Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 73–82
http://www.ssca.org.in/journal

Development of Survey Weighted Composite Indices under Complex Surveys

Deepak Singh[1], Pradip Basak[2], Tauqueer Ahmad[1], Raju Kumar[1] and Anil Rai[1]
[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal*

Received: 01 July 2022; Revised: 01 June 2023; Accepted: 05 June 2023

Abstract

An index is constructed by a mathematical model representing multi-dimensional variables into a single value. Multi-dimensional variables are often correlated with each other which is referred as the problem of multicollinearity. Most of the present indices except principal component analysis (PCA) based method do not consider the effect of multicollinearity among the variables. For survey data, even though the PCA based indices are able to tackle the problem of multicollinearity but do not use survey weights and auxiliary information which leads to erroneous ranking of the survey units like households, districts, states, *etc.* Therefore, the present study proposes some new methods of index construction which are capable to incorporate the survey weights and auxiliary information available in the complex survey data as well as removes the effect of multicollinearity among variables.

Key words: Index; Multicollinearity; Survey weight; Auxiliary information; Complex surveys.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Indicators are helpful in recognizing patterns and attracting consideration regarding specific issues. A composite indicator is framed when singular indicators are assembled into a single index on the premise of a basic model. Composite indicators/index are much similar to mathematical or computational models which should ideally measure multi-dimensional concepts, which can't be caught by a single indicator alone, *e.g.*, intensity, industrialization, *etc.*

An index is a single measure to rank the units based on the multi-dimensional concepts which are presented in form of multivariable. Indices are quite useful to measure multi-dimensional concepts, which can't be caught by a single indicator alone like to summarize complex (elusive) or multi-dimensional processes into a single figure to benchmark the performance of countries, states, districts, *etc.* for policy formulation. The primary

Corresponding Author: Pradip Basak
Email: pradipbasak.99@gmail.com

role of index is to simplify otherwise complicated comparisons. Niti Aayog, Government of India develops composite indices like water management index, district hospital index, export preparedness index, India innovation index, multidimensional poverty index, school education quality index, SDG Index, state energy index and state health index. A comprehensive survey of different indicators of economic and social well-being has been provided by Sharpe (1999). The quantification of development efforts affected in various socio-economic fields was studied by constructing composite index of development based on information on fourteen important indicators by Narain *et al.,* (1991). The economic growth of China using social indicators has been estimated by Klein and Ozmucur (2003). Potential agro-forestry areas using Objective Analytic Hierarchy Process was identified by Ahmad *et al.,* (2003). Economic development in Karnataka, hilly states and Jammu and Kashmir was evaluated by Narain *et al.,* (2003, 2004, 2005). Livelihood index for different agro-climatic zones of India was developed by Rai *et al.,* (2008). The food insecurity in urban India was reported by developing the food insecurity index by Athreya *et al.,* (2010). The Human Development Index (HDI) developed by United Nations Development Programme is the geometric mean of the three-dimension indices *i.e.*, Health, Education and Income (Human Development Report, 2016).

Therefore, in most of the situations, composite index are based on simple or weighted average method which does not consider the effect of multicollinearity among the indicator variables that are used for index construction. PCA based index accounts for the effect of multicollinearity among the indicator variables through the eigen values and eigen vectors derived from the variance-covariance matrix using maximum likelihood or ordinary least squares methods of estimation. Dahal (2007) developed soil quality index by using PCA in which all those principal components (PCs) for which the eigen value is greater than one are retained. Agricultural development index was developed by Kumar (2008) using the principal component technique. Medical expenditure panel survey from 1996 to 2011 was used to develop principal component-based index by Chao and Wu (2017). Water poverty index was developed by Senna *et al.,* (2019) using PCA.

However, the above PCA based index methods are based on the assumption that sample elements, on which the indicator variables are measured, are independent and identically distributed. This assumption of independence holds good if the data are collected through simple random sampling with replacement. However, it does not hold good for other sampling designs where the inclusion probability and survey weight are attached with each sampling unit and hence the above PCA based index methods lacks representativeness of the population when the data is collected with complex survey design. Now a day, most of the survey designs are complex in nature involving stratification, unequal probabilities of selection, clustering, multi-stages, multi-phases and auxiliary information. In case of large-scale surveys, stratified multistage sampling design is widely used where the units in a stratum are relatively homogenous which violates the assumption of independence of sample elements. Any deviation from independence assumption leads to erroneous estimation of variance covariance matrix which in turn leads to erroneous estimation of eigenvalues and eigenvectors, and thereby resulting in poor PCA based index. Therefore, in case of complex survey data there is a need to develop PCA based index using survey weights to tackle the problem of representativeness of population and auxiliary information which leads to development of efficient indices.

