.copernicus.eu/#!/home.

Hassell D, Gregory J, Blower J, Lawrence BN, Taylor KE (2017). "A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1)." *Geoscientific Model Development*, **10**(12), 4619–4646. ISSN 1991-959X. doi:10.5194/gmd-10-4619-2017. URL https://gmd.copernicus.org/articles/10/4619/2017/.

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, *et al.* (2020). "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society*, **146**(730), 1999–2049.

Hufkens K, Stauffer R, Campitelli E (2019). "The **ecwmfr** Package: An Interface to ECMWF API Endpoints." URL https://bluegreen-labs.github.io/ecmwfr/.

Lu M, Appel M, Pebesma E (2018). "Multidimensional Arrays for Analysing Geoscientific Data." *ISPRS International Journal of Geo-Information*, **7**(8), 313. ISSN 2220-9964. doi:10.3390/ijgi7080313. URL http://www.mdpi.com/2220-9964/7/8/313.

McIntosh AI, Jenkins HE, White LF, Barnard M, Thomson DR, Dolby T, Simpson J, Streicher EM, Kleinman MB, Ragan EJ, *et al.* (2018). "Using Routinely Collected Laboratory Data to Identify High Rifampicin-Resistant Tuberculosis Burden Communities in the Western Cape Province, South Africa: A Retrospective Spatiotemporal Analysis." *PLoS Medicine*, **15**(8).

Michna P, Woods M (2021). ***RNetCDF**: Interface to 'NetCDF' Datasets*. R package version 2.5-2, URL https://CRAN.R-project.org/package=RNetCDF.

Müller K, Wickham H (2021). ***tibble**: Simple Data Frames*. R package version 3.1.6, URL https://CRAN.R-project.org/package=tibble.

NOAA (2022). "National Centers for Environmental Information (NCEI)." Online; accessed 24 June 2022, URL https://www.ncei.noaa.gov/.

Pebesma E (2012). "**spacetime**: Spatio-Temporal Data in R." *Journal of Statistical Software*, **51**(7), 1–30. URL https://doi.org/10.18637/jss.v051.i07.

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

Pebesma E (2021). ***stars**: Spatiotemporal Arrays, Raster and Vector Data Cubes*. R package version 0.5-2, URL https://CRAN.R-project.org/package=stars.

Pebesma E, Bivand R (2022). "CRAN Task View: Handling and Analyzing Spatio-Temporal Data." Version 2022-03-07, URL https://CRAN.R-project.org/view=SpatioTemporal.

Pierce D (2019). ***ncdf4**: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files*. R package version 1.17, URL https://CRAN.R-project.org/package=ncdf4.

Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J (2021). ***GGally**: Extension to **ggplot2***. R package version 2.1.2, URL https://CRAN.R-project.org/package=GGally.

Simmons A, Hortal M, Kelly G, McNally A, Untch A, Uppala S (2005). "ECMWF Analyses and Forecasts of Stratospheric Winter Polar Vortex Breakup: September 2002 in the Southern Hemisphere and Related Events." *Journal of the Atmospheric Sciences*, **62**(3), 668 – 689. doi:10.1175/JAS-3322.1. URL https://journals.ametsoc.org/view/journals/atsc/62/3/jas-3322.1.xml.

Simmons A, Soci C, Nicolas J, Bell B, Berrisford P, Dragani R, Flemming J, Haimberger L, Healy S, Hersbach H, Horányi A, Inness A, Munoz-Sabater J, Radu R, Schepers D (2020). "Global Stratospheric Temperature Bias and Other Stratospheric Aspects of ERA5 and ERA5.1." (859). doi:10.21957/rcxqfmg0. URL https://www.ecmwf.int/node/19362.

Stuart EA (2010). "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, **25**(1), 1.

Sumner M (2020). ***tidync**: A Tidy Approach to NetCDF Data Exploration and Extraction*. R package version 0.2.4, URL https://CRAN.R-project.org/package=tidync.

Sutherland P, Rossini A, Lumley T, Lewin-Koh N, Dickerson J, Cox Z, Cook D (2000). "**Orca**: A Visualization Toolkit for High-dimensional Data." *Journal of Computational and Graphical Statistics*, **9**(3), 509–529. URL https://www.tandfonline.com/doi/abs/10.1080/10618600.2000.10474896.

