Tidy data structure and visualisation to support exploration and modeling of temporal data

A thesis submitted for the degree of

Doctor of Philosophy

by

Earo Wang

B.Comm. (Hons), Monash University

Department of Econometrics and Business Statistics

Monash University

Australia

February 2019

Contents

Acknowledgements

This document was created with bookdown (Xie, 2016), and the raw files contain all the R code to produce the plots and tables.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Earo Wang

Preface

Chapter 2 has been submitted to *Journal of Statistical Software* and is currently under review. It has won the 2018 ASA Statistical Graphics Student Paper Award.

Abstract

Temporal-context data sets often include a richness of information that is not possible to include in the typical data formats that are used in time series analysis. They also present some complications for modelling and visualisation, such as a long time spans, multiple factor variables, heterogeneous data types, low time resolutions, implicit missing values, and multilevel temporal components. In this thesis, we extend the conceptual tidy data framework, which provides the foundation for good modern data practice, to encompass time series data, and develop new graphics for displaying temporal data. First, we develop a new calendar view for visualising sub-daily time series data, which is particularly useful for viewing people patterns. Second, we have extended the tidy data concept to temporal data, and note that the "molten" data structure is flexible enough to handle the full richness of these complex data sets, heterogeneous variables, missing values. The tidy temporal data also builds on ideas of a data pipeline which improves and supports an organised worflow for data analysis. The third topic will focus on providing a visualisation framework for time series with nested and crossed factors. All the methods have been implemented in the R packages **sugrrants** for visualisation and **tsibble** for tidy time series data.

Chapter 1

Introduction

The conventional plot for time series is a line plot, where the measured variable is plotted against a time variable, and consecutive points are connected by lines. The line plot assumes: (1) there are no missing values, (2) the series is not too long so that seasonality can be viewed along with trend, (3) the time series stands alone without any complementary information, and (4) there is a single measured variable. The problem is that data rarely comes in this form. Quite often the temporal component is just one aspect of multi-facetted data, some observations are missing, the series could be very long, so that seasonality is lost relative to any long-term trend, measurements might be taken at irregular intervals, and we may have a big collection of time series. This thesis addresses these broader issues of better data structures and visualisation for big temporal-context data.

1.1 Research framework

Data visualisation in statistics has evolved in recent years to have graphics formally described using a grammar (Wilkinson, 2005**wickham2010layered**), and also data structures for statistical analysis have been organised into a conceptual framework of "tidy data" (Wickham, 2014). Both aspects are critical for evaluating the strength of the research, and for building on a foundation of sound data practice.

1.1.1 Tidy data structure

Wickham (2014) developed a set of tidy data principles to standardise and facilitate the data analysis process, and also suggested that each variable is a column, each observation is a row. An attribute describing a unit's properties (such as temperature, precipitation, pedestrian counts) is thought of as a variable. An observation refers to the same unit measured across each variable. A variable together with an observation defines the values that essentially comprise of a dataset. The newer concept of tidy data is the "long" form: every measured value is described by a unique set of identifiers. From the long form, data can be summarised, and shaped, and arranged in many different forms, making different types of analysis on the same data easier.

However, in some ways it is the classical statistical format of a rectangular table where each time series is a column, and each time index is a row. This format is the "wide" form of tidy data, and it is what is typically required for many statistical models as it proves computationally efficient when doing matrix operations.

The common data structure for time series data in statistical software, like R (**r**), is actually short-hand. For example, a time series defined as a `ts` object in R is a vector of values, annotated with a starting time and frequency. Associated with the data object are methods that can be applied to plot, summarise and model. However, this approach is model-centric rather than data-centric. Data typically doesn't come in this form. Data typically has a lot more with it, and to get it in this form requires extracting a small part of all the data, and maybe, even massaging it into regular time measurements. Working from a tidy format this specialist form could be created with several wrapper functions, whilst maintaining the connections to the complete data.

1.1.2 The grammar of graphics

A grammar of graphics was first proposed by Wilkinson (2005) and extended by **wickham2010layered**. These methods establish a conceptual framework for mapping data to graphical form. It is the analogue to thinking of statistics as a mapping of random

variables. For example, a mean is the mapping of n independent and identically distributed random variables, $X_1, \ldots, X_n$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. With this functional mapping, the statistic can be computed on a sample, and the properties of the statistic can be analytically studied.

The grammar formulates a set of rules to map variables to visual quantities. Therefore, we can make the comparisons between different displays and describe the differences under a standard framework. The grammar enables one to understand the process of creating visual graphics and the data structure behind the statistical plots. It is efficient and replicable to make any type of graphic display. The R package **ggplot2** (Wickham et al., 2018) is an implementation of the grammar of graphics.

1.2 Scope

This thesis research proposes a data-centric structure to represent time series by extending the "tidy data" principles. It allows rich data information to be included, which is not currently supported by time series model objects. It also provides a mapping bridging the semantics of a dataset to its physical representation, and consequently fosters ideas of a data pipeline that supports thinking of operations on the data variables. A collection of verbs are developed to smooth out the workflow for analysis of temporal data.

A new calendar-based graphics have been created and implemented to better visualise sub-daily data related to human behaviours. Tidy temporal data builds the foundation of calendarising the data using the data re-structuring approach. The grammar of graphics further adds the plotting capacities to the calendar format.

Tidy temporal data also integrates nested and crossed factors. A visualisation framework will be proposed and formulated for time series with nested and crossed factors.

Chapter 2

Calendar-based graphics for visualizing people's daily schedules

Abstract

Calendars are broadly used in society to display temporal information, and events. This paper describes a new R package with functionality to organize and display temporal data, collected on sub-daily resolution, into a calendar layout. The function `frame_calendar` uses linear algebra on the date variable to restructure data into a format lending itself to calendar layouts. The user can apply the grammar of graphics to create plots inside each calendar cell, and thus the displays synchronize neatly with **ggplot2** graphics. The motivating application is studying pedestrian behavior in Melbourne, Australia, based on counts which are captured at hourly intervals by sensors scattered around the city. Faceting by the usual features such as day and month, was insufficient to examine the behavior. Making displays on a monthly calendar format helps to understand pedestrian patterns relative to events such as work days, weekends, holidays, and special events. The layout algorithm has several format options and variations. It is implemented in the R package **sugrrants**.

Bibliography

Wickham, H (2014). Tidy Data. *Journal of Statistical Software* **59**(10), 1–23.

Wickham, H, W Chang, L Henry, TL Pedersen, K Takahashi, C Wilke, and K Woo (2018). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. http://ggplot2.tidyverse.org, https://github.com/tidyverse/ggplot2.

Wilkinson, L (2005). *The Grammar of Graphics (Statistics and Computing)*. Secaucus, NJ: Springer-Verlag New York, Inc.

Xie, Y (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. ISBN 978-1138700109. Boca Raton, Florida: Chapman and Hall/CRC. https://github.com/rstudio/bookdown.