2. Methodology

2.1. Estimators of variance-covariance matrix

Let us consider a finite population $U = (1, 2, \ldots, k, \ldots, N)$ of size N units having l subpopulations/blocks/states such that the h^{th} subpopulation has N_h units and $\sum_{h=1}^{l} N_h = N$, $(h = 1, 2, \ldots, l)$. Let s be a probabilistic sample of size n drawn from this population such that $\sum_{h=1}^{l} n_h = n$ where n_h is the number of units belongs to the h^{th} sub-population with assumption that $n_h \neq 0$ and d_{hi} denotes the survey weight associated with i^{th} unit of the sample in h^{th} subpopulation such that $\sum_{h=1}^{l}\sum_{i=1}^{n_h} d_{hi} = 1$. Let $\mathbf{y} = (y_{1,}y_{2,...,}y_q)'$ and $\mathbf{x} = (x_{1,}x_{2,...,}x_q)'$ be the p and q set of standardised indicators and auxiliary variables respectively. Let, $\mathbf{y}_{hi} = (y_{hi1,}y_{hi2,...,}y_{hip})'$ and $\mathbf{x}_{hi} = (x_{hi1,}x_{hi2,...,}x_{hip})'$ be values of the variables $\mathbf{y}$ and $\mathbf{x}$ corresponding to i^{th} sample unit of h^{th} sub-population where, $h = 1, 2, \ldots, l$ and $i = 1, 2, \ldots, n_h$. The ordinary least squares estimator of variance-covariance matrix, Σ_{yy} is given as

$$\hat{\sum}_{yy} = \mathbf{V}_{yys} = (n-1)^{-1}\sum_h\sum_i (\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)^T \tag{1}$$

where, $\bar{\mathbf{y}}_s = \sum_h\sum_i \mathbf{y}_{hi}/n$.

Following Skinner *et al.,*(1986) and Smith & Holmes (1989), the survey-weighted estimator of $\sum_{yy}$ is given by

$$\hat{\sum}_{yyw} = \mathbf{V}^*_{yys} = \sum_h\sum_i d_{hi}\mathbf{y}_{hi}\mathbf{y}_{hi}^T - \mathbf{y}^*_s\mathbf{y}^{*T}_s \tag{2}$$

where, $\bar{\mathbf{y}}^*_s = \sum_h\sum_i d_{hi}\mathbf{y}_{hi}$, and in the presence of auxiliary information $\mathbf{x}$, unweighted regression estimator is given by

$$\hat{\sum}_{yyr} = \mathbf{V}_{yys} + \mathbf{b}_{yx}(\sum_{xx} - V_{xxs})\mathbf{b}_{yx}^T \tag{3}$$

where,

$$\mathbf{b}_{yx} = \mathbf{V}_{xys}\mathbf{V}_{xxs}^{-1},$$
$$\mathbf{V}_{xxs} = (n-1)^{-1}\sum_h\sum_i (\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)(\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)^T,$$
$$\mathbf{V}_{xys} = (n-1)^{-1}\sum_h\sum_i (\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)^T.$$

In the case of survey data with auxiliary information, following Skinner *et al.,* (1986) and Smith & Holmes (1989), survey-weighted regression estimator is given by

$$\hat{\sum}_{yywr} = \mathbf{V}^*_{yys} + \mathbf{b}_{yxw}(\sum_{xx} - \mathbf{V}^*_{xxs})\mathbf{b}_{yxw}^T \tag{4}$$

where,

$$\mathbf{b}_{yxw} = \mathbf{V}^*_{xys}\mathbf{V}^{*-1}_{xxs},$$

$$\mathbf{V}^{*}_{xxs} = \sum_{h}\sum_{i} d_{hi}\mathbf{x}_{hi}\mathbf{x}^{T}_{hi} - \bar{\mathbf{x}}^{*}_{s}\bar{\mathbf{x}}^{*T}_{s},$$

$$\mathbf{V}^{*}_{xys} = \sum_{h}\sum_{i} d_{hi}\mathbf{x}_{hi}\mathbf{y}^{T}_{hi} - \bar{\mathbf{x}}^{*}_{s}\bar{\mathbf{y}}^{*T}_{s}.$$

