SCC Client Report
Multivariate Statistical Methods for Computing a Small–Scale Fisheries Vulnerability Index

December 7, 2023

1 Background

Small-Scale Fisheries (SSFs) are crucial to the livelihoods and sustenance of a significant portion of the global population. At least 492 million people are partially dependent on SSFs, and these fisheries account for 90% of global capture fisheries employment and 37 million tons of catch, representing 40% of the global fisheries catch (FAO, 2023). The coexistence and often overlap of SSF and Large-Scale Fisheries (LSF) operations is another notable issue - namely, 83.7% of SSF catch occurs within 20 kilometers from the shore, a zone where large-scale fishing activities also take place. This proximity leads to potentially harmful interactions, with anecdotal reports of conflicts between these two sectors. You plan to utilize data from Global Fishing Watch (industrial fishing effort), Illuminating Hidden Harvests (SSF data), and national indicators like governance scores and the World Bank data to create a country-level SSF vulnerability index. These data provide insights into the intensity of nearshore large-scale fishing activity, the dependence of nations on small-scale fisheries, and the nutritional supply available from SSFs. The data also delves into the vulnerability of small-scale fisheries to large-scale operations, considering factors like exposure to LSF, sensitivity of SSFs, and the adaptive capacity of nations. The approach includes assessing the nation's ability to mediate the impacts of LSF and SSF interactions, looking at development levels and governance quality - providing deeper insights into the intricate relationship between SSFs and LSFs. Our goal is to determine an appropriate country-level LSF vulnerability index that can assist the client in determining the potential food security impacts stemming from LSF and SSF spatial overlap.

2 Recommendations

2.1 Overview

In constructing a SSF Index, this report's recommendations outline several different weighting schema for combining the data, as well as highlighting variable choice, transformation, and index evaluation. We describe advanced statistical methods such as Principal Components Analysis (PCA), Factor Analysis (FA), Structural Equation Models (SEMs), and Path Models (PM) as tools to determine variable weights in a composite index, while also highlighting the advantages

and disadvantages of each. PCA serves to reduce dimensionality, revealing correlated variables' underlying patterns, while FA identifies latent variables, aiding in data reduction and simplification. Additionally, we advocate for considering a nation's ability to mediate large-scale and small-scale fishing interactions using SEMs or PMs. SEMs provide a robust framework for examining interactions/relationships between latent variables, which offers versatility in analyses. PMs, a specific type of SEM, are especially useful when the focus of an analysis is more on prediction rather than hypothesis testing. PMs operate by fitting a composite model and representing variables as a linear combination of their components, which makes using PMs advantageous for small sample sizes, missing data, or non-normal distributions. Our recommendations reflect the importance of data quality, correlation between variables, and adherence to assumptions for optimal results. This report also includes a discussion of operational issues that may arise as a result of these methods, and ways to remediate those issues. Ultimately, these advanced statistical methods can help guide the creation of a comprehensive SSF Vulnerability Index, while still making sure nuances and anomalies in the data are accounted for – crucial for effective fisheries management.

2.2 Composite Index Weighting via Principal Components Analysis (PCA)

2.2.1 High-level Description of Principal Components Analysis

Principal components analysis ("PCA") is a statistical method useful for high-dimensional data that can transform originally correlated variables into a new smaller set of uncorrelated variables ("principal components"), depending on the desired amount of variance to be explained (Nardo et al., 2005). Each principal component captures an independent source of variance in the data, and when there is high correlation among the original variables, a high proportion of the original variance can be represented with few principal components (Nardo et al., 2005). Proper use of PCA requires that the original variables are correlated, centered, typically standardized and a linear relationship exists (Nardo et al., 2005). PCA can help reduce dimensionality, reveal which variables share a significant proportion of the common variance, aid feature extraction and thus help increase interpretability of high-dimensional data (Budaev, 2010). However, pitfalls exist when there is only a marginal dimensionality reduction, significant outliers exist, the principal components lack interpretability / intuitive meanings and the runtime of calculating these principal components is high (Nardo et al., 2005).

