A Tidy Framework and Infrastructure to Systematically Assemble Spatio-temporal Indexes from Multivariate Data

H. Sherry Zhang[1], Dianne Cook[1], Ursula Laa[2], Nicolas Langrené[3], Patricia Menéndez[1]

ARTICLE HISTORY
Compiled August 20, 2023

[1] Department of Econometrics and Business Statistics, Monash University, Melbourne, Victoria, Australia
[2] Institute of Statistics, University of Natural Resources and Life Sciences, Vienna, Austria
[3] Department of Mathematical Sciences, BNU-HKBU United International College, Zhuhai, Guangdong, China

ABSTRACT
Indexes are useful for summarizing multivariate information into single metrics for monitoring, communicating, and decision-making. While most work has focused on defining new indexes for specific purposes, more attention needs to be directed towards making it possible to understand index behavior in different data conditions, and to determine how their structure affects their value and variation in values. Here we discuss a modular data pipeline recommendation to assemble indexes. It is universally applicable to index computation and allows investigation of index behavior as part of the development procedure. One can compute with different the index with different parameter choices, adjust steps in the index definition by adding, removing, and swapping them to experiment with various index designs, calculate uncertainty measures, and assess an index's robustness. The paper presents three examples to illustrate the pipeline framework usage: comparison of two different indexes designed to monitor the spatio-temporal distribution of drought in Queensland, Australia; the effect of dimension reduction choices on the Global Gender Gap Index (GGGI) on a country's ranking; and how to calculate bootstrap confidence intervals for XXX index. The methods are supported by a new R package, called `tidyindex`.

KEYWORDS
indexes; data pipeline; software design; uncertainty; decision-making

1. Introduction

Indexes are commonly used to combine and summarize different sources of information into a single number for monitoring, communicating, and decision-making. They serve as critical tools across the natural and social sciences. Examples include the Air Quality Index, El Niño-Southern Oscillation Index, Consumer Price Index, QS University

CONTACT: H. Sherry Zhang. Email: huize.zhang@monash.edu.

Rankings, and the Human Development Index. In environmental science climate indexes are produced by major monitoring centers, like the United States Drought Monitor and National Oceanic and Atmospheric Administration, to facilitate agricultural planning and early detection of natural disasters. In economics, indexes provide insight into market trends through combining prices of a basket of goods and services. In social sciences, indexes are used to monitor human development, gender equity, or university quality. The problem is that every index is developed in its own unique way, by different researchers or organizations, and often indexes designed for the same purpose cannot easily be compared.

To construct an index, experts typically start by defining a concept of interest that requires measurement. This concept often lacks a direct measurable attribute or can only be measured as a composite of various processes, yet it holds social and public significance. To create an index, once the underlying processes involved are identified, relevant and available variables are then defined, collected, and combined using statistical methods into an index that aims to measure the process of interest. The construction process is often not straightforward, and decisions need to be made, such as the selection of variables to be included, which might depend on data availability and the statistical definition of the index to be used, among others. For instance, the indexes constructed from a linear combination of variables need to decide on the weight assigned to each variable. Some indexes have a spatial and/or temporal component, and variables can be aggregated to different spatial resolutions and temporal scales, leading to various indexes for different monitoring purposes. Hence, all these decisions can result in different index values and have different practical implications.

To be able to test different decision choices systematically for an index, the index needs to be broken down into its fundamental building blocks to analyze the contribution and effect of each component. We call this process of breaking the index construction into different steps the index pipeline. Such decomposition of index components provides the means to standardize index construction via a pipeline and offers benefits for comparing among indexes and calculating index uncertainty.

In this work, we provide statistical and computational methods for developing a data pipeline framework to construct and customize indexes using data. The proposed pipeline comprises various modules, including temporal and spatial aggregation, variable transformation and combination, distribution fitting, benchmark setting, and index communication. Given the decisions analysts need to make when combining multivariate data into indexes, the proposed pipeline enables the evaluation of how the specific choice can affect the index, as well as how the index may appear under alternative options. Furthermore, uncertainty calculation can also flow through the pipeline, providing the index with confidence measures.

The rest of the paper is structured as follows: Section 3 reviews the tidy framework in R and how index construction can benefit from such a framework. The details of the pipeline modules are presented in Section 4. Section 5 explains the design of the `tidyindex` package that implements the modules. Examples are given in Section 6 to illustrate the use cases of the pipeline.

2. Background to index development

There are many documents providing advice on how to construct indexes for different fields, and review articles describing the range of available indexes for specific purposes. The OECD handbook (OECD, European Union, and Joint Research Centre - European Commission 2008) provides a comprehensive guide for computing socio-economic composite indexes, with detailed steps and recommendations. The drought index handbook (Svoboda, Fuchs, et al. 2016) provides details of various drought indexes and recommendations from the World Meteorology Organization. Zargar et al. (2011), Hao and Singh (2015) and Alahacoon and Edirisinghe (2022) are review papers describing the range of possible drought indexes.

There is also some attention being given to the diagnosis of indexes, and incorporation of uncertainty. Jones and Andrey (2007) investigates the methodological choices made in the development of indexes for assessing vulnerable neighborhoods. Saisana, Saltelli, and Tarantola (2005) describes incorporating uncertainty estimates and conducting sensitivity analysis on composite indexes. Tate (2012), similarly, makes a comparative assessment of social vulnerability indexes based on uncertainty estimation and sensitivity analysis. (XXX Something about Ursula's colleagues paper here too.)

There are also R packages supporting index calculation. The `SPEI` package (Vicente-Serrano, Beguería, and López-Moreno 2010) computes two drought indexes. The `gpindex` package (Martin 2023) computes price indexes, and the `fundiversity` package (Grenié and Gruson 2023) computes functional diversity indexes for ecological study. The package `COINr` (Becker et al. 2022) is more ambitious, making a start on following the broader guidelines in the OECD handbook to construct, analyze, and visualize composite indexes.

From reviewing this literature, and in the process of developing methods for making it easier to work with multivariate spatiotemporal data, it seems possible to think about indexes in a more organised, cohesive and standard manner. Actually, it seems that the area could benefit from a *tidy* approach.

3. Tidy framework

The tidy framework consists of two key components: tidy data and tidy tools. The concept of tidy data (Wickham 2014) prescribes specific rules for organizing data in an analysis, with observations as rows, variables as columns, and types of observational units as tables. Tidy tools, on the other hand, are concatenated in a sequence through which the tidy data flows, creating a pipeline for data processing and modeling. These pipelines are data-centric, meaning all the tidy tools or functions take a tidy data object as input and return a processed tidy data object, directly ready for the next operations to be applied. Also, the pipeline approach corresponds to the modular programming practice, which breaks down complex problems into smaller and more manageable pieces, as opposed to a monolithic design, where all the steps are predetermined and integrated into a single piece. The flexibility provided by the modularity makes it easier to modify certain steps in the pipeline and to maintain and extend the code base.