2.2. Methodology of proposed indices

Let us assume that $\hat{\Sigma}_{yy}$ is a real positive definite matrix. Let, the non-zero eigenvalues of $\hat{\Sigma}_{yy}$ are $\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_{\mathrm{p}}$ and the corresponding eigen vectors are $\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_{\mathrm{p}}$. For distinct λ_{j}'s ($j = 1, 2, 3, \ldots, p$), an orthogonal matrix of order pxp can be formed as

$$\Gamma = [\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_{\mathrm{p}}], \tag{5}$$

such that, $\hat{\Sigma}_{yy} = \Gamma A \Gamma^{T}$, where $\mathrm{A} = \mathrm{diag}\left(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{\mathrm{p}}\right) = \Gamma^{T}\hat{\Sigma}_{yy}\Gamma$. Now let us consider an orthogonal transformation of $\mathbf{y}$ such that

$$\mathbf{P} = \Gamma^{T}\mathbf{y} \tag{6}$$

where $\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_{\mathrm{p}}$ are the p components of $\mathbf{P}$ and are called as PCs. The composite index corresponding to i^{th} sample unit of h^{th} sub-population is given as

$$C_{hi} = \frac{\sum_{j=1}^{p} \lambda_j PC_{hij}}{\sum_{j=1}^{p} \lambda_j} \tag{7}$$

where the PC_{hij}'s are principal component scores of j^{th} variable corresponding to the i^{th} sample unit of h^{th} sub-population $\forall\ h = 1, 2, \ldots, l; i = 1, 2, \ldots, n_h; j = 1, 2, \ldots, p$. The average of C_{hi}'s within h^{th} sub-population gives the composite index value for h^{th} sub-population as

$$C_h = \sum_{i=1}^{n_h} C_{hi} \Big/ n_h. \tag{8}$$

The composite index values of sub-populations are re-scaled by using the following formula as

$$CI_h = \frac{C_h - \min(C_h)}{\max(C_h) - \min(C_h)}. \tag{9}$$

The ranking of l sub-populations is done based on the re-scaled composite index values (CI_h). All the composite index values (CI_h) lie between 0 and 1, where one denotes the highest rank and zero denotes the lowest rank.

The existing PCA based index uses the non-zero eigenvalues derived from the ordinary least squares estimator of variance-covariance matrix, $\hat{\Sigma}_{yy}$. Here, indices are proposed based on survey weighted estimator, unweighted regression estimator and survey weighted regression estimator of variance-covariance matrix. The index that uses the non-zero eigenvalues derived from the survey weighted estimator of variance-covariance matrix, $\hat{\Sigma}_{yyw}$ is referred as the survey weighted PCA based index and it is used when the data are collected through complex survey designs in which the inclusion probability for all the units is not same. The index developed based on the non-zero eigenvalues derived from unweighted regression estimator of variance-covariance matrix, $\hat{\Sigma}_{yyr}$ is referred as the unweighted regression PCA based index and it is useful when auxiliary information is available in the data. One more index has been developed that uses the non-zero eigenvalues derived from survey weighted regression estimator of variance-covariance matrix, $\hat{\Sigma}_{yywr}$ and is referred as the survey-weighted regression PCA based index. This index is particularly useful when there is the presence of auxiliary information under complex survey designs.

3. Empirical evaluations

This Section summarizes the simulation studies conducted to evaluate the empirical performance of the developed indices. Two types of simulation studies, namely design based simulation and model-based simulation are considered. In case of design-based simulation, real survey dataset is used as a finite population. From this fixed population, repeated random samples are drawn. In the case of model-based simulation, at each simulation run a synthetic population data is first generated under the model and then a sample is drawn from this simulated population, and process is repeated several times. In the simulation studies, the following indices are considered

i) Unweighted PCA based index (denoted as PCA Index),

ii) Survey weighted PCA based index (denoted as SW-PCA Index),

iii) Unweighted regression PCA based index (denoted as REG-PCA Index), and

iv) Survey weighted regression PCA based index (denoted as SW-REG-PCA Index).

3.1. Design-based simulation

The household consumer expenditure survey data of NSS 68$^{\text{th}}$ round is used for design based simulation study. The data of five states namely, Jammu and Kashmir, Orissa, Kerala, Sikkim and Jharkhand, and one union territory, *i.e.*, Andaman and Nicobar Islands have been considered for the study. The survey data of these five states and one union territory are considered as independent populations and then samples (10 % of the population) are drawn from each of these populations. The primary units of the survey are households. Therefore, within a state, sample size is allocated among the districts using proportional allocation and then from each of the districts, households are selected by simple random sampling without replacement (SRSWOR). Here, the variables considered are Cereals (Z1), Pulses and pulse products (Z2), Milk and milk products (Z3), Salt and sugar (Z4), Edible oil (Z5), Egg, Fish and meat (Z6), Vegetables (Z7), Fruits (fresh) (Z8), Spices (Z9), Beverages (Z10), Served processed food (Z11) and Packaged processed food (Z12). Following Smith and Holmes (1989), a new variable is created by summing up all the twelve variables which is considered as auxiliary variable.

