

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 18, 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 applicable 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 problem with current index construction

From last week: indices constructed separately, but there are lots of are same [!], but hard to see from the current construction. “messy data are messy in there own way”. Describe how they are same, so we can modular them

Currently, index construction lacks structure. It is often up to the researchers to decide whether to provide open source code on the new indices, what would be the best user interface for the new indices, and how easily the new indices can be used by end users, along with other existing indices. This causes the problem that indices are often designed separately and many indices simply lump all the steps into a [single command function]. While this simplifies the amount of code needed to derive index values from a set of inputs, many problems arise when the code is adopted outside the research group that wrote the index. First of all, indices created by a single command provide little room for modification. For example, many indices involve a step to fit variables into a particular distribution. Index researchers may provide options for different fits, but it is generally hard to experiment with options outside the provided distributions. Also, many indices share similar or identical steps in part of their construction, a [single command function] eliminates the possibility to borrow codes written by others and forces researchers to rebuild the wheels when creating new indices. Lastly, users may wish to inspect intermediate results during the index construction to validate the output or to compare results with other indices, but [single command functions] do not support this demand unless the users can customise the source code as they need.

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

Ideally each command corresponds to a user action

While jumbled indices jumble in their own ways, perspicuous indices are all alike. A clear pipeline for indices construction brings transparency to the algorithm. Many often, official websites like National Oceanic and Atmospheric Administration (NOAA) will regularly updates the raw data and the indices calculated for users’ consumption and provides description of the methodology. However, the source code and the scripts is usually not publicly available and there is no tool, or users need to be its own tool, to verify the calculation. Once a data pipeline is built, users can follow the steps in the index construction to validate the numbers published by the authorities. Once a data pipeline is constructed, it opens possibility to do statistics on indices. For example, bootstrap and other methods could be applied into the pipeline to calculate uncertainty on the index value with randomisation. Properties of existing indices can be examined through swapping the steps involved in the construction. For example, some indices prefer to standardise the input variables in the beginning of the workflow while other prefer to move the standardisation towards the end and directly apply it on the index value. A flexible pipeline approach gives the freedom to construct the index in both ways by placing the standardisation step in different stages of the pipeline. The importance of the position of a single steps can be compared through indices constructed by placing the step at different stages of the pipeline. Also, the pipeline approach motivates new indices through combining existing steps from different indices.

who would benefit from this paper

this work could be interested to scientists actively developing new 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. The pipeline proposed in the paper opens up the possibility for general scientists to construct indices customised to their research purpose.
- 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 `dp1yr` (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.

- easy to operate properly: should be designed that users know what is appropriate to do
- promote good scientific methodology: should protect users from doing stupid things, doing right thing

good statistical practice

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 $j = 1, 2, \dots, J$ 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} ???

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](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](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, 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>.