Wang E, Cook D, Hyndman RJ (2020a). "Calendar-based Graphics for Visualizing People's Daily Schedules." *Journal of Computational and Graphical Statistics*, **29**(3), 490–502.

Wang E, Cook D, Hyndman RJ (2020b). "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics*, **29**(3), 466–478. URL https://doi.org/10.1080/10618600.2019.1695624.

Wickham H (2014). "Tidy Data." *Journal of Statistical Software*, **59**(10), 1–23. URL https://doi.org/10.18637/jss.v059.i10.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wickham H (2019). *Advanced R*. CRC press. ISBN 978-0815384571. URL https://adv-r.hadley.nz/rcpp.html.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H

(2019). "Welcome to the **tidyverse**." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

Wickham H, Cook D, Hofmann H, Buja A (2011). "**tourr**: An R Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software*, **40**(2). ISSN 1548-7660. URL http://www.jstatsoft.org/v40/i02/.

Wickham H, François R, Henry L, Müller K (2022). ***dplyr**: A Grammar of Data Manipulation*. R package version 1.0.8, URL https://CRAN.R-project.org/package=dplyr.

Wickham H, Hofmann H, Wickham C, Cook D (2012). "Glyph-Maps for Visually Exploring Temporal Patterns in Climate Data and Models." *Environmetrics*, **23**(5), 382–393. doi:10.1002/env.2152.

Xie Y (2015). *Dynamic Documents with R and **knitr***. 2nd edition. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963, URL https://yihui.name/knitr/.

Xie Y, Allaire J, Grolemund G (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138359338, URL https://bookdown.org/yihui/rmarkdown.

Xie Y, Hofmann H, Cheng X (2014). "Reactive Programming for Interactive Graphics." *Statistical Science*, **29**(2), 201 – 213. URL https://doi.org/10.1214/14-STS477.

Affiliation:

H. Sherry Zhang
Monash University
21 Chancellors Walk, Clayton VIC 3800 Australia
E-mail: huize.zhang@monash.edu

Dianne Cook
Monash University
21 Chancellors Walk, Clayton VIC 3800 Australia
E-mail: dicook@monash.edu

Ursula Laa
University of Natural Resources and Life Sciences
Gregor-Mendel-Straße 33, 1180 Wien, Austria
E-mail: ursula.laa@boku.ac.at

Nicolas Langrené
BNU-HKBU United International College
2000 Jintong Road, Tangjiawan, Zhuhai, Guangdong Province, China
E-mail: nicolaslangrene@uic.edu.cn

Patricia Menéndez
Monash University
21 Chancellors Walk, Clayton VIC 3800 Australia
E-mail: patricia.menendez@monash.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

Appendix to 'cubble: An R Package for Organizing and Wrangling Multivariate Spatio-temporal Data'

Sherry Zhang, Dianne Cook, Ursula Laa, Nicolas Langrené, Patricia Menéndez

2022-06-14

This is the supplementary material for the main paper, containing an extended example following example 5.2 to highlight how **cubble** can be used to deal with data that has a hierarchical structure. It also describes the process to create linking plots. Furthermore, this appendix contains additional information about the data sources and the necessary code for extracting and preparing the data sets used in the paper with the goal to ensure reproducibility.

1 Extension of example 5.2: Australian precipitation pattern in 2020

In example 5.2, some overlapping of the glyphs occurred for a number of stations on the right hand side of the map in Figure 6. This is a problem when mapping time series or other glyphs corresponding to locations that are geographically closed on the map. In some cases it is better to display such information at an aggregated level by grouping the data adequately before exhibiting the information on a figure. The example below shows how can spatio-temporal data be organised at different hierarchical levels so that information can be grouped both temporal and spatially using `switch_key()`. The goal of this example is to highlight how easy is to move across the data hierarchy using **cubble**.