The composite index values are computed for each households using different methods of index construction and the average of composite index values of all the households within a district is taken as the index value for that district. The index values computed for each of the districts within a state are compared with the composite index value of districts based on the population variance covariance matrix. The Monte Carlo simulation was run S=5000 times. Simulation studies are carried out in R software. The developed indices are evaluated by percentage relative root mean squared error (RRMSE), defined by

$$RRMSE(\widehat{\theta}) = \sqrt{\frac{1}{s}\sum_{s=1}^{S}\left[\sum_{i=1}^{k}\left(\frac{\widehat{\theta}_{si}-\theta_i}{\theta_i}\right)^2\right]} * 100 \tag{10}$$

where, $\widehat{\theta}_{si}$ is the sample index value for i^{th} district at s^{th} simulation run and θ_i is the population index value for i^{th} district. The values of the percentage relative root mean square error of different indices are reported in Table 1.

Table 1: Percentage relative root mean square error (RRMSE %) of different indices considered in the design-based simulation

State	PCA Index	SW-PCA Index	REG-PCA Index	SW-REG-PCA Index
Jammu & Kashmir	1242.00	879.14	1230.95	863.70
Orissa	363.18	362.21	357.90	352.35
Kerala	228.27	227.91	222.10	223.26
Andaman & Nicobar Islands	208.97	207.47	206.27	205.00
Sikkim	126.49	125.59	121.51	118.10
Jharkhand	466.21	452.80	432.25	424.95

From Table 1, it is clear that the proposed indices perform better than the unweighted PCA based index in terms of RRMSE. Among the proposed indices, SW-REG-PCA Index performs best followed by REG-PCA Index and SW-PCA Index. Since SW-REG-PCA Index utilises the auxiliary information as well as survey weights available through survey design, therefore, it performs best. However, for the state of Jammu & Kashmir, SW-PCA Index performs better than the REG-PCA Index. The Table 1 indicates that the proposed methodologies of indices development are efficient in comparison to the existing PCA based index method for complex survey designs.

3.2. Model based simulation

In model-based simulation, the methodology given by Smith & Holmes (1989) were followed where they have assumed the design variable Z as a sum of other variables. Thus, Z is a continuous variable to form population design groups such as strata to investigate the performance of different estimators of population variance covariance matrix for complex survey design. In this simulation study, an artificial population is generated using multivariate normal distribution $\mathbf{X} = \left(\mathbf{Y}^T, Z\right)^T$ having mean vector $\mathbf{u}_x$ and variance covariance matrix $\sum_{xx}$ satisfying the linearity and homoscedasticity assumptions. The vector $\mathbf{Y}$ comprises of twelve variables and Z is the design variable which is the sum of all the twelve variables. The mean vector $\mathbf{u}_x$ and variance-covariance matrix $\sum_{xx}$ are estimated from the NSS 68$^{\text{th}}$ round household consumption expenditure survey data on the twelve set of variables, *i.e.*, Cereals (Z1), Pulses and pulse products (Z2), Milk and milk products (Z3), Salt and sugar (Z4), Edible oil (Z5), Egg, Fish and meat (Z6), Vegetables (Z7), Fruits (fresh) (Z8), Spices (Z9), Beverages (Z10), Served processed food (Z11) and Packaged processed food (Z12). A finite population of one lakh units is generated at each simulation run. Then the population is stratified into five strata, each having equal number of units based on the ordered z-values of the design variable Z.

Table 2: Mean of variables considered for population data generation

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
807.93	213.31	721.86	127.64	65.13	166.69	120.73	70.09	58.70	49.10	97.85	56.63

Table 3: Variance-covariance matrix of variables considered for population data generation

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
Y1	263931	34252	76232	19551	11007	26131	19112	6828	6776	6280	3673	7081
Y2	34252	22964	41505	9415	4393	4090	4918	3003	2995	2189	1594	2829
Y3	76232	41505	620633	54028	14295	8103	18088	16563	8402	11975	10450	15495
Y4	19551	9415	54028	14862	3476	2290	3434	1884	1883	1560	896	1835
Y5	11007	4393	14295	3476	3891	3192	2433	973	1508	982	555	1601
Y6	26131	4090	8103	2290	3192	28266	4843	3285	2466	2187	3389	2588
Y7	19112	4918	18088	3434	2433	4843	6531	1726	1520	1361	924	1772
Y8	6828	3003	16563	1884	973	3285	1726	4607	887	1316	1763	1731
Y9	6776	2995	8402	1883	1508	2466	1520	887	1852	717	733	890
Y10	6280	2189	11975	1560	982	2187	1361	1316	717	2733	1582	1184
Y11	3673	1594	10450	896	555	3389	924	1763	733	1582	101389	2261
Y12	7081	2829	15495	1835	1601	2588	1772	1731	890	1184	2261	6203

Samples of size 2000 are selected from this population using SRSWOR within each stratum and allocated sample sizes in the strata are provided in Table 4.

