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

September 30, 2022

ABSTRACT

- The paper describes a framework for constructing indices
- use drought indices as examples, but appliable 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

Constructing a pipeline that divides a complex procedure into steps that can be concatenated has been adopted widely in the computational statistics community. One particular example is a machine learning pipeline, tidymodel (Kuhn and Wickham 2020). [more details on pipeline]

Statistical pipeline in interactive graphics

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.

Statistical pipeline in other domains

TODO: read papers on other pipeline in R

4 A pipeline for natural hazard indices

uniform workflow to work with index construction.

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

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

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, 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.