MONASH
BUSINESS
SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

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

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

Earo Wang, Dianne Cook, Rob J Hyndman

May 2018

Working Paper no/yr

Tidy data structure 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

2 May 2018

JEL classification: C10,C14,C22

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

Abstract

Temporal-context data is often rich with information and time formats, for example multiple observational units, different time lengths, heterogeneous data types, nested and crossed factors and etc. This work presents a cohesive and conceptual framework for organizing and manipulating temporal data, which in turn flows into visualization and modelling routines. Tidy data principles are applied and extended to temporal data: (1) mapping the semantics of a dataset into its physical layout, (2) an explicitly declared index variable representing time, (3) a "key" comprised of single or multiple variables to uniquely identify units over time, using a syntatical and user-oriented approach in which it imposes nested or crossed structures on the data. This tidy data representation most naturally supports thinking of operations on the data as building blocks, forming part of a "data pipeline" in time-based context. A sound pipeline practice facilitates a succinct and transparent workflow for analyzing temporal data. Applications are included to illustrate tidy temporal data structure, data pipeline ideas and usage. The architecture of tidy temporal data has been implemented in the R package **tsibble**.

Keywords: temporal context, time series, data structure, R

1 Introduction

Temporal-context data consist of observational units indexed at different time points X_{jt}, where the j^{th} unit takes measurements of X on over time t, for $j = 1, \dots, N_i$ and $1 \leq t \leq T$. Time primarily forms the contextual basis of temporal data, but it could arrive in many possible formats. For example, data recorded at fine time resolutions (hours, minutes, and seconds) are typically associated with different time zones and daylight savings. Temporal data often carries with rich information other than the time: multiple observational units of different time lengths, multiple and heterogeneous measured variables, multiple grouping factors involving nested or crossed structures, linking to other data tables, and etc.

In the literature, time series and panel (longitudinal) data are common terms referred to temporal-context data, depending on the research fields. Researchers who are concerned with modelling large *T* and small *N* would name as "time series" (serial correlation); those who are interested in modelling small *T* and large *N* as "panel data" (asymptotic). The data format is two-dimensional array, but different modelling focuses lend the data input to different representations. A matrix is used to represent multivariate time series where each row represents observations measured at a time point and each column represents a series ("wide" form). This matrix representation requires homogeneity (that is, all the columns must be of same type.), time indices implicitly inferred as attributes or meta-information, series of same time length, and explicit missingness. By contrast, panel data are organised in rectangular form of heterogeneous column types where multiple study subjects are stacked and repeated for its time indices in a single column ("long" form), due to commonly arisen unbalanced panels. This specification requires explicitly declared panel variable and index, which has been implemented in Stata's time series module and R package **plm**. Evidently, this data organisation appears more flexible than matrix in supporting explicit time index, multiple subjects of different time lengths, and implicit missing values. [R]

Temporal data can often be aggregated in a manner that exhibits a nested or crossed structure, also known as hierarchical or grouped time series (Hyndman & Athanasopoulos 2017). For example, in a manufacturing setting, a company can add up every store's sales by region, by state and by country, which gives a strictly hierarchical time series; alternatively they can group the sales for each product together based on common attributes such as store, brand, price range and so forth, which leads to a non-hierarchical structure—a grouped time series. The R package **hts** is the implementation of this type of time series. But it is frustrating to create and work with such data objects due to non-extensible and non-modified. Variational data aggregation (some groups no need aggregation). [R]

Wickham (2014) coined the term "tidy data" and formalized the processing from messy data to "tidy data". These principles attempt to standardize the mapping from the semantics of a dataset to its structure and facilitate data analysis in a coherent way. Based on the systematic structuring principles, a grammar of data manipulation. [R]

This paper proposes a unified data representation of temporal-context data, blending time series of nested and crossed factors into a two-dimensional column-homogeneous array in a long format. By leveraging the "tidy data" principles, observations and variables position and bridge their meanings in both physical and internal structures. Data manipulation involves in transforming either observations or variables, or both, which can be described and achieved with a collection of shorthand operators. A chain of data transformations lend itself to a data pipeline.

2 Data semantics

The choice of representation of temporal-context data is made from a data-centric perspective, which is taken in the light of the operations that are to be performed on the data. Firstly, a data set must be structured in a "tidy" layout. Secondly, declaring the data set to contain temporal observations is determined by an "index" that represent time and a "key" that uniquely identifies each unit that measurements take place on over time. The "key" works similarly as the panel variable in the Stata's **tsset** command to define the units or subjects, but it is expanded to include multiple variables rather than a single one. A syntax is introduced to express a key consisting of nested and crossed factors. The composition of index and key uniquely defines each observation in a data table, which is equivalent to a primary key (Codd 1970) in a relational database.

Given the nature of temporal ordering, a temporal data set must be sorted by time index. If a key is explicitly declared, the key will be sorted followed by arranging time in ascending order.

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

3 Data pipeline

Literature about data pipeline.

A pipeline exhibits a hierarchy of data operations: (1) atomic (1-dimensional) vectors (`mean(variable)`) –> (2) (2-dimensional) data table (`summarise()`) –> repeat step (1) and (2) to form a chain. Reversely, a data pipeline is decomposed into rectangular blocks, and then into atomic strips. (UNIX pipe)

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()`
- **time-wise**: `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 Conclusion and future work

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.

References

Codd, EF (1970). A relational model of data for large shared data banks. *Communications of the ACM* **13**(6), 377–387.

Hyndman, RJ & G Athanasopoulos (2017). *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts. `OTexts.org/fpp2`.

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