MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Tidy data structures to support exploration and modeling of temporal-context data

Earo Wang, Dianne Cook, Rob J Hyndman

March 2018

Working Paper no/yr

Tidy data structures to support exploration and modeling of temporal-context data

Earo Wang
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: earo.wang@monash.edu
Corresponding author

Dianne Cook
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: dicook@monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: rob.hyndman@monash.edu

26 March 2018

JEL classification: C10,C14,C22

Tidy data structures to support exploration and modeling of temporal-context data

Abstract

ATemporal-context data is often rich with information, which is not supported by current time series model objects. The data may have heterogeneous variable types, implicit missings, multiple levels of of temporal structure, many different grouping variables. The conventional matrix structure that underlies time series models requires numerical variables, and implicit times, which have the effect of discarding some elements of the rich data available. This work addresses the broader issues of better data structures and modern data pipelines for analysing and visualising temporal-context data. We extend the tidy data concept to temporal data, and note that the 'molten' data structure is flexible enough to handle heterogeneity, low time resolutions, and implicit missing values. There are two supporting components required to turn 'molten' data into a valid temporal data: (1) an explicitly declared index variable containing time-stamps; and (2) a unique identifier the multiple groups of measurements, which is referred to as a 'key'. A syntactical approach is introduced to describe nested or crossed data structures, which employ the 'key'. The tidy data structure also supports thinking of operations on the data as part of a data pipeline, which facilitates the workflow for analysis of temporal data. A case study is included to illustrate the tidy structure, the data pipeline ideas and usage. A new representation of tidy temporal data along with the supporting methods has been implemented in the R package **tsibble**.

Keywords: blah, blah

1 Introduction

Data measured or collected at different time points are referred to as temporal-context data or time series. Figure **??** shows a glimpse of on-time performance of domestic flights in U.S.A from 2016 to 2017. It keeps track of on-time records for each flight's at its scheduled time over years. This dataset arrives in a rectangular form, which features:

- heterogeneous data columns, such as numbers, characters and timestamps;
- multiple measured variables including departure/arrival delay in minutes, flight time, distance between origin and destination, and etc.;
- low time resolutions, for example scheduled departure time in minutes;
- multi-levels of grouping like airports nested within cities, and cities further nested within states.

However, there is lack of data management tools for time series involving those characteristics in most statistical software, for example data wrangling, visualisation, analysing and forecasting in time-based context. It becomes more challenging in handling time series as increased data complications. The bottleneck that makes time series data frustrating to work with is the conventional data structure that a vector or matrix underlies.

Although a matrix is a tabular form, it assumes homogeneity (that is, all the columns must be real values.) and time indices implicitly inferred as attributes (or meta-information). This format of data input is model-oriented rather than data-oriented as it proves computationally efficient for many statistical models when doing matrix operations. However, it is has the effect of discarding intersting information in the data, is more opaque in the handling of the temporal components, and less efficient for many other data analysis tasks, such as data transformation, visualisation and model diagnostics, like forcing the fitting of a square peg into a round hole.

In addition, time indices form the basis of temporal data and the source of contextual information to understand temporal dependence. Time as attributes limits the use of time context, since it is difficult to access and manipulate. For example, the components of date-times, such as months, day of week, and time of day, cannot be easily extracted in order to examine seasonal fluctuations. It is also impossible to switch between different time zones and daylight savings to compare time across places when needed.

wickham2014tidy developed and formulated the conceptual framework of tidy data: (1) each variable forms a column; (2) each observation forms a row; (3) each type of observational unit forms a table. These principles attempt to standardise the mapping from the semantics of a dataset to its structure and facilitate data analysis in a coherent way. We shall extend "tidy data" into the time series domain.

This paper proposes a tidy data structure and a modern data pipeline for storing, managing and analysing time series data, using a collection of fluent and fluid tools to help with exploitation in temporal context.

2 Tidy temporal-context data

A modern reimagining of time series should provide heterogeneous data types and time indices as explicitly declared data column. This can be achieved using a "data frame" in R or other statistical languages to represent a tabular format, instead of "matrix". Tidy temporal-context data at least consists of:

1. index: an explicitly declared data variable contains time indices, such as date-times, year-months, years and etc.
2. key: a set of grouping factors uniquely identifies each unit that measurements take place on over time, which may include single or multiple columns.
3. interval: data with regular time interval results in a common time interval in one table.