The data `climate_full`, also extracted from the GHCN, records daily precipitation and maximum/minimum temperature for 640 stations in Australia from 2016 to 2020. A simple k-means algorithm based on the distance matrix between stations is used to create 20 clusters. The data `station_nested` is a nested cubble with a cluster column indicating the group to which each station belongs. More advanced clustering algorithms can be used as well, as long as they provide a mapping from each station to a cluster.

```
station_nested <- climate_full %>% mutate(cluster = ...)
```

To create a group-level cubble, use `switch_key()` with the new key variable, `cluster`:

```
cluster_nested <- station_nested %>% switch_key(cluster)
```

With the group-level cubble, `get_centroid()` is useful to compute the centroid of each cluster, which will be used as the major axis for the glyph map later:

```
cluster_nested <- cluster_nested %>% get_centroid()
```

Long form cubble at both levels can be accessed through stretching the nested form. With access to both station and cluster-level cubbles, various plots can be made to understand the cluster. Figure 1 shows two example plots that can be made with this data. Subplot A is a glyph map made with the cluster level cubble in the long form and subplot B inspects the station membership of each cluster using the station level cubble in the nested form.

Figure 1: Profile of aggregated precipitation at 639 weather stations in Australia. Subplot A shows the glyph map of the weekly averaged precipitation of each cluster. The group number is printed in the middle of the y minor axis and can be used as a reference line to read the magnitude. Subplot B shows the station membership of each cluster.

2 Extension to section 5.5: Interactive graphics

Figure 2 in this section can also be made using `cubble` and `leaflet`, in which case the temperature range is displayed as a small subplot upon clicking on the map. This procedure involves first creating the popup plots from the long form cubble as a vector and then adding these plots to a leaflet map created from the nested cubble, with `leafpop::addPopupGraphs()`:

```
# data pre-processing
clean <- climate_full %>% ...

# use the long form to create subplots for each station
df_id <- unique(clean$id)
p <- map(1:length(df_id), function(i){
  dt <- clean %>% filter(id == df_id[i])
  ggplot(dt) %>% ...
})

# create nested form leaflet map with temperature band as subplots
nested <- face_spatial(clean)
leaflet(nested) |>
  addTiles() |>
  addCircleMarkers(group = "a", ...) |>
  leafpop::addPopupGraphs(graph = p, ...)
```

Figure 2 shows the same information as Figure 9 but using leaflet and popups.

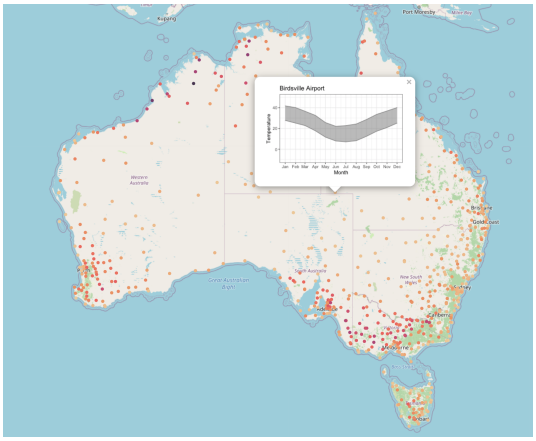

Figure 2: Same as Figure 11 except the temperature variation is now shown as a popup in the leaflet map.

3 Additional illustration on multiple linked plots

This figure is a supplement to Section 4.3 of the main paper, illustrating how linking from the time series plot to the map is achieved.