Examples of using a pipeline approach for data analysis can be traced back to the interactive graphics literature, including Buja et al. (1988); Sutherland et al. (2000); Xie, Hofmann, and Cheng (2014); Wickham et al. (2009). Wickham et al. (2009) argue that whether made explicit or not, a pipeline has to be presented in every graphics

program, and making them explicit is beneficial for understanding the implementation and comparing between different graphic systems. While this comment is made in the context of interactive graphics programs, it is also applicable generally to any data analysis workflow. More recently, the tidyverse suite (Wickham et al. 2019) takes the pipeline approach for general-purpose data wrangling and has gained popularity within the R community. The pipeline-style code can be directly read as a series of operations applied successively on tidy data objects, offering a method to document the data wrangling process with all the computational details for reproducibility.

Since the success of tidyverse, more packages have been developed to analyze data using the tidy framework for domains specific applications, a noticeable example of which is `tidymodels` for building machine learning models (Kuhn and Wickham 2020). To create a tidy workflow tailored to a specific domain, developers first need to identify the fundamental building blocks to create a workflow. These components are then implemented as modules, which can be combined to form the pipeline. For example, in supervised machine learning models, steps such as data splitting, model training, and model evaluation are commonly used in most workflow. In the `tidymodels`, these steps are correspondingly implemented as package `rsample`, `parsnip`, and `yardstick`, agnostic to the specific model chosen. The uniform interface in tidymodels frees analysts from recalling model-specific syntax for performing the same operation across different models, increasing the efficiency to work with different models simultaneously.

For constructing indexes, the pipeline approach adopts explicit and standalone modules that can be assembled in different ways. Index developers can choose the appropriate modules and arrange them accordingly to generate the data pipeline that is needed for their purpose. The pipeline approach provides many advantages:

- makes the computation more transparent, and thus more easily debugged.
- allows for rapidly processing new data to check how different features, like outliers, might affect the index value.
- provides the capacity to measure uncertainty by computing confidence intervals from multiple samples as generated by bootstrapping to original data.
- enables systematic comparison of surrogate indexes designed to measure the same phenomenon.
- it may even be possible to automate diagrammatic explanations and documentation of the index.

Adoption of this pipeline approach would provide uniformity to the field of index development, research, and application.

4. Details of the index pipeline

In constructing various indexes, the primary aim is to transform the data, often multivariate, into a univariate index. Spatial and temporal considerations are also factored into the process when observational units and time periods are not independent. However, despite the variations in contextual information for indexes in different fields, the underlying statistical methodology remains consistent across diverse domains. Each index can be represented as a series of modular statistical operations on the data. This allows us to decompose the index construction process into a unified pipeline workflow with a standardized set of data processing steps to be applied across different indexes.

An overview of the pipeline is presented in Figure 1, illustrating the nine available modules designed to obtain the index from the data. These modules include operations for temporal and spatial aggregation, variable transformation and combination, distribution fitting, benchmark setting, and index communication. Analysts have the flexibility to construct indexes by connecting modules according to their preferences.

Now, we introduce the notation used for describing pipeline modules. Consider a multivariate spatio-temporal process,

$$\mathbf{x}(s;t) = \{x_1(s;t), x_2(s;t), \cdots, x_p(s;t)\} \qquad s \in D_s \subseteq \mathbb{R}^m, t \in D_t \subseteq \mathbb{R}^n \tag{1}$$

where:

- $x_j(s,t)$ represents a variable of interest for example precipitation, $j = 1, \cdots, p$ and
- s represents the geographic locations in the space $D_s \subseteq \mathbb{R}^m$. Examples of geographic locations include a collection of countries, longitude and latitude coordinates or regions of interest and,
- t denotes the temporal order in $D_t \subseteq \mathbb{R}^n$. For instance, time measurements could be recorded hourly, yearly, monthly, quarterly, or by season.

In what follows when geographic or temporal components of the $x_j(s,t)$ process are fixed we will be using suffix notation. For example, $x_{sj}(t)$ represents the data for a fixed location s as a function of time t. While $x_{tj}(s)$ denotes the spatial varying process for a fixed t. An overview of the notation for pipeline input, operation, and output is present in Table 1.

Table 1. An notation overview of the input, operation, and output of each pipeline module.

Section	Module	Input	Operation	Output
3.1	Temporal processing	$x_{sj}(t)$	$f[x_{sj}(t)]$	$x_{sj}^{\text{Temp}}(t') \quad t' \in D_{t'}$
3.2	Spatial processing	$x_{tj}(s)$	$g[x_{tj}(s)]$	$x_{tj}^{\text{Spat}}(s') \quad s' \in D_{s'}$
3.3	Variable transformation	$x_j(s;t)$	$T[x_j(s;t)]$	$x_j^{\text{Trans}}(s;t)$
3.4	Scaling	$x_j(s;t)$	$[x_j(s;t) - \alpha]/\gamma$	$x_j^{\text{Scale}}(s;t)$
3.5	Dimension reduction	$\mathbf{x}(s;t)$	$h[\mathbf{x}(s;t)]$	$\mathbf{y}(s;t) \quad \mathbf{y} \subseteq \mathbb{R}^d, d < q$
3.6	Distribution fit	$x_j(s;t)$	$F[x_j(s;t)]$	$p_j(s;t) \quad p(.) \in [0,1]$
3.7	Normalising	$x_j(s;t)$	$\Phi^{-1}[x_j(s;t)]$	$z_j(s;t)$
3.8	Benchmarking	$x_j(s;t)$	$u[x_j(s;t)]$	$b_j(s;t)$
3.9	Simplification	$x_j(s;t)$	$v[x_j(s;t)]$	$A_j(s;t) \in \{a_1, a_2, \cdots, a_j\}$

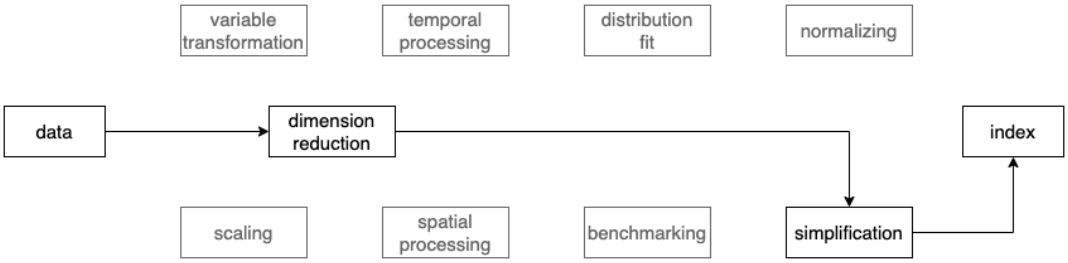

Figure 1. Diagram of pipeline modules for index construction. The highlighted path illustrates one possible construction using the dimension reduction and simplification module.

