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 25, 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

Keywords spatio-temporal data • indices • data pipeline

1 Introduction

Why index is useful, why people care about indices

incorporate the following in why using index: multiple pieces of information (variables) that need to be taken into account

Many concepts relevant to decision making cannot be directly measured, however, they are crucial for resource allocation, early prevention, and other operational purpose. For example, fire authorities would be interested to quantify fire risk since bushfires can have a huge impact on monetary loss, health, and the local ecosystem. Climatologists would be interested in monitoring the change in global climate since variability in atmospheric and oceanic conditions has a direct impact on global weather and climate. Usually this concept of interest is associated with more than one variables and these variables need to be integrated to make decisions on the subject matter. A common approach to quantify concepts like these is to construct an index using these relevant variables. This allows researchers to compare the quantity of interest across entities (i.e. countries, regions) and also cross time.

Define what is an index, what is not

In this article, an index is defined as a tool to quantify a concept of interest that does not have a direct measure. The concept of interest doesn't have a direct measure can because it is impractical to measure 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, so 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 belonging to this category are the economic indices like the Consumer Price Index, where price changes of a basket of items are weighted to measure inflation. The lack of direct measure could also because the concept itself is an unobservable human construction, rather than a physical quantity that can be measured. Many natural hazard and social concepts falls into this category. This includes drought indices constructed from meteorological, agricultural, hydrological, and social-economic variables, e.g. Standardised Precipitation Index (SPI) (McKee et al. 1993) and Aggregated Drought Index (ADI) (Keyantash and Dracup 2004) among others. Social development indices like Human Development Index (United Nations Development Programme 2022) and Global Liveability Index (Economist Intelligence Unit 2019) measure various aspects of the quality of human capital and urban life.

still need to tweak the tone a bit: "they are called index, they are not the index we will talk about"

Despite many quantity having the term *index* in their name, they cannot be technically classified as indices according to the definition given above. The reason for these quantities to lose their index memberships is that they are variables can be accurately measured given the instrument precision. This includes quantities like precipitation of the driest month or percentage of days when maximum temperature is below 10th percentile. They are measures of precipitation and percentage of days under specific conditions (dries month, maximum temperature below 10th percentile). They are variables, or indicators, that can be used to construct indices but are not indices themselves. Similarly, a set of remote sensing indices are not indices, since they are measures of electromagnetic wave reflectance. This includes Normalized Difference Vegetation Index (NDVI) (Tucker 1979), derived from the ratio of difference over sum on two segments in the spectrum, also called band: near-infrared (NIR) and red. So are the "indices" derived from NDVI, e.g. Vegetation Condition Index (Kogan 1995). Notice that this does not exclude all the construction derived from remotes sensor variables to be valid indices. For example, Vegetation Drought Response Index (Brown et al. 2008) is a valid index since it integrates climate, satellite, and biophysical variables to quantify vegetation stress.

What is the challenges with current index construction

see if there is any paper describing this type of pains

useful to reference tidy data and tidy model that makes the workflow on modelling tidy somewhere in introduction

Currently, index construction lacks a standardised workflow. It is often up to researchers or research institutions to decide whether to provide open source code on the new indices, what would be the best user interface for other researchers to use the new indices, and how easily the new indices can be compared with other existing indices. This makes the computation lack transparency and indices cumbersome to experiment with:

- Researchers who wish to validate the indices calculated from large institutes need to reinvent the wheels themselves since the source code used for computing is often not available for public consumption;
- Open-source code provided by research groups has a narrow margin for exploring other options outside the provided;
- Similar steps used by different indices are difficult to spot since the design of the user interface for indices often includes all the steps under a single function call; and

- It is generally hard to inspect intermediate results during the index construction if users wish to check the output of a certain step.

what can be done if people adopt this pipeline/ why it is beneficial?

first sentence: tailor to their indices without thinking about the big picture

~~While every jumbled index jumbles in its own way, [perspicuous] clearly-laied out indices are all alike.~~ This paper proposes a data pipeline for index construction. By recognising the common steps shared by many indices, we develop a pipeline that breaks down index construction into multiple modules and allow operations in various modules to be combined like building blocks to construct indices. The pipeline approach is general while adaptable to most index construction. It allows indices to be created, studied, and compared in a structured tidy form and enables statistical analysis of indices to be performed easily: More specifically, it enables researchers to 1) validate the indices calculated from external organisations, 2) unify various indices under the same framework for computing, 3) swap or adjust individual steps in the index construction to study their contribution, 4) calculate uncertainty on indices through bootstrap or others, 5) enhance existing indices through comparing and studying their statistical properties, and finally, 6) propose new indices from combining different steps in existing indices.

who would benefit from this paper

This work is of interest to researchers actively developing new indices since it encourages new indices to be delivered in an easy-to-reproduce design. It would also provide analysts who wish to compute a range of indices in their analysis a uniform interface to build relevant indices from raw data. For statisticians and software developing engineers, this work frames the process of index construction in a more user-oriented workflow and could motivate similar research for other process in scientific computing.