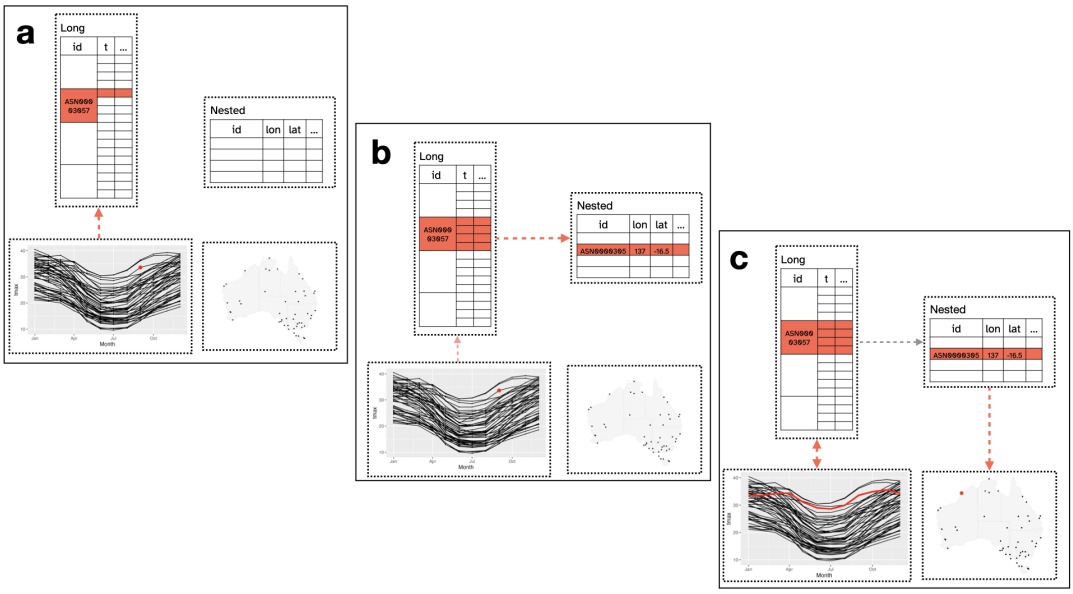

Figure 3: Linking between multiple plots. The line plots and the map are constructed from shared crosstalk objects (long and nested cubbles). When a point on the time series is selected, the corresponding row in the long cubble will be activated (a). This will link to all the rows with the same id in the long cubble and the row in the nested cubble with the same id (b). Both plots will be updated with the full line selected and the point highlighted on the map (c).

4 Scripts for creating the example data

This section contains the codes for extracting the data for the examples discussed in the main manuscript.

4.1 Historical maximum temperature

The script below presents the codes required to obtain the data `historical_tmax` used in the example of Section 5.2 *Australian historical maximum temperature*. The function `rnoaa::meteo_pull_monitors()` may take a while to query a large number of stations in the first time. A copy of the data is provided in the data folder of the paper repository at: https://github.com/huizezhang-sherry/paper-cubble.

```
library(tidyverse)
library(cubble)
all_stations <- rnoaa::ghcnd_stations() %>%
  filter(str_starts(id, "ASN")) %>% # Australian stations start wiht "ASN"
  filter(last_year >= 2020) %>%
  mutate(wmo_id = as.numeric(wmo_id), name = str_to_lower(name)) %>%
  select(-state, -gsn_flag) %>%
  select(id, longitude, latitude, elevation, name,
         wmo_id, element, first_year, last_year) %>%
  rename(long = longitude, lat = latitude, elev = elevation)

tmax_stations <- all_stations %>%
  filter(element == "TMAX", first_year < 1970, !is.na(wmo_id))

raw_tmax <- all_stations %>%
  rowwise() %>%
  mutate(ts = list(rnoaa::meteo_pull_monitors(
    monitors = id, var = "TMAX",
    date_min = glue::glue("{first_year}-01-01"),
    date_max = glue::glue("{last_year}-12-31")
    ) %>%
      select(-id)
    )
  )

historical_tmax <- raw_tmax %>%
  select(-element) %>%
  unnest(ts) %>%
  mutate(tmax = tmax/10) %>%
  filter(lubridate::year(date) %in% c(1971: 1975, 2016:2020)) %>%
  as_cubble(index = date, key = id, coords = c(long, lat))

save(historical_tmax, file = here::here("data/historical_tmax.rda"))
```

4.2 Australian 2016-2020 climate data

The data `climate_full`, used in the examples in Sections 5.3, 5.4, and 5.5 of the main paper and in **??** here, can be obtained in a similar fashion with a slight change on the selected variable and date parameter in `rnoaa::meteo_pull_monitors()` as shown below. The full script is provided below and a copy of the data is also available in the data folder of the paper GitHub repository linked above.

```
# all the Australian stations have all of the three PRCP, TMAX, and TMIN recorded
aus_stations <- all_stations %>%
  filter(element %in% c("PRCP", "TMAX", "TMIN")) %>%
  nest(element: last_year) %>%
  rowwise() %>%
  filter(nrow(data) == 3) %>%
  select(-data)
```

```
aus_climate_raw <- aus_stations %>%
  rowwise() %>%
  mutate(ts = list(
    rnoaa::meteo_pull_monitors(
      monitors = id, var = c("PRCP", "TMAX", "TMIN"),
      date_min = "2016-01-01", date_max = "2020-12-31"
      ) %>%
      select(-id)
    )
  )

climate_full <- aus_climate_raw %>%
  unnest(ts) %>%
  mutate(tmax = tmax/10, tmin = tmin/10) %>%
  cubble::as_cubble(key = id, index = date, coords = c(long, lat))

save(climate_full, file = here::here("data/climate_full.rda"))
```