4.1. *Temporal processing*

The temporal processing module takes as input argument a single variable $x_{sj}(t)$ at location s as a function of time. In this step the original time series can be transformed or summarized into a new one via time aggregation, the transformation is represented by the function f, $x_{sj}^{\text{Temp}}(t') = f[x_{sj}(t)]$ where t' refers to the new temporal resolution after aggregation. An example of temporal processing done in the computation of the Standardized Precipitation Index (SPI) (McKee et al. 1993), consists of summing the monthly precipitation series over a rolling time window of size k. That is also known as the time scale. For SPI, the choice of the time scale k is used to control the accumulation period for the water deficit, enabling the assessment of drought severity across various types (meteorological, agricultural, and hydrological).

4.2. *Spatial processing*

The spatial processing module takes a single variable with a fixed temporal dimension, $x_{tj}(s)$, as input. This step transforms the variable from the original spatial dimension s into the new dimension $s' \in D_{s'}$ through $x_{tj}^{\text{Spat}}(s') = g[x_{tj}(s)]$. The change of spatial dimension allows for the alignment of variables collected from different measurements, such as in-situ stations and satellite imagery, or originating from different resolutions. This also includes the aggregation of variables into different levels, such as city, state, and country scales.

4.3. *Variable transformation*

Variable transformation takes the input of a single variable $x_j(s;t)$ and reshapes its distribution using the function $T[.]$ to produce $x_j^{\text{Trans}}(s;t)$. When a variable has a skewed distribution, transformations such as log, square root, or cubic root can adjust the distribution towards normality. For example, in Human Development Index (HDI), a logarithmic transformation is applied to the variable Gross National Income per capita (GNI), to reduce its impact on HDI, particularly for countries with high GNI values.

4.4. *Scaling*

Unlike variable transformation, scaling maintains the distributional shape of the variable. It includes techniques such as centering, z-score standardization, and min-max

standardization and can be expressed as $[x_j(s;t) - \alpha]/\gamma$. In Human Development Index (HDI), the three dimensions (health, education, and economy) are converted into the same scale (0-1) using min-max standardization.

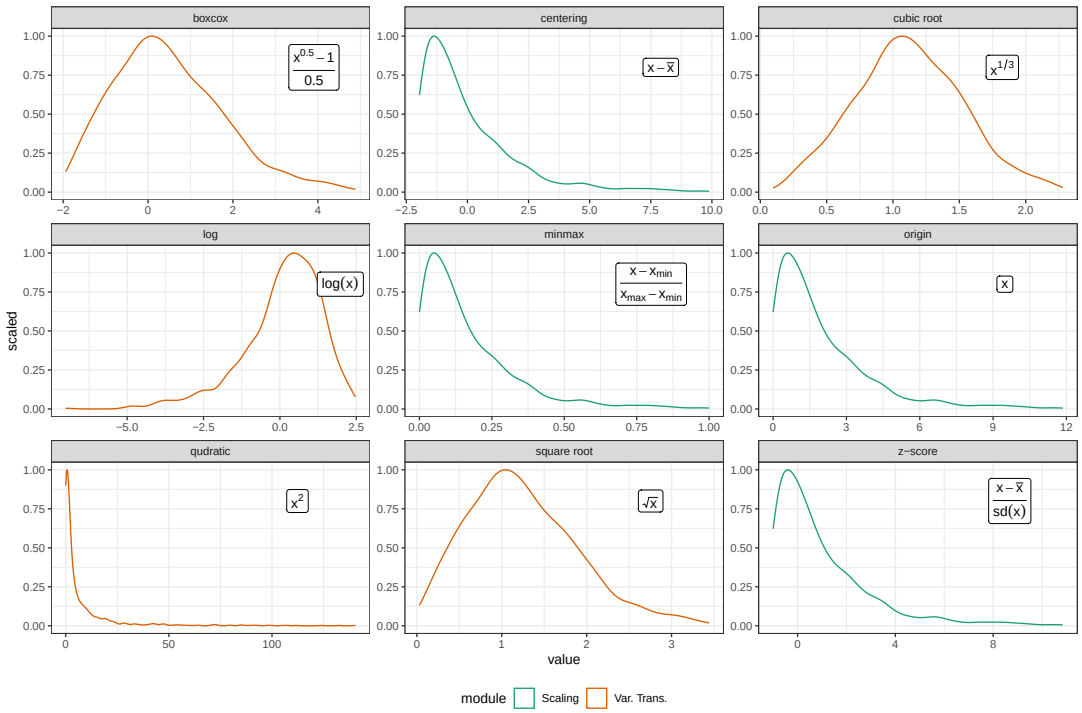

Figure 2. Comparison of the module scaling (green) and variable transformation (orange). While both modules change the variable range, scaling maintains the same distributional shape, which is not the case with variable transformation.

4.5. *Dimension reduction*

Dimension reduction takes the multivariate information $\mathbf{x}(s;t)$, where $\mathbf{x} \subseteq \mathbb{R}^q$, $q \leq p$, as the input. It summarises the high-dimensional information into a lower-dimension representation $\mathbf{y}(s;t)$, where $y \subseteq \mathbb{R}^d$ and $d < q$, as the output. The transformation can be based on domain-specific knowledge, originating from theories describing the underlying physical processes, or guided by statistical methods. For example, the Standardized Precipitation-Evapotranspiration Index (SPEI) (Beguería and Vicente-Serrano 2017) calculates the difference D between precipitation (P) and potential evapotranspiration (PET), using a water balance model ($D = P - \text{PET}$). This is the only step that differs from the Standardized Precipitation Index (SPI).

4.6. *Distribution fit*

Distribution fit applies the Cumulative Distribution Function (CDF) F of a distribution on the variable $x_j(s;t)$ to obtain the probability values $p_j(s;t) \in [0, 1]$. In SPEI, many distributions, including log-logistic, Pearson III, lognormal, and general extreme distribution, are candidates for the aggregated series. Different fitting methods and different goodness of fit tests may be used to compare the distribution choice on the index value.

4.7. *Normalising*

Normalizing applies the inverse normal CDF Φ^{-1} on the input data to obtain the normal density $z_j(s;t)$. Normalizing can sometimes be confused with the scaling or variable transformation module, which does not involve using a normal distribution to transform the variable. It is arguably whether normalizing and distribution fit should be combined or separated into two modules. A decision has been made to separate them into two modules given the different types of output each module presents (probability values for distribution fit and normal density values for normalizing).

4.8. *Benchmarking*

Benchmark sets a value $b_j(s,t)$ for comparing against the original variable $x_j(s;t)$. This benchmark can be a fixed value consistently across space and time or determined by the data through the function $u[x_j(s;t)]$. Once a benchmark is set, observations can be highlighted for adjustments in other modules or can serve as targets for monitoring and planning.

4.9. *Simplification*

Simplification takes a continuous variable $x_j(s;t)$ and categorises it into a discrete set $A_j(s;t) \in \{a_1, a_2, \cdots, a_j\}$ through a piecewise function,

$$v[x_i(s;t)] = \begin{cases} a_0 & C_1 \leq x^i(s;t) < C_0 \\ a_1 & C_2 \leq x^i(s;t) < C_1 \\ a_2 & C_3 \leq x^i(s;t) < C_2 \\ \cdots & \\ a_z & C_z \leq x^i(s;t) \end{cases} \tag{2}$$

This is typically used at the end of the index pipeline to simplify the index to communicate to the public the severity of the concept of interest measured by the index. An example of simplification is to map the calculated SPI to four categories: mild, moderate, severe, and extreme drought.

5. Software design

The R package `tidyindex` implements a proof-of-concept of the index pipeline modules described in Section 4. These modules compute an index in a sequential manner, as shown below:

```
DATA |>
  module1(...) |>
  module2(...) |>
  module3(...) |>
  ...
```

Each module offers a variety of alternatives, each represented by a distinct function. For example, within the `dimension_reduction()` module, three methods are available: `aggregate_linear()`, `aggregate_geometrical()`, and `manual_input()` and they can be used as:

```
dimension_reduction(V1 = aggregate_linear(...))
dimension_reduction(V2 = aggregate_geometrical(...))
dimension_reduction(V3 = manual_input(...))
```

Each method can be independently evaluated as a recipe, for example,

```
manual_input(~x1 + x2)
```

takes a formula to combine the variables `x1` and `x2` and return:

```
[1] "manual_input"
attr(,"formula")
[1] "x1 + x2"
attr(,"class")
[1] "dim_red"
```

This recipe will then be evaluated in the pipeline module with data to obtain numerical results. The package also offers wrapper functions that combine multiple steps for specific indexes. For instance, the `idx_spi()` function bundles three steps (temporal aggregation, distribution fit, and normalizing) into a single command, simplifying the syntax for computation. Analysts are also encouraged to create customized indexes from existing modules.

```
idx_spi <- function(...){
  DATA |>
    aggregate(...) |>
    dist_fit(...) |>
    augment(...)
}
```

(more changes) The accompanied package, tidyindex, is not intended to offer an exhaustive implementation for all indexes across every domains. Instead, it provides a realization of the pipeline framework proposed in the paper. When adopting the pipeline approach to construct indexes, analysts may consider developing software that can be readily deployed in the cloud for production purposes.

6. Examples

This section uses the example of drought and social indexes to show the analysis made possible with the index pipeline. The drought index example computes two indexes (SPI and SPEI) with various time scales and distributions simultaneously using the pipeline framework to understand the flood and drought events in Queensland. The second example focuses on the dimension reduction step in the Global Gender Gap Index to explore how the changes in linear combination weights affect the index values and country rankings.

6.1. *Every distribution, every scale, every index all at once*

The state of Queensland in Australia frequently experiences natural disaster events such as flood and drought, which can significantly impact its agricultural industry. This example uses daily data from Global Historical Climatology Network Daily (GHCND), aggregated into monthly precipitation, to compute two drought indexes – SPI and SPEI – at various time scales and fitted distributions, for 29 stations in the state of Queensland in Australia, spanning from January 1990 to April 2022. This example showcases the basic calculation of indexes with different parameter specifications within the pipeline framework. The dataset used in this example is available in the `tidyindex` package as `queensland` and blow prints the first few rows of the data:

```
# A tibble: 5 x 9
  id                ym  prcp  tmax  tmin  tavg  long   lat name
  <chr>          <mth> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 ASN00029038 1990 Jan  1682  34.3  24.7  29.5  142. -15.5 KOWANYAMA ~
2 ASN00029038 1990 Feb   416  35.2  23.2  29.2  142. -15.5 KOWANYAMA ~
3 ASN00029038 1990 Mar  2026  32.5  23.6  28.0  142. -15.5 KOWANYAMA ~
4 ASN00029038 1990 Apr   597  32.9  17.7  25.3  142. -15.5 KOWANYAMA ~
5 ASN00029038 1990 May   244  31.8  20.1  25.9  142. -15.5 KOWANYAMA ~
```

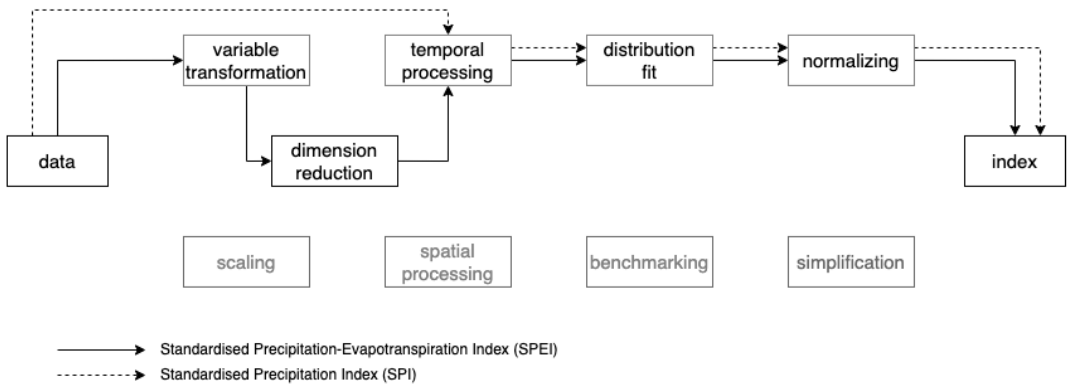

Figure 3. Index pipeline for two drought indexes: the Standardized Precipitation Index (SPI) and the Standardized Precipitation-Evapotranspiration Index (SPEI). Both indexes share similar construction steps with SPEI having two steps additional steps (variable transformation and dimension reduction) to convert temperature into evapotranspiration and combine it with the precipitation series.

Figure 3 illustrates the pipeline steps of the two indexes. The two indexes are similar with the distinct that SPEI involves two additional steps – variable transformation and dimension reduction – prior to temporal processing. As introduced in Section 5, wrapper functions are available for both indexes as `idx_spi()` and `idx_spei()`, which allows for the specification of different time scales and distributions for fitting the aggregated series. In `tidyindex`, multiple indexes can be calculated collectively using the function `compute_indexes()`. Both SPI and SPEI are calculated across four time scales (6, 12, 24, and 36 months). The SPEI is fitted with two distributions (log-logistic and general extreme value distribution) and the gamma distribution is used for SPI:

```
.scale <- c(6, 12, 24, 36)
idx <- queensland %>%
  init(id = id, time = ym) %>%
  compute_indexes(
    spei = idx_spei(
      .pet_method = "thornthwaite", .tavg = tavg, .lat = lat,
      .scale = .scale, .dist = c(gev(), loglogistic())),
    spi = idx_spi(.scale = .scale)
  )
```

```
# A tibble: 128,586 x 19
   .idx  .period id             ym  prcp  tmax  tmin  tavg  long   lat
   <chr>   <dbl> <chr>       <mth> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 spei        6 ASN0002~ 1990 Jun   170  29.7  16.2  23.0  142. -15.5
 2 spei        6 ASN0002~ 1990 Jun   170  29.7  16.2  23.0  142. -15.5
 3 spei        6 ASN0002~ 1990 Jun     0  23.0  11.8  17.4  139. -20.7
 4 spei        6 ASN0002~ 1990 Jun     0  23.0  11.8  17.4  139. -20.7
 5 spei        6 ASN0003~ 1990 Jun   794  25.8  18.1  21.9  146. -16.9
 6 spei        6 ASN0003~ 1990 Jun   794  25.8  18.1  21.9  146. -16.9
 7 spei        6 ASN0003~ 1990 Jun   504  23.0  13.8  18.4  145. -17.1
 8 spei        6 ASN0003~ 1990 Jun   504  23.0  13.8  18.4  145. -17.1
 9 spei        6 ASN0003~ 1990 Jun  1970  23.9  16.4  20.2  146. -17.6
10 spei        6 ASN0003~ 1990 Jun  1970  23.9  16.4  20.2  146. -17.6
# i 128,576 more rows
# i 9 more variables: name <chr>, pet <dbl>, diff <dbl>,
#   .scale <dbl>, .agg <dbl>, .method <chr>, .fitted <dbl>,
#   .dist <chr>, .index <dbl>
```

The output contains the original data, index values (`.index`), parameters used (`.scale`, `.method`, and `.dist`), and all the intermediate variables (`pet`, `.agg`, and `.fitted`). This data can be visualized to investigate the spatio-temporal distribution of the drought/flood events, as well as the response of index values to different time scales and distribution parameters at specific single locations. Figure 4 and Figure 5 exemplify two possibilities. Figure 4 presents the spatial distribution of SPI during two periods: October 2010 to March 2011 for the 2010/11 Queensland flood and October 2019 to March 2020 for the 2019 Australia drought, which contributes to the notorious 2019/20 bushfire. Figure 5 displays the sensitivity of the SPEI series at the Texas post office to different time scales and fitted distributions. Larger time scales produce a smoother index across time, however, all time scales indicate an extreme drought (corresponding to -2 in SPEI) in 2020, confirming the severity of the drought across different time horizons. Moreover, the chosen distribution has less influence on the index, with general extreme value distribution tending to produce more extreme outcomes than log-logistic distribution for the extreme events (index $>$ 2 or $<$-2).

6.2. *Does a minor change in variable weights cause a tornado?*

The Global Gender Gap Index (GGGI), published annually by the World Economic Forum, measures gender parity by assessing relative gaps between men and women in four key areas: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment (World Economic Forum 2023). The

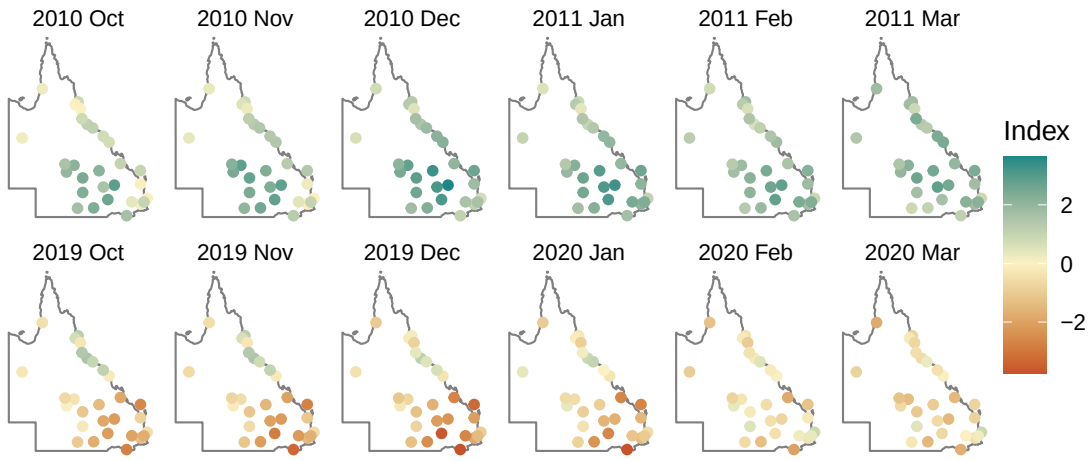

Figure 4. Spatial distribution of Standardized Precipitation Index (SPI-12) in Queensland, Australia during two major flood and drought events: 2010/11 and 2019/20. The map shows a continuous wet period during the 2010/11 flood period and a mitigated drought situation, after its worst in 2019 December and 2020 January, likely due to the increased rainfall in February from the meteorological record.

index, compiled from 14 variables expressed as female-to-male ratios, first aggregates these variables into four dimensions through linear combinations. These dimensions are then combined through another linear combination to form the index. Figure 6 illustrates this pipeline construction by applying the dimension reduction module twice on the data to generate the index. Table 2 presents the variable weights (`V-weight`) and dimension weights (`D-weight`) used in the two dimension reduction steps. In the table, the variable weights are computed as the inverse of the standard deviation of each variable, scaled to sum to 1. These weights ensure that a one percentage point change in the standard deviation of each variable contributes equally to the index. Dimension weights are equal across the four dimensions and the last column, `weight`, multiples the variable and dimension weights to produce a single set of weights.

Table 2. Weights used to compute the Global Gender Gap Index

Variable	Dimension	V-weight	D-weight	Weight
Labour force participation	Economy	0.199	0.25	0.050
Wage equality for similar work	Economy	0.310	0.25	0.078
Estimated earned income	Economy	0.221	0.25	0.055
Legislators senior officials and managers	Economy	0.149	0.25	0.037
Professional and technical workers	Economy	0.121	0.25	0.030
Literacy rate	Education	0.191	0.25	0.048
Enrolment in primary education	Education	0.459	0.25	0.115
Enrolment in secondary education	Education	0.230	0.25	0.058
Enrolment in tertiary education	Education	0.121	0.25	0.030

Variable	Dimension	V-weight	D-weight	Weight
Sex ratio at birth	Health	0.693	0.25	0.173
Healthy life expectancy	Health	0.307	0.25	0.077
Women in parliament	Politics	0.310	0.25	0.078
Women in ministerial positions	Politics	0.247	0.25	0.062
Years with female head of state	Politics	0.443	0.25	0.111

The 2023 GGGI data is available from the Global Gender Gap Report 2023 in the country's economy profile and can be accessed in R via the `tidyindex` package as `gggi` and Table 2 as `gggi_weights`. The index can be reproduced with the package as:

```
gggi %>%
  init(id = country) %>%
  add_meta(gggi_weights, var_col = variable) %>%
  dimension_reduction(
    index_new = aggregate_linear(
      ~labour_force_participation:years_with_female_head_of_state,
      weight = weight))
```

After initialzing the `gggi` object and attaching the `gggi_weights` as metadata, a single linear combination step within the dimension reduction module is applied to the 14 variables (from column `labour_force_participation` to `years_with_female_head_of_state`), using the weight specified in the `weight` column of the attached metadata. While computing the index from the original 14 variables, it remains unclear how the missing values are handled within the index, which impacts 68 out of the total 146 countries. However, after aggregating variables into the four dimensions, where no missing values exist, the index is reproducible for all the countries.

A linear combination can also be interpreted as a linear projection of multivariate information with a weight vector. Projecting data from higher to lower dimensions unavoidably leads to information loss and the weights used require careful examination to understand their effects on the index and implications for interpretation and decision-making. By making slight adjustments to the weight vector, we can observe how index values and country rankings change. For illustration, we select a set of countries including:

1) Top-ranked countries with GGGI > 0.85,
2) Countries ranked between 57 and 62, with GGGI values from 0.72 to 0.73, and
3) Low-ranked countries with GGGI < 0.6.

We slightly vary the weight of the politics dimension from the original 0.25 while keeping the weights constant for the other three dimensions. This process creates an animation showing the movement of index values in response to changing weights. This visualization technique, which presents a sequence of data projections, is referred to as a "tour". The specific kind of tour that moves between the original and nearby projections is known as a "local tour".

In Figure 7, six frames have been chosen from the animation available at https://vimeo.com/847874016. When the weight of politics is reduced (Frame 1 and 6 vs. Frame 12), the difference in GGGI values between the top Nordic countries and mid- or low-ranked countries narrows, suggesting a smaller variation among countries

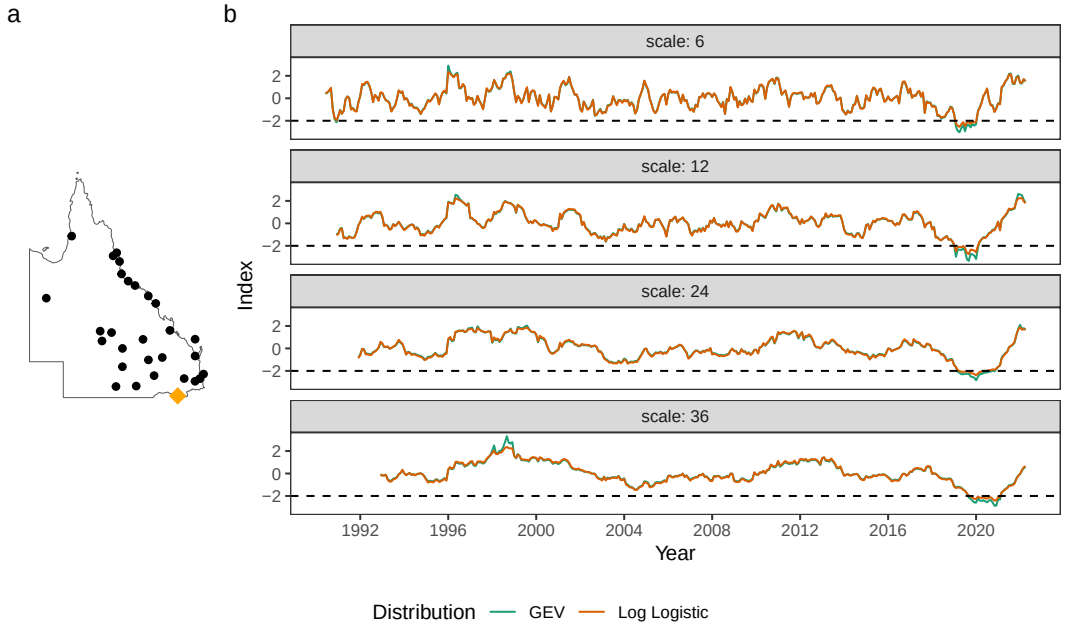

Figure 5. Time series plot of Standardized Precipitation-Evapotranspiration Index (SPEI) at the Texas post office station (highlighted by a diamond shape in panel a). The SPEI is calculated at four time scales (6, 12, 24, and 36 months) and fitted with two distributions (Log Logistic and GEV). The dashed line at -2 represents the class "extreme drought" by the SPEI. A larger time scale gives a smoother index series, while also taking longer to recover from an extreme situation as seen in the 2019/20 drought period. The SPEI values from the two distribution fit mostly agree, while GEV can result in more extreme values, i.e. in 1998 and 2020.

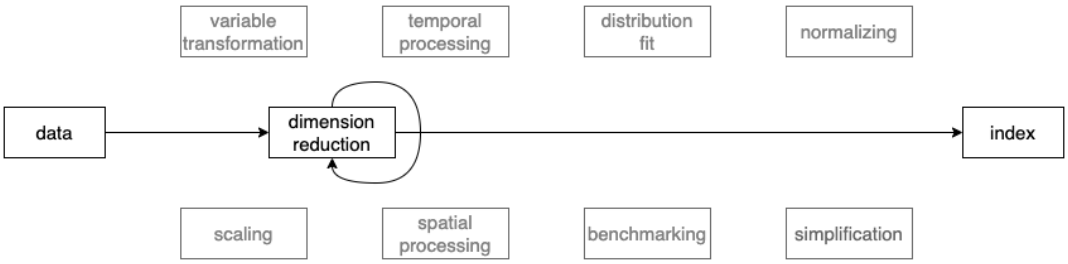

Figure 6. Index pipeline for the Global Gender Gap Index (GGGI). The index is constructed as applying the module dimension reduction twice on the data.

in achieving gender parity. Conversely, when the weight of politics increases (Frame 18, 24, and 29 vs. Frame 12), nearly all the countries in the three categories experience a decrease in GGGI values. A noticeable exception to this trend is Bangladesh, where its index value moves in the same direction as the politics weight. This leads to its index value being almost similar to those of the top-ranked Nordic countries when the politics dimension is assigned a weight of 0.52 in Frame 29.

In GGGI, the index value has a direct interpretation as the percentage of the gender gap that has been closed. When countries exhibit closer or consistent decreasing index values, it can be directly interpreted as a reflection of a country's progress toward gender parity. Ideally, an index should be robust against minor changes in its construction components. This example provides researchers with means to observe changes in the index resulting from variations in the weights used in linear combinations. This approach can be applied broadly to different sets of weights of interest, extending beyond the change in a single dimension illustrated by the example.

TODO: increase font for variable names

6.3. *Confidence interval [TO BE CHANGED]*

Index uncertainties may arise from various stages of the pipeline, including from different parameterisation choices, as illustrated from Section 6.1, or from the statistical procedures in the pipeline. Given the lack of direct instruments for measurement, true values for indexes can never be obtained to compare the index against. Nevertheless, an index can be provided with confidence interval to report it accuracy. In this example, the Texas post office station highlighted in Figure 5 is used to illustrate the calculation of confidence interval for the Standardized Precipitation Index (SPI). Bootstrapping is used to account for the sampling uncertainty in the distribution fit step of the index pipeline and to assess its impact on the SPI series.

In SPI, the distribution fit step fits the gamma distribution to the aggregated precipitation series separately for each month. This results in 31 or 32 points, from 1990 Jan to 2022 Apr, for estimating each set of distribution parameters. To account for this sampling uncertainty with these samples, bootstrapping is used to generate replicates of the aggregated series. In the `tidyindex` package, this bootstrap sampling is activated when the argument `.n_boot` is set to a value other than the default of 1. In the following code, the Standardized Precipitation Index (SPI) is calculated using a time scale of 24. The bootstrap procedure samples the aggregated precipitation (`.agg`) for 100 iterations (`.n_boot = 100`) and then fits the gamma distribution. The resulting gamma probabilities are then transformed into normal densities in the normalizing step with `agument()`.

```
DATA %>%
  aggregate(.var = prcp, .scale = 24) %>%
  dist_fit(.dist = gamma(), .var = .agg, .n_boot = 100) %>%
  augment(.var = .agg)
```

The confidence interval can then be calculated using the quantile method from the bootstrap samples. Figure 8 presents the 80% and 95% confidence interval for the Texas post office station, in Queensland, Australia. From the start of 2019 to 2020, the majority of the confidence intervals lie below the extreme drought line (SPI = -2), suggesting a high level of certainty that the Texas post office is suffering from a drastic drought. The

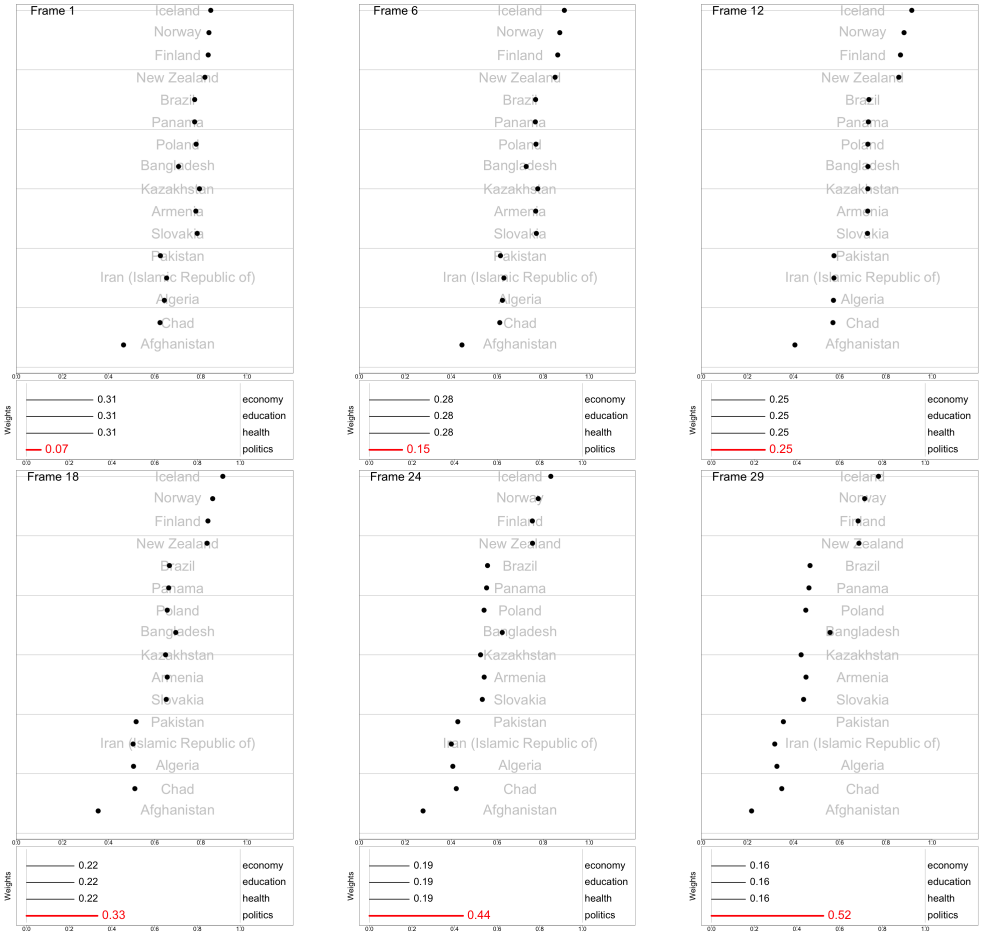

Figure 7. Six frames selected to explore how varying the weights of the politics dimension changes the index values and country rankings in Global Gender Gap Index (GGGI). The top panel shows the GGGI value against the country, ranked by its original index value in Frame 12. The bottom panel displays the weight used to produce the index values in the top panel, with each frame corresponds to a set of weights. Countries selected includes 1) top-ranked countries with GGGI > 0.85, 2) countries ranked between 57 and 62 with GGGI from 0.72 to 0.73, and 3) low-ranked countries with GGGI < 0.6. Compared to Frame 12 where equal weigts are used for the four dimension, a reduced weight in politics (Frame 1 and 6) shows narrower gaps between top and mid- or lower-ranked countries, while an increase in the politics weight (Frame 18, 24, and 29) leads to a systematic decrease of GGGI values across all the countries, except for Bangladesh. Full animation is available at https://vimeo.com/847874016.