The rest of the paper is structured as follows: Section 2 reviews the concept of data pipeline in R. The pipeline framework for index construction is presented in Section 3. 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

As a scripted language for data analysis, R provides a interactive environment for users to code with data without the abstraction a typical programming language requires. This flexible environment is paramount for data analysis since in practice, data analysis is an iterative process that involves trials and errors to solve the problems. Analysts need the freedom to quickly experiment with different ideas when working with data. The tidyverse suite provides a collection of R package for conducting data analysis. These tools directly operate on data, and users, even new users, can easily read from the code on what operation is performed on the data and what to expect as the output. All of these are built upon a simple prior knowledge of the tidy data principle (Wickham 2014), which prescribes three rules to structure tabular data: 1) Each variable forms a column, 2) Each observation forms a row, and; 3) Each type of observational unit forms a table.

Constructing indices also involves this working pattern of trials and errors on different options before finding the final solution. A tidy workflow will give researchers more flexibility to tune the steps as they wish and while maintain a readable and intuitive code base.

2.2 Data pipeline

The concept of pipeline (and pipe) is not new concept to computing. It refers to a set of data processing elements connected in series, where the output of one element is the input of the next one. Wickham et al. (2009) argues that whether made explicit or not, pipeline has to be presented in every graphics program. The paper also argues that breaking down graphic rendering into steps is beneficial for understand the implementation and compare between different graphic systems. Both arguments are applicable to index construction. Papers that proposing new indices always present a workflow diagram to illustrate how data is transformed into the final indices, but not much implementation has explicitly present this this step-by-step transformation in the code design.

To better understand the pipeline approach in statistic computing, examples in interactive graphics and machine learning models are given. In interactive graphics, the data pipeline is used to describe how the data is processed to be rendered on the screen. Examples of early work on building pipeline for interactive graphics includes Buja et al. (1988), Sutherland et al. (2000), and Xie, Hofmann, and Cheng (2014). Steps in these pipelines include non-linear transformation, variable standardization, randomization and dimension reduction. A more familiar example from tidymodels (Kuhn and Wickham 2020) presents a set of packages for machine learning models following the tidyverse principles (Wickham et al. 2019). The machine learning pipeline constitutes data resampling, feature engineering, model fitting, model tuning, and model evaluation each being responsible by a separate package. Before tidymodel, analysts who wish to fit different models need to learn the workflow of each individual algorithm and need to rewrite most of the model fitting, tuning, and evaluation code when switching from one model to another. The tidymodel approach unifies a wide range of models under a single pipeline where users can substitute different models while maintain the same pipeline. This allows users to easily experiment with different machine learning models a lot easier.

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 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.

Normalising is usually the last step

3.5 Variable transformation

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

change the shape of the variable

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

3.6 Temporal processing

3.7 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

3.8 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

Brown, Jesslyn F., Brian D. Wardlow, Tsegaye Tadesse, Michael J. Hayes, and Bradley C. Reed. 2008. "The Vegetation Drought Response Index (VegDRI): A New Integrated Approach for Monitoring Drought Stress in Vegetation." *GIScience & Remote Sensing* 45 (1): 16–46. https://doi.org/10.2747/1548-1603.45.1.16.

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.

Economist Intelligence Unit. 2019. "The Global Liveability Index 2019." The Economist. https://www.cbeinternational.ca/pdf/Liveability-Free-report-2019.pdf.

Keyantash, John A., and John A. Dracup. 2004. "An Aggregate Drought Index: Assessing Drought Severity Based on Fluctuations in the Hydrologic Cycle and Surface Water Storage." *Water Resources Research* 40 (9). https://doi.org/10.1029/2003WR002610.

Kogan, F. N. 1995. "Application of Vegetation Index and Brightness Temperature for Drought Detection." *Advances in Space Research*, Natural Hazards: Monitoring and Assessment Using Remote Sensing Technique, 15 (11): 91–100. https://doi.org/10.1016/0273-1177(95)00079-T.

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

McKee, Thomas B, Nolan J Doesken, John Kleist, et al. 1993. "The Relationship of Drought Frequency and Duration to Time Scales." In *Proceedings of the 8th Conference on Applied Climatology*, 17:179–83. 22. Boston, MA, USA.

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.

Tucker, Compton J. 1979. "Red and Photographic Infrared Linear Combinations for Monitoring Vegetation." *Remote Sensing of Environment* 8 (2): 127–50. https://doi.org/10.1016/0034-4257(79)90013-0.

United Nations Development Programme. 2022. "Human Development Report 2021-22." New York. http://report.hdr.undp.org.

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, 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.

Xie, Yihui, Heike Hofmann, and Xiaoyue Cheng. 2014. "Reactive Programming for Interactive Graphics." *Statistical Science* 29 (2): 201–13. https://www.jstor.org/stable/43288470?seq=1.