2.2.2 Guidelines for Using PCA to Determine Variable Weights in a Composite Index

Principal Components Analysis can be implemented simply in R using the prcomp() function from the 'stats' package, in addition to calculating manually. This function uses Singular Value Decomposition to return the corresponding standard deviations of the principal components as well as the final matrix with eigenvector columns corresponding to the weights of our co-variates (RDocumentation, 2019).

When calculating a composite index using PCA, the correct construction of variable weights can maximize the accuracy of our composite index. In cases where indices are weighted, one must confirm the validity of the data measurement and the generalizability of each component index. There must be statistical significance or predictive power between the index and external outcomes to ensure the index can be applied more generally to subsequent topics. After breaking into principal components using zero-centered, normalized data, we can generate an index from the

absolute values of PCA loadings (weightings between -1:1). We can iteratively dilate these variable weightings through the optimization process of "index mining" (Chao, 2017).

Note that after determining co-variate weights and sorting them, not all of the variance needs to be explained by the principal components. Rather, the main factors must explain enough of the variance to accurately reflect general patterns in the data (Morekonda, 2020).

2.2.3 Advantages and Potential Pitfalls / Disadvantages to Using PCA as a Tool to Construct a Composite Index

The PCA has a number of excellent mathematical properties - most importantly that the index obtained from the first principal component explains the largest portion of variance in the data. When these PCA indices capture a large amount of the variance and make analysis more interpretable it is an extremely useful tool. However, when the first principal component accounts for a limited part of the variance in the data, we can lose a significant amount of information since the first principal component is often the only component that is used. Moreover, the PCA based index is often elitist, with a strong tendency to represent highly intercorrelated indicators and to neglect the others, irrespective of their possible contextual importance (Mazziotta and Pareto, 2015). This poses a significant challenge for PCA-based asset indices, emphasizing the importance of ensuring a diverse range of asset variables to prevent issues related to clumping or truncation (Vyas and Kumaranayake, 2006). In addition, it is often the case that the weights assigned to variables in PCA might be similar, leading to the equal weighting of variables in the composite index that may lack a clear and intuitive interpretation in terms of the original variables. Finally, PCA has a linearity assumption, meaning that there must be a linear relationship among the variables (which is not always the case). Although PCA is quite useful in its dimensionality reduction, the procedure is by no means 'perfect' for deriving composite development indices and is often very case specific (Ram, 1982).

2.3 Determining Index Weights via Factor Analysis

2.3.1 High-level Description of Exploratory and Confirmatory Factor Analysis

Factor Analysis shares the same goals as Principal Component Analysis (PCA). It is used to define a weighting scheme to form composite indicators. Like PCA, Factor Analysis has a linearity assumption. Unlike PCA, it is based on a linear model. Nardo et. al (2005) describe the goal as to "account for the highest possible variation in the indicators set using the smallest possible number of factors" and the paper outlines the linear model and its respective variables in detail. The main goal of PCA is to identify new factors that explain your variables and to be able to represent your original variables using these factors.

There are two main types of Factor Analysis. Exploratory Factor Analysis (EFA) is used as to identify latent variables, which are hidden or unmeasured variables, and as a data reduction technique. Confirmatory Factor Analysis (CFA) is commonly utilized after EFA to assess the validity of the measures and improve the factor model.

EFA aims to identify a subset of factors that can explain the most covariance from the measured data. It is a data reduction technique because it can identify correlated sub-indicators to correct for overlapping information. In general, you start with your covariance matrix, identify the amount of factors to retain, and then rotate factors for interpretability. Nardo et. al (2005) explains each step of this procedure in detail.

Unlike EFA, CFA starts with a hypothesized model for weights. In general, you specify your model and then choose how to assess its goodness of fit. Floyd and Widaman (1995) detail best practices and considerations for this procedure in detail.

