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 4, 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 and reused in research or in practice.
- One of the most commonly used drought index, SPI and SPEI, are available through the `SPEI` package and its readily avaiability to be applied on user's data
- However, even if most of the indices have available existing implementation available to users, there would be hundreds of them and it will be burdensome to remember
- The paper describes a general pipeline framework to construct indices from spatio-temporal data, so that indices can be constructed through [a set of general steps].
- The benefit of such a framework is that ...
- 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

Index construction is a way to summarise complicated information (used in environmental data). The complexity of these information may involve a spatial distribution, a temporal scale that defines the frequency of the data, and a multivariate perspective that collects different climate variables.

Various methods are proposed to extract multivariate information on the spatial coverage, across time.

Indices with readily available code to implement have more potential to be computed by researchers, i.e. the package `SPEI` (Beguería and Vicente-Serrano 2017) has implemented the SPI and SPEI index for calculating drought indices. [Another package in a different domain]

Each individual index follows its own data pipeline and it can be difficult to evaluate how an index can be affected by tweaking parameters in a certain step, rearranging the order of steps, using to different method. This paper proposes a data pipeline for constructing indices using multivariate spatio-temporal data. The steps involved in the pipeline are general in nature and flexible to be adopted to most index construction for environmental data.

2 Natural hazard indices

2.1 Climate indices

2.2 Drought indices

2.3 Economics/ social indices

3 Data pipeline in R

3.1 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.

3.2 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). [expand on tidymodels]

4 A pipeline for building natural hazard 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

reusable

4.1 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}???

4.2 Spatial aggregation

4.3 Variable transformation

Take one or multiple variables to create a new variable, linearly? 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.]

4.4 Temporal processing

4.5 Dimension reduction

4.6 Normalising

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.

4.7 Benchmarking

4.8 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

5 Extending the pipeline

6 Examples

6.1 Constructing Standardised Precipitation Index (SPI)

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

Reference

Beguería, Santiago, and Sergio M. Vicente-Serrano. 2017. *SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index*. https://CRAN.R-project.org/package=SPEI.

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.