relatively wide confidence interval, as well as during the excessive precipitation events in 1996-1998 and 1999-2000, suggests a high variation of the gamma parameters estimated from the bootstrap samples and its difficulty to accurately quantifying the drought/ flood severity in extreme events.

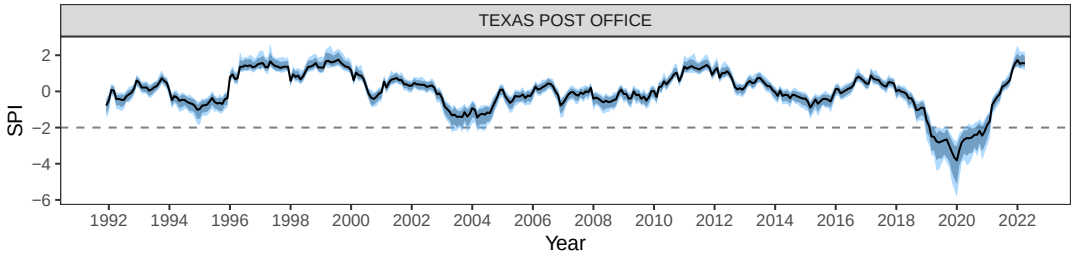

Figure 8. 80% and 95% confidence interval of the Standardized Precipitation Index (SPI-24) for the Texas post office station, in Queensland, Australia. A bootstrap sample of 100 is taken from the aggregated pricipitation series to estimate gamma parameters and to calculate the index. The dashed line at SPI = -2 represents an extreme drought as defined by the SPI. Most parts of the confidence intervals from 2019 to 2020 sit below the extreme drought line (-2) and are relatively wide compared to other time periods. This suggests that while it is certain that the Texas post office is suffering from a drastic drought, there is considerable uncertainty in quantifying its severity, given the extremity of the event.

7. Conclusion

The paper introduces a data pipeline comprising nine modules designed for the construction and analysis of indexes within the tidy framework. The pipeline offers a modular workflow to allow compute index with different parameterizations, to test minor changes to the original index definition, and to quantify uncertainties. The framework proposed in the paper is universal to index across diverse domains. Examples are provided, including the drought indexes (SPI and SPEI) and Global Gender Gap Index (GGGI), to demonstrate the index calculation with different time scales and distributions, to illustrate how slight adjustment of linear combination weights impact the index, and to calculate confidence intervals on the index.

8. Acknowledgement

This work is funded by a Commonwealth Scientific and Industrial Research Organisation (CSIRO) Data61 Scholarship. The article is created using Quarto (Allaire et al. 2022) in R (R Core Team 2021). The source code for reproducing this paper can be found at: https://github.com/huizezhang-sherry/paper-tidyindex.

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.

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. *Quarto* (version 1.2). https://doi.org/10.5281/zenodo.5960048.

Becker, William, Giulio Caperna, Maria Del Sorbo, Hedvig Norlen, Eleni Papadimitriou, and Michaela Saisana. 2022. "COINr: An r Package for Developing Composite Indicators." *Journal of Open Source Software* 7 (78): 4567. https://doi.org/10.21105/joss.04567.

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.

Grenié, Matthias, and Hugo Gruson. 2023. *fundiversity: Easy Computation of Functional Diversity Indices*. https://doi.org/10.5281/zenodo.4761754.

Hao, Zengchao, and Vijay P. Singh. 2015. "Drought Characterization from a Multivariate Perspective: A Review." *Journal of Hydrology* 527 (August): 668–78. https://doi.org/10.1016/j.jhydrol.2015.05.031.

Jones, Brenda, and Jean Andrey. 2007. "Vulnerability Index Construction: Methodological Choices and Their Influence on Identifying Vulnerable Neighbourhoods." *International Journal of Emergency Management* 4 (2): 269–95. https://doi.org/10.1504/IJEM.2007.013994.

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

Martin, Steve. 2023. *Gpindex: Generalized Price and Quantity Indexes*. https://CRAN.R-project.org/package=gpindex.

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.

OECD, European Union, and Joint Research Centre - European Commission. 2008. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD. https://doi.org/10.1787/9789264043466-en.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Saisana, M., A. Saltelli, and S. Tarantola. 2005. "Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators." *Journal of the Royal Statistical Society Series A: Statistics in Society* 168 (2): 307–23. https://doi.org/10.1111/j.1467-985X.2005.00350.x.

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.

Svoboda, Mark, Brian Fuchs, et al. 2016. "Handbook of Drought Indicators and Indices." *Drought and Water Crises: Integrating Science, Management, and Policy*, 155–208.

Tate, Eric. 2012. "Social Vulnerability Indices: A Comparative Assessment Using Uncertainty and Sensitivity Analysis." *Natural Hazards* 63 (2): 325–47. https://doi.org/10.1007/s11069-012-0152-2.

Vicente-Serrano, Sergio M., Santiago Beguería, and Juan I. López-Moreno. 2010. "A Multiscalar Drought Index Sensitive to Global Warming: The Standardized

Precipitation Evapotranspiration Index." *Journal of Climate* 23 (7): 1696–1718. https://journals.ametsoc.org/view/journals/clim/23/7/2009jcli2909.1.xml.

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.

World Economic Forum. 2023. "The Global Gender Gap Report 2023." https://www3.weforum.org/docs/WEF_GGGR_2023.pdf.

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.

Zargar, Amin, Rehan Sadiq, Bahman Naser, and Faisal I Khan. 2011. "A Review of Drought Indices." *Environmental Reviews* 19 (NA): 333–49. https://www.jstor.org/stable/envirevi.19.333.