2.3.2 Guidelines for Using FA to Determine Variable Weights in a Composite Index.

Factor analysis relies on several assumptions for accurate results. Violating these assumptions may lead to factors that are misleading or hard to interpret. In general, the most basic requirement for optimal use of factor analysis is high-quality data that are measured on interval or quasi-interval scales, with some correlation between variables (Floyd and Widaman, 1995). Ideally, these data would also be distributed in a multivariate normal fashion. For certain methods of parameter estimation, like maximum likelihood factor analysis, multivariate normality is a strict assumption. In contrast, others, such as the principal axes (or least squares) method–which is by far the most commonly used approach for exploratory factor analysis–do not require or assume multivariate normality (Floyd and Widaman, 1995). However, both exploratory and confirmatory factor analysis appear to be relatively robust against violations of normality in practice (Floyd and Widaman, 1995). Another key assumption in factor analysis is linearity. This means that the relationships among variables are assumed to be linear, ensuring that changes in values are consistent (Nardo et al., 2005). Finally, it is important to note that larger sample sizes ensure more stable estimates of factor structure. While explicit guidelines have always been in flux, experts typically recommend a minimum participants-to-variables ratio of between 5:1 and 10:1, and an overall sample size greater than 100-200 (Budaev, 2010). When the sample size is smaller than that, adequate solutions can still be obtained with a greater participants-to-variables ratio.

2.3.3 Advantages and Potential Pitfalls/Disadvantages to Using FA as a Tool to Construct a Composite Index

Factor analysis, particularly Exploratory Factor Analysis (EFA), is beneficial for hypothesis generation when researchers have no pre-existing hypotheses about the underlying factor structure of the data (Floyd and Widaman, 1995). By identifying the factors which are responsible for the correlations among a set of variables, EFA can reveal underlying patterns and structures in the data that aren't immediately apparent. This can be helpful in deciding what variables to include in a composite index. Additionally, EFA excels at data reduction and simplification, reducing complex datasets into more easily understandable and interpretable "super-variables" (i.e., latent variables) (Qualtrics, 2023). This is especially beneficial in fields like psychology, sociology, and market research where identifying latent constructs is vital. EFA is also particularly valuable in cross-country analyses, highlighting key elements that contribute most to variability in the dataset (Nardo et al., 2005). It assigns higher factor loadings– numerical values that indicate the strength of association between a sub-indicator (variable) and a factor– to sub-indicators that show the greatest variation across different units of analysis (i.e., countries). In this way, it efficiently pinpoints sub-indicators which are more likely to yield insights in comparative studies, as opposed to those with minimal variation (Nardo et al., 2005). In general, loading values of 0.3 or 0.4 are the recommended, but not universally agreed upon, minimum cutoff to be considered when interpreting factors (Budaev, 2010). With regard to creating a composite index, one method to consider would be *loading-based weighting*, where weights are assigned based on the factor loading of each variable (i.e., a higher loading means a higher weight for that variable).

In terms of disadvantages, factor analysis is sensitive to problems with small sample sizes, and this is particularly relevant when the focus is on a limited set of countries because factor analysis won't take into account the cultural context of different countries and generalization is difficult. This can lead to difficulties in generalizing the results and potentially misrepresenting the weights of various components in the index. It's also sensitive to the presence of outliers, because they can contribute to the identification of additional factors or obscure the presence of genuine factors, which can lead to overfitting or underfitting of the factor model. This can subsequently skew the weights assigned to different components of the index, and thus can compromise the reliability of the composite index (Nardo, 2005). Another disadvantage is that correlations don't necessarily represent the real influence of the sub-indicators on the problem being measured because it doesn't provide you much information about the direction of influence in bidirectional relationships. Such a limitation can lead to misinformed weighting decisions, where the influence of certain indicators might be either overstated or understated. Furthermore, factor analysis's assumption of linearity can be particularly problematic in constructing composite indices. Real-world data often exhibit non-linear relationships, and the presumption that all relationships are linear (Hair, 2009) can lead to inaccurate weightings of the components within the index. There is a large amount of assumptions that are made. One of these is that factor analysis assumes that there is no bias when selecting sub-indicators. This assumption can lead to skewed weightings if certain relevant indicators are overlooked or if included indicators are not representative of the dimensions they are supposed to measure. And lastly, it assumes there are strong intercorrelations, because conducting factor analysis on a dataset with very little intercorrelation means that there are almost the same amount of factors as there are original variables, which defeats the entire purpose of factor analysis. This can lead to a composite index where the weights do not accurately represent the underlying structure of the data, thus questioning the validity and utility of the index.

