

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 1, 2022

ABSTRACT

- The paper describes a framework for constructing indices
- use drought indices as examples, but applicable in general to environmental indices constructed from multivariate spatio-temporal data

Keywords 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.

Numerous indices have been proposed by researchers and practitioners to monitor natural hazard, for example, Alahacoon and Edirisinghe (2022) reviews 111 drought indices derived from traditional and remote sensing data; [review of index construction in other area, climate indices, many other reviews in drought].

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

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

3 Data pipeline in R

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. 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. 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.

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 $j = 1, 2, \dots, J$ for variable, $s = 1, 2, \dots, S$ for location, and $t = 1, 2, \dots, T$ 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

4.4 Temporal processing

4.5 Dimension reduction

4.6 Normalising

4.7 Benchmarking

4.8 Simplification

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

- Alahacoon, Niranga, and Mahesh Edirisinghe. 2022. “A Comprehensive Assessment of Remote Sensing and Traditional Based Drought Monitoring Indices at Global and Regional Scale.” *Geomatics, Natural Hazards and Risk* 13 (December): 762–99. <https://doi.org/10.1080/19475705.2022.2044394>.
- 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, 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, 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>.