Review for JCGS-19-044

A new tidy data structure to support exploration and modeling of temporal data

The authors propose a new data infrastructure for temporal data with methods to aid the data wrangling of raw data to visualisation or model ready data adhering to the tidy data principles by Wickham (2014). Their methods are implemented in the `tsibble` R package.

The manuscript is beautiful, well-written and make a valuable contribution to the abstract representation of temporal data and the pipeline methods for intuitive data wrangling. Their `tsibble` R package follow closely to the spirit of `tidyverse` R package and thus those familiar with `tidyverse` will have less resistance to learning the new semantics. The existing popularity of `tsibble` package (over 180 stars on github and 20K downloads on CRAN) demonstrate its impact.

Having used `tsibble` myself, I believe the software design is well thoughtout, however, I believe that the authors can do better by improving the narrative in the manuscript. For example, authors may like to consider points below.

- More mini examples embeded throughout the manuscript. Mainly it didn't showcase to me `tsibble`'s capability well enough but understandably this may make the manuscript long and may be better left to vignettes.
- Finding a duplicate is a nice touch but I don't understand the significance of the Figure 5 mosaic plot in demonstration for the capability of `tsibble`. I understand that finding duplicate/wrong entry is made easier by `tsibble` but besides filtering those duplicate/wrong entries, the data wrangling for mosaic plot is easily done without `tsibble`. More specifically what is the significance `tsibble::index_by` below? Could the authors explain what is the benefit of explicit defining index at this stage as opposed to explicit specification at a later stage?

```
delayed_carrier <- us_flights %>%
  mutate(delayed = dep_delay > 15) %>%
  group_by(carrier) %>%
  index_by(year = year(sched_dep_datetime)) %>%  #<<
  summarise(
    Ontime = sum(delayed == 0),
```

```
    Delayed = sum(delayed)
  ) %>%
  gather(delayed, n_flights, Ontime:Delayed)
```

- In relation to the above point, perhaps it would be helpful to have a definition of tidy temporal data much like the three rules in tidy data (every row is an observation etc..). E.g. in `tsibble` the semantics `index` is a variable with some inherent ordering from past to present.
- I hold a similar sentiment regarding the calendar plot in Figure 12. It's a great plot but I don't necessary think this is a plot that show cased the benefit of `tsibble`.
- Figure 11 with missing data was a great demonstration for `tsibble` and the data times series pipeline shown in Figure 3. It may be helpful to reduce the opacity of the points as there is a lot of overplotting. The top line which lumps a number of customers together is not particuarly informative here. Perhaps it is better replaced with a histogram.

General comment:

- Authors may like to add/change `spread`/ `gather` verbs to `pivot_longer`/`pivot_wider` with the recent move to deprecate the former in `tidyr`. Likewise authors may like to consider to replace "long data" to "longer data" and "wide data" to "wider data" as long/wide implies absolute terms. https://tidyr.tidyverse.org/dev/articles/pivot.html
- Manuscript does implicitly feel more time series data oriented rather than a general temporal data. E.g. references of time points as `index` seems inherent from other time series data infrastructure (i.e. `zoo` and `xts`) and the presented case studies is not representative of any longitudinal study. This does not mean it should be changed but `index` to me was not as intuitive to be the temporal variable.
- Figure 4 animation doesn't play so authors may like to ensure that the final publication works. Github version works.
- Have the authors thought about how to deal with mixed time resolutions for `index`? E.g. some subjects have exact day recorded; some missing day but month & year recorded; some only date range recorded.
- The emphasis of the manuscripts is in R but do other languages also tend to define time index implicitly?

Minor comments:

- “summer time” -> “daylight saving time”
- P2 L22 “numeric data” -> “numerical data”
- P7 L42 “analytic point of view” -> “analytical point of view”
- P16 L53 “explicitly repeated references” -> “explicitly repeating references”