2.4 Structural Equation / Path Models (SEMs/PMs) for Composite Indicators/Indexes

2.4.1 High-level Description of Structural Equation and Path Models

Structural equation modeling (SEM) is a technique that allows for explanation of relationships between latent variables, or variables that cannot be directly measured (i.e. exposure or sensitivity in your case.) SEM relies on a structure of predetermined hypothesized relationships between variables; as such, it is a useful tool for comparing hypothetical models. SEM differs from more standard regression techniques as it can be used to simultaneously estimate multiple relationships. Additionally, it allows for the modeling of causal relationships and it can handle more abstract variables. For a primer on SEM, see Beran and Violato (2010a).

SEM processes can be best understood with a diagram, with arrows between variables denoting that they are related (for example, an arrow between corruption estimate and governance in your case.) SEM then estimates coefficients for each of these arrows/relationships, allowing you to quantify and compare them. For example, if the coefficient on the arrow between corruption estimate and governance was higher in magnitude than the coefficient on the arrow between rule of law and governance, we could infer that a change in corruption will have a bigger impact on governance than a change in rule of law. For examples of basic SEM diagrams, see Fan et al. (2016).

Partial least squares path modeling (PLS-PM) is a type of SEM that is often used when the

focus is more on prediction rather than hypothesis testing. Unlike other variants of SEM, it can fit a composite model rather than a factor model. A composite model is one where variables are represented as a linear combination of their component parts. Using PM can be useful for working with small sample sizes, missing data, or non-normal data distributions (Lauro et al., 2018). For more information on PMs, see Henseler et al. (2016). One should note, however, that the PLS-PM methodology for SEMs is a relatively recent development and is still viewed critically by some researchers. There is contention regarding the efficacy of its use in small sample sizes, introduction of bias, and the ad-hoc nature of how PM has been developed (Sarstedt et al., 2016).

2.4.2 Guidelines for Using SEMs/PMs for the Construction of Composite Indicators

When using SEM or PLS-PM to construct composite indicators, it's essential to start with a strong theoretical foundation or prior research to define the model. This entails identifying latent variables, their interrelationships, and the observable indicators for these constructs. SEM can explain covariances among observed variables through relationships between composites and individual variables. (Grace, 2008)

SEM can be employed to construct a composite index and its subindexes by first defining the latent constructs or factors that represent the underlying concepts of interest, such as economic development, social well-being, or environmental sustainability. Once these latent constructs are established, observed variables or indicators related to each construct are chosen and loaded onto their respective factors in the SEM. This step involves specifying the measurement model, ensuring that the indicators effectively represent the latent constructs. After confirming the measurement model's adequacy, the structural model is developed, specifying the relationships and paths between the latent constructs to capture the complex interactions and dependencies among them. The resulting SEM provides a comprehensive framework for constructing the composite index and subindexes, incorporating both the measurement and structural aspects of the model.(Grace, 2008)

SEM can be employed in both exploratory (model-building) and confirmatory (hypothesis-testing) modes. In model-building, SEM facilitates the exploration of the data to develop a model, while in hypothesis-testing, it is used to test pre-specified models against the data. PMs specifically shine in model-building as they fit a composite model which enables simultaneous assess to the relationships between multiple observed variables and their latent constructs. This dual capability makes SEM highly versatile in handling complex data structures often encountered in the construction of composite indicators.

2.4.3 Strengths and Weaknesses of this Approach to Creating Composite Indicators

When using SEMs/PMs in general to construct composite indicators/indices, the first strength is the SEM itself. SEMs are able to examine many relationships while partialing out measurement error. It can also look at correlation within this error and identify the degree of which unknowns influence shared error in your variables. This allows you to reliably create composite indicators baed on multiple variables. It is relatively robust to missing data, since it uses raw data to fit an SEM instead of summary statistics, and even moreso in the case of PM (Lauro et al., 2018). A SEM allows us to analyze dependent observations, as well as longitudinal data such as time series data or growth models. This works well if you plan to incorporate logintudinal data within your index. Some limitations of SEM relate with the use of latent variables. Although a latent variable is designed to closely approximate a measured variable, it might differ in it's variance, which is a

combination of the measured variable's variance along with the variance of the shared error between measured variables. This means that when creating composite indicators, there will be a difference when using latent variables. To observe and analyze multiple variables at a time comes with the cost of a larger sample size. Since we must meet a minimum sample size threshold for each variable we analyze, the more variables we analyze, the larger sample size we need for calculations to work out. Another limitation is that, before fitting your model, you need to verify if your hypothesized model is fit by the data. That is, to check if there are different solutions for parameters that produce the same covariance matrix. Having multiple non-unique solutions in this case means the model is not identified. This implies a few conditions. First, the t-rule states that the number of parameters t must be less than one half times the number of the observations times the number of observations plus one. Next, the scale of the latent variables must be standardized to unit variance, or scaled to a reference variable. Lastly, we must follow the two-indicator rule: having 2 indicators for each latent variable for cases with 2 or more factors, and the three-indicator rule: having 3 indicators for each latent variable for cases with 1 factor (Bartholomew, 2021). PMs are considered a specific solution to the sample size issue, but that is contested in literature (Sarstedt et al., 2016). Some specific weaknesses of PMs also arise: it uses OLS regression, and thus regression assumptions should be fulfilled (Lauro et al., 2018).

In SEM, various types of errors can occur, including errors of model specification, problematic data, errors of analysis, and errors of interpretation. When addressing model specification issues, it's crucial to carefully consider composite indicator design and model specification since even a strong statistical method like SEM can fail if the design is flawed. In terms of data, it's important to accurately assess data sources and their suitability for composite index construction, as well as identify any potential data collection flaws. Regarding data analysis, caution should be exercised when making causal inferences with SEM, as the order of variables matters, and causal relationships are often complex to establish. Interpretation is generally less problematic since indices are numerical and easily comparable on a suitable scale. (Beran and Violato, 2010b; Grace, 2008)

Moreover, SEMs are simplified representations of reality, and sometimes important factors may be omitted due to limitations in knowledge or measurement capabilities. This omission can lead to misleading interpretations. It's important to be open to alternative models that fit the data and recognize that initial models may be rough approximations. Another thing to consider is the adequacy of the data. Ensuring the quality of the data along with a sufficient sample size is crucial. Recommendations suggest at least 5-20 observations per model parameter. Overall, it's essential to thoroughly consider whether the data is fit for SEMs before analysis. (Grace, 2008)

2.5 Operational Issues When Constructing a Composite Index: Variable Choice, Transformation, Evaluation

2.5.1 Choice of Informative Variables

Composite indices simplify complex phenomena into a single numerical value, aiding comparisons. Constructing these indices involves assigning weights to variables, representing their empirical importance. In the assessment of composite indices, variable importance gauges how individual variables contribute to the overall score, addressing concepts like economic development or social progress (Booysen et al., 2002).

Evaluating composite indices requires understanding variable importance, achieved through assigning weights based on perceived significance and conducting sensitivity analysis. This practice

enhances interpretability, clarifying the factors influencing outcomes. Practical applications extend to policy making, where insights into influential variables guide effective interventions and strategies to address specific challenges or areas in need of improvement (Dan et al., 2023).

Schlossarek's paper examines the issue of weights and importance in composite indices of development. The nominal weight assigned to a variable often differs from the degree to which the variable affects the scores of the overall index (Schlossarek et al., 2019). The newly suggested notion of importance is based on the idea that an important indicator, if omitted from the index, causes large changes in countries' results. It is argued that for sake of transparency, it is necessary to communicate to the users of indices the difference between the nominal weights and the importance. Additionally, the authors insist that the information about the empirical importance is crucial for the constructors of indices who may take it into account when deciding about the inclusion of variables into an index and about the final weights of variables (Schlossarek et al., 2019). The nominal weights do not provide credible information on the importance of variables if it is not accompanied by other methodological procedures, e.g. the method of data normalization. The communication of nominal weights alone implicitly creates a perception that the nominal weight is a measure of the importance of the variable (Schlossarek et al., 2019).

The stronger the correlation is between the original and modified composite index, the lower the importance of the examined indicator. The correlation coefficient equal to 1 means that the exclusion of the indicator does not affect (up to a linear transformation) the composite index and therefore has no importance. Concept of importance of a given indicator is based on correlation between the original composite index (OCI) which includes the examined indicator, and the modified composite index (MCI) which excludes the examined indicator (Schlossarek et al., 2019). The two metrics of MAI (Measure of absolute importance) and MRI (Measure of relative importance) are then constructed based on correlation between OCI and MCI to determine the relationship between nominal weights and importance. The interpretation of the strength of the relationship depends not just on the correlation coefficient, but also on what relationship we expect.

2.5.2 Variable Transformation

Normalization is a crucial step when dealing with variables in a dataset with different measurement units, aiming to standardize indicators by transforming them into dimensionless numbers. However, the choice of normalization method can be influenced by scale transformations, leading to varied outcomes based on measurement units. For example, expressing temperature in Celsius or Fahrenheit can result in different composite indicator outcomes. While standardization methods, like z-scores, maintain invariance to changes in measurement units, other techniques, such as logarithmic scaling, introduce non-invariance. Common normalization approaches include ranking indicators, which is simple but loses absolute level information, and standardization, providing comparability with a mean of zero and standard deviation of one. Re-scaling is another method that transforms indicators to a uniform range between 0 and 1, but it may magnify the impact of extreme values (Nardo et al., 2005).

2.5.3 Techniques for Index Evaluation

For our clients seeking to construct and interpret composite indices, understanding the methodologies of sensitivity analysis, variable importance, and quality assessment is crucial. These analyses provide a robust framework for evaluating and refining composite indices, ensuring their reliability

and relevance in decision-making processes.

Sensitivity Analysis (SA) is key in composite indices, particularly for understanding the impact of variations in input factors (X_i) on the output. These indices often show non-linear behaviors, making SA vital. It primarily utilizes variance-based sensitivity measures, such as the First-Order Sensitivity Measure (S_i), reflecting the direct effect of an input factor on output variance, and the Total-Effect Sensitivity Measure (ST_i), accounting for all direct and interaction effects of an input factor. Techniques like the Sobol' method, especially Saltelli's adaptation, are instrumental in these calculations.

The next critical step is assessing the importance of individual variables, crucial for determining their significance in the composite index. The method proposed by Schlossarek, Syrovátka, and Vencálek (2019) begins with examining each variable's nominal weight and its influence on the index score. It also involves analyzing data variability and the impact of normalization methods, such as Z-scores and min-max normalization. Correlation analysis, particularly using the Pearson correlation coefficient, is then conducted to evaluate the unique or shared impact of each variable on the index. This comprehensive approach helps identify significant differences between nominal weights and actual impact, facilitating necessary adjustments.