Table 4: Allocation of sample size in the strata

Stratum	1	2	3	4	5
Sample size	600	300	200	300	600

Then the various indices are computed using this sample data. The Monte Carlo simulation was run S=5000 times. Simulation studies are carried out in R software. The developed indices are evaluated by the criterion of percentage relative root mean squared error (RRMSE, %), defined by

$$RRMSE(\widehat{\theta}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left[\sum_{i=1}^{k}\left(\frac{\widehat{\theta}_{si}-\theta_{si}}{\theta_{si}}\right)^2\right]} * 100 \tag{11}$$

where, $\widehat{\theta}_{si}$ and θ_{si} are the sample and population index values of i^{th} strata at s^{th} simulation run.

Table 5: Percentage relative root mean squared error (RRMSE, %) of different indices considered in model-based simulation

Indices	% RRMSE
Unweighted PCA based index	22.84
Survey weighted PCA based index	19.46
Unweighted regression PCA based index	19.20
Survey weighted regression PCA based index	19.03

Table 5 reports the performance of all the proposed indices obtained from the simulation study. From Table 5, it is clear that the SW-REG-PCA Index, which utilises the auxiliary information as well as survey weights, performs best followed by REG-PCA Index and SW-PCA Index. Therefore, all the developed indices performs better than the existing PCA based Index in terms of the criterion of RRMSE, %.

From the empirical evaluations in section 3, it is inferred that when sample is selected through complex survey design in which there is unequal selection probabilities of sample units, the indices that incorporate survey weights perform better in comparison to the traditional PCA based index method which is incapable to incorporate the survey weights. The proposed REG-PCA Index which is capable to incorporate the auxiliary information performs better than the traditional PCA based index method which does not utilize the auxiliary information even when it is available.

4. Conclusions

Most of the large-scale surveys conducted by different Government agencies, NGOs, research organisations and private firms use complex survey designs which involve unequal probabilities of selection, stratification, clustering, multistage, multiphase, nonresponse and other post stratification adjustments. Ignoring these aspects of complex survey data while constructing indices may lead to biased estimates and greater standard errors which leads to erroneous ranking of the survey units under consideration. Thus, it may result in erroneous inferences. Therefore, in the present study, different indices are developed which are capable to incorporate the survey weights and auxiliary information available in the complex survey data as well as removes the effect of multicollinearity among the index variables. The improved performance of the developed indices in comparison to the existing PCA based index have been demonstrated through simulation studies using both real and artificially generated data. Therefore, the developed indices will give better inferences in the case of complex survey data.

Acknowledgements

The authors are profoundly thankful to the editors and the anonymous referee for their valuable comments and suggestions which led to improvements in the article. The authors also gratefully acknowledge the valuable contributions made by Late Dr. Hukum Chandra, National Fellow and Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi.

References

Ahmad, T., Singh, R., and Rai, A. (2003). Development of GIS based technique for identification of potential agro forestry areas. *Project Report*, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

Athreya, V. B., Rukmani, R., Bhavani, R. V., Anuradha, G., Gopinath, R., and Velan S. A. (2010). Report on the state of food insecurity in urban India. *Report*, M. S. Swaminathan Research Foundation.

Chao, Y. S. and Wu, C. J. (2017). Principal component-based weighted indices and a framework to evaluate indices: Results from the medical expenditure panel survey 1996 to 2011. *PLoS ONE*, **12**, e0183997.

Dahal, H. (2007). Factor analysis for soil test data: A methodological approach in environment friendly soil fertility management. *Journal of Agriculture and Environment*, **8**, 8–19.

Human Development Report (2016). *Human Development for Everyone*. United Nations Development Programme.

Klein, L. R. and Ozmucur, S. (2003). The estimation of China's economic growth. *Journal of Economic and Social Measurement*, **28**, 187–202.

Kumar, M. (2008). *A Study on Development Indices and their Sensitivity Analysis*. M.Sc. Thesis, ICAR-Indian Agricultural Research Institute, New Delhi, India.

Narain, