DEMO ARXIV TEMPLATE

A PREPRINT

H. Sherry Zhang
Department of Econometrics and Business Statistics
Monash University
Melbourne, VIC
huize.zhang@monash.edu

Collaborators
Department of Econometrics and Business Statistics
Monash University
Melbourne, VIC

October 11, 2022

ABSTRACT

- Indices, useful, quantify severity, early monitoring,
- A huge number of indices have been proposed by domain experts, however, a large majority of them are not being adopted, reused, and compared in research or in practice.
- One of the reasons for this is the plenty of indices are quite complex and there is no obvious easy-to-use implementation to apply them to user's data.
- The paper describes a general pipeline framework to construct indices from spatio-temporal data,
- This allows all the indices to be constructed through a uniform data pipeline and different indices to vary on the details of each step in the data pipeline and their orders.
- The pipeline proposed aim to smooth the workflow of index construction through breaking down the complicated steps proposed by various indices into small building blocks shared by most of the indices.
- The framework will be demonstrated with drought indices as examples, but appliable in general to environmental indices constructed from multivariate spatio-temporal data

***K*eywords** spatio-temporal data • indices • data pipeline

1 Introduction

Why index is useful, why people care about indices

Quantities that can't be directly measured is ubiquitous in our society and they can often be relevant for policy making, resource allocation, early prevention and operational purpose. For example, fire authority would be interested to quantify fire risk since bushfire can have a huge impact on monetary loss, health, and local ecosystem among others. Climatologists would be interested in monitoring the change in global condition since variability in atmospheric and oceanic condition has a direct impact on the global weather and climate. Usually more than one variable is related to the concept of interest and they often needs to be integrated to make decisions. A commonly used approach here is to construct an index with the observable variables. This allows analysts to compare the quantity of interest across countries and also cross time.

Define what is an index, what is not

The meaning of indices varies across disciplines and in this article, an index is defined as a tool to quantify a concept of interest that does not have direct measure. The concept of interest doesn't have direct measure can because it is impractical to measure the concept of interest at the population level. For example, it would be nearly impossible to include all the available stocks in the market to characterise stock market behavior, indices like Dow Jones Industrial Average, S&P 500, and Nasdaq Composite select a representative set of stocks to measure the overall market behavior. Also belong to this category is the economic indices like Consumer Price Index, where price changes of a basket of items are weighted to measure inflation. The lack of direct measure could also due to the absence of a universal definition on the concept of interest. Many natural and social concepts are fundamentally multi-facets and some indices combines measures from different domains to characterise the concept of interest. This includes drought indices constructed from meteorological, agricultural, hydrological, and social-economic variables [give some names?]. Social development indices like Human Development Index [ref] and Global Liveability Index [ref] measure various aspects on the quality of human capital and urban life.

Despite many measurements have index in its name, they cannot be technically classified as indices according to the definition given above. A major reason for these measures to lose its index membership is that the underlying concept of interest is straight forward to measure once defined. Sometime variables like precipitation of driest month or percentage of days when maximum temperature is below a certain percentile are classified as climate indices, however, precipitation and temperature can be directly measured by rain gauge and thermometer. These variables are transformation on precipitation and temperature. They can be used to construct indices but are not indices themselves. What is also not indices are many remote sensing based drought measures as they are transformation on one or more remote sensor bands. This includes Normalized Difference Vegetation Index (NDVI) [ref], a transformation difference over sum on near-infrared (NIR) and red visible channel, and Vegetation Condition Index [ref], a rescale of NDVI into [0, 1], among others.

TODO: confirm the actual climate indices are indices by our definition Questions: need reference for stock market indices, CPI?

What is the problem with current workflow on index construction

There has not yet been a standardised workflow on index construction. For simple indices, users may be able to work out the calculation from the equations or workflow provided, while for more complicated construction, software is not even available for open-source consumption. Even when the software implementation is available, most indices are computed through a single index command, for example, the package `SPEI` provides two drought indices: SPI and SPEI through function `spi()` and `spei()`. This software design focuses on the name of the index and includes all the steps in a single index function. With this design, there is little room for modification. For example, fit a new distribution that is not provided in the initial code. Also, when scaled to a large number of indices, it increases the cognitive load of remembering hundreds of index names. A better approach is to modularise these steps and construct a pipeline for users to build indices through choosing different building blocks from each module. With this design, users focus on the operations involved in the index construction. If an operation from another index is useful, users can borrow it to modify an existing index.

who would benefit from this paper

The article contributes a pipeline framework to construct general indices.

- domain scientists who propose new indices and directly use indices in analysis. [use multiple indices to compare]
- general scientists who uses indices as part of analysis, modelling process. Official indices are usually published in a public table with values calculated. Although methods of calculation are provided by the relevant

index provider, raw data and scripts are usually not available for public consumption. The pipeline proposed in the paper opens up the possibility for general scientists to construct indices customised to their research purpose.
- decision makers, towards understand the
- statisticians, so as to [draw focus/ advocate] on statistical workflow and software design.

The rest of the paper is structured as follows: the concept of data pipeline in R is reviewed in Section 2. Section 3 presents the data pipeline for index construction. Section 4 explains how to include a new building block in each pipeline module. Examples are given in Section 5 to demonstrate the index construction with the pipeline built.

2 Data pipeline in R

2.1 Tidy data

Before the concept of tidy data (Wickham 2014), tabular data arrive at data analysts in all different ways. Different analysts would write customised scripts for analysing the specific data. These scripts can be extended to other data analysed by the same people or group but this is not generalizable directly to another dataset. When the tidy data concept comes, variables and values are arranged so that 1) Each variable forms a column, 2) Each observation forms a row, and; 3) Each type of observational unit forms a table. With this specific layout, wrangling on tabular data can be standardised into a grammar of data manipulation in `dplyr` (Wickham et al. 2022).

A similar issue happens with index construction where researchers construct their own indices in their own ways and there has not yet been a tidy principle on index construction. Also, this tidy principle on index construction is more complex than those in tidy data and the `dplyr` package. It has to encompass the workflow of transformation from the raw data towards the final index series.

2.2 Data pipeline

Constructing a pipeline that divides a complex procedure into steps that can be concatenated has been adopted widely in the R community.

The data pipeline in interactive graphics is a set of steps that transform the raw data to the plots displayed on the screen. The initial pipeline proposed by Buja et al. (1988) involves the following steps: non-linear transformation, variables standardization, randomization, projection engine, and viewporting. The initial pipeline proposed by Buja et al. (1988) involves the following steps: non-linear transformation, variables standardization, randomization, projection engine, and viewporting. Another example in the early work of pipeline by Sutherland et al. (2000) describes a three-step pipeline: variable standardization, dimension reduction, and scaling data into the viewing window. This pipeline also includes the transformation on spatial and temporal variables, i.e. computing time lag on temporal variables. This pipeline also includes the transformation on spatial and temporal variables, i.e. computing time lag on temporal variables. Wickham et al. (2009) argues that whether made explicit or not, pipeline has to be presented in every graphics program and breaking down graphic rendering into steps is also beneficial for understand the implementation and compare between different graphic systems.

The data pipeline concept is further enhanced by the pipe operator (`%>%`) in R where a set of operations, or steps, can be chained together to form a set of instructions.

A more recent data pipeline is tidymodels (Kuhn and Wickham 2020), a set of packages for machine learning models following the tidyverse principles (Wickham et al. 2019). Steps: Exploratory data analysis (EDA), feature engineering, model tuning and selection, and model evaluation.

3 A pipeline for building statistical indices

The construction of natural hazard indices also follows a set of steps, which is usually illustrated using a flowchart in the paper. However, every researcher follows a certain design philosophy and steps taken in the index constructed by different researchers are not aligned. This discourages experiment with multiple indices. Initiate a new workflow when computing a new index.