Finally, conducting a quality assessment of the composite index, primarily through Uncertainty Analysis (UA), is essential. UA explores uncertainties in input factors, using simulation techniques like the Monte Carlo method, involving stochastic simulation, to assess their effect on the output. Key elements in UA include the selection of sub-indicators, data selection and editing, data normalization, weighting schemes, and the composite indicator formula. We advise you to focus on a few critical input factors for efficient analysis (Saisana et al., 2005). This involves assigning Probability Density Functions (PDFs) to input factors, generating random combinations of these factors, selecting normalization methods and weighting schemes, and computing the output for each combination. In plain terms, PDF entails a mathematical function that describes the likelihood of different outcomes or values of the factor. The final phase involves analyzing the output vector to construct the empirical PDF of the output, which helps estimate characteristics like variance and higher-order moments of the composite index. Incorporating the Sobol' method, particularly Saltelli's version, enriches the sensitivity analysis. This comprehensive approach to UA and sensitivity analysis substantially improves the composite index's dependability, making it a more effective tool for analysis and decision-making.

3 Stylistic Instructions – Delete Prior to Circulation

NOTE 1: Here is an example of how to cite a reference in–line and have it appear in the references section:

We propose that you use the R programming language (R Core Team, 2020) for your analysis.

Note that there is a BibTex citation entry call 'Rcitation' in the bibliography file SCC.bib (see menu on left of OverLeaf Screen). Most online links to journal articles will have a link that provides citations in a variety of formats; usually BibTex is an option. There are different entry formats for books, journal articles, software manuals and online links, so be sure to use the appropriate format.

NOTE 2: Please use typewriter font (as shown `here`) for references to `R`, `R` functions and variable names.

References

(2023). *Illuminating Hidden Harvests*. ://www.fao.org/fishery/en/publication/288864: Food and Agriculture Organization of the United Nations, Duke University, WorldFish.

Bartholomew, D. J. (2021). *Analysis of Multivariate Social Science Data*. CRC Press.

Beran, T. N. and C. Violato (2010a). Structural equation modeling in medical research: a primer. *BMC research notes 3*(1), 1–10.

Beran, T. N. and C. Violato (2010b). Structural equation modeling in medical research: A primer. *BMC Research Notes 3*(1).

Budaev, S. V. (2010). Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. *Ethology 116*(5), 472–480.

Chao, Y.-S. (2017). Principal component-based weighted indices and a framework to evaluate indices: Results from the medical expenditure panel survey 1996 to 2011. *PubMed*.

Fan, Y., J. Chen, G. Shirkey, R. John, S. R. Wu, H. Park, and C. Shao (2016). Applications of structural equation modeling (sem) in ecological studies: an updated review. *Ecological Processes 5*, 1–12.

Floyd, F. J. and K. F. Widaman (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment 7*(3).

Grace, J. B. (2008). Structural equation modeling for observational studies. *The Journal of Wildlife Management 72*(1), 14–22.

Henseler, J., G. Hubona, and P. A. Ray (2016). Using pls path modeling in new technology research: updated guidelines. *Industrial management & data systems 116*(1), 2–20.

Lauro, N. C., M. G. Grassia, and R. Cataldo (2018). Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. *Social Indicators Research 135*, 421–455.

Mazziotta, M. and A. Pareto (2015). Composite index construction by pca? no, thanks. In *Conference Paper*, pp. 1–9.

Morekonda, H. (2020). Building index using principal component analysis. *Github*.

Nardo, M., M. Saisana, A. Saltelli, S. Tarantola, et al. (2005). Tools for composite indicators building. *European Comission, Ispra 15*(1).

Qualtrics (2023). Factor analysis and how it simplifies research findings. https://www.qualtrics.com/experience-management/research/factor-analysis/. Accessed: 14 Nov 2023.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ram, R. (1982). Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component'representation. *Journal of Development Economics 11*(2), 227–247.

RDocumentation (2019). prcomp: Principal components analysis. *RDocumentation*.

Sarstedt, M., J. F. Hair, C. M. Ringle, K. O. Thiele, and S. P. Gudergan (2016). Estimation issues with pls and cbsem: Where the bias lies! *Journal of Business Research 69*(10), 3998–4010.

Vyas, S. and L. Kumaranayake (2006). Constructing socio-economic status indices: how to use principal components analysis. *Health policy and planning 21*(6), 459–468.