In SQL database, a primary key, which uniquely defines each record in a database table, is equivalent to the composition of "index" and "key". In multivariate time series notation as Equation (1), X_{jt} represents series j, for $j = 1, \ldots, p$ and $1 \leq t \leq T$, in the form of a $T \times p$ matrix.

$$\begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1T} & X_{2T} & \cdots & X_{pT} \end{bmatrix} \tag{1}$$

2.1 Time index and interval

Time index forms an integral component and contextual base of temporal data. In temporal data frame, time-based index must be clearly stated as a data column rather than inferred as attributes, and thus can be accessible. This creates flexibility in handling time indices.

(1) temporal elements can be created, and then exploratory data visualisation and analysis (not specialist plotting) and multiple seasonality modelling for sub-daily data
(2) convert to the same time zone and thus compare

(3) join other data tables using the index as common key

For data indexed in regular time space, the time interval is obtained by computing the minimal positive time distance in a data table. This suggests that each observational unit collected at the same interval forms a table.

2.2 Keys

Key variables are usually discrete descriptors, and are typically variables that were created during the data collection to uniquely define the measured values. For instance, to distinguish the performance of each flight in the dataset, the "key" is the `flight` variable, allowing separation of multiple time series in one data table. The "key" not only identifies the unit to be measured over time, but also incorporates structures of data. Without a key, a data table can be considered as a univariate time series; in other words, the key is implicit. With a single key of more than one categories, it lends itself to a collection of time series in a table. But when there are at least two keys in the table, it indicates nested or crossed structures.

In experimental designs, a variable is crossed with another when every category of one variable co-occurs with every category of the other, while a variable is nested within another when each category of the former variable co-occurs with only one category of the latter. Regarding faceted plots in statistical graphics, **wilkinson2006grammar** discussed the difference between nesting and crossing structure in depth. The distinction between two graphical layouts may appear subtle but convey completely different meanings of the data. The crossing example could be gender with marital status: married women, married men, single women, and single men. It takes all the possible combinations into account, resulting in 2 by 2 drawing panels in a graphical device. If a panel goes empty, there exists no observation for that category in the dataset, but it would be filled with observations. However, considering pregnancy status and gender, there are only three possible combinations: pregnant women, non-pregnant women, and non-pregnant men, since pregnant status is nested under the gender variable. In this case, any physical layout of such nesting structure must not occur to an impossible category.

It appears more useful in making this distinction in statistical analysis including visualisation and modeling, compared to data manipulation.

3 Data pipeline

Tidy data builds a concrete foundation to enable pipeline data analysis, which provides a coherent and fluent framework to work with data. It helps (1) break up a big problem to into manageable blocks, (2) generate human readable analysis workflow, (3) avoid introducing mistakes as many as possible.

- **row-wise**: `filter()`, `slice()`, `arrange()`, `fill_na()`
- **column-wise**: `mutate()`, `select()`, `summarise()`, `tsummarise()`
- **rolling window**: `slide()`, `tile()`, `stretch()`
- **statistics**: `lag()`, `diff()`, `acf()`

4 Application: U.S.A domestic flights on-time performance (2016-2017)

A dataset of on-time performance of domestic flights in U.S.A from 2016 to 2017 is studied and explored for illustration of tidy data and data pipeline.

5 Discussion

A tidy representation of time series data, and data pipelines to facilitate data analysis flow have been proposed and discussed. It can be noted that tidy temporal data gains greater flexibility in keeping data richness, making data transformation and visualisation easily. A set of verbs provides a fluent and fluid pipeline to work with tidy time series data in various ways.

The ground of time series modelling or forecasting is left untouched in this paper. The future plan is to bridge the gap between tidy data and model building. Currently, it is required to casting to matrix from tidy data and therefore building a model. But time series models should be directly applied to tidy data as other wrangling tools do, without such an intermediate step. In particular, a univariate time series model, like arima and exponential smoothing, can be applied to multiple time series independently. A tidy format to represent model summaries and forecasting objects will be developed and implemented in the future. Model summaries include coefficients, fitted values, and residuals; forecasting objects include future time path and distributions generating prediction intervals.