The most popular indices (i.e. SPI, SPEI, etc) have existing software implementation (`SPEI`) to be applied to a different set of data.

constructing time series index should also be encapsulated in my framework

Here we assume a concept of interest is determined, relevant variables/ indicators are identified and available to construct indices.

3.1 Raw data

Another section on original data directly downloaded, can have different spatial resolution, temporal granularity, data quality problem. After processing them and align them together they become the "raw data"

The data used to construct the natural hazard index usually have three dimensions, one for location, one for time, and one for multivariate. Mathematically, it can be written as $X_{j,s,t}$, where $j = 1, 2, \cdots, J$ for variable, $s = 1, 2, \cdots, S$ for location, and $t = 1, 2, \cdots, T$ for time.

The location s can refer to vector points or areas characterised by longitude-latitude coordinates, or raster cells obtained from satellite images.

The time dimension t can be daily, weekly, biweekly (14-16 days), monthly, or even quarterly

Variables

This multidimensional array structure is commonly used in geospatial analysis

Given the variety of data sources at different spatial resolution and temporal granularity, the raw data may first come in multiple pieces. Sometimes, even a considerable amount of work is needed to align the spatial and temporal extent of multivariate data.

A notation for different variables have different spatial and temporal granularity X_{j_1,s_1,t_1}???

3.2 Spatial aggregation

mostly happen with raster data

3.3 Scaling

A specific transformation on the scale of the data

z-score standardising, min-max standardisation into [0, 1] or [0, 100], percentage change on the baseline close to variable transformation step

3.4 Variable transformation

Restrict it to single variable, square root, log etc could be linearly, also non-linear

GAM, can you do additive model pairwise/ three-way

3.5 Temporal processing

3.6 Dimension reduction

sometimes called feature extraction in the machine learning community With drought indices, the extraction of meaningful variables from the original data is usually supported by the water balance model, for example, in SPEI, the step that create d out of precipitation and potential evapotranspiration (PET) has theoretical backup from [see paper.]

Also include weighting

3.7 Normalising

The purpose of normalising is for cross-comparison. This step can get criticism from analysts for …

specifically for converting from a fitted distribution to normal score via reverse CDF function, non-parametric formula, or empirical approximation, a common step in many index: SPI, SSI, Z score. The purpose of normalising is to convert the index into a standardised series after all the steps for the ease of comparison.

3.8 Benchmarking

3.9 Simplification

Discretise the continuous index into a few labelled categories. For communicating the severity of natural hazard to general public.

uniform workflow to work with index construction.

- illustration
- math notation
- benefit of the pipeline approach
  - index diagnostic
  - uncertainty

4 Incorporating new buliding blocks into the pipeline

5 Examples

5.1 Constructing Standardised Precipitation Index (SPI)

- a basic workflow and congruence with results in the `SPEI` pkg
- allow multiple distribution fit
- allow bootstrap uncertainty

5.2 Calculating SPEI with raster data

Reference

Buja, A, D Asimov, C Hurley, and JA McDonald. 1988. "Elements of a Viewing Pipeline for Data Analysis." In *Dynamic Graphics for Statistics*, 277–308. Wadsworth, Belmont.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Sutherland, Peter, Anthony Rossini, Thomas Lumley, Nicholas Lewin-Koh, Julie Dickerson, Zach Cox, and Dianne Cook. 2000. "Orca: A Visualization Toolkit for High-Dimensional Data." *Journal of Computational and Graphical Statistics* 9 (3): 509–29. https://www.jstor.org/stable/1390943.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*.

Wickham, Hadley, Michael Lawrence, Dianne Cook, Andreas Buja, Heike Hofmann, and Deborah F. Swayne. 2009. "The Plumbing of Interactive Graphics." *Computational Statistics* 24 (2): 207–15. https://doi.org/10.1007/s00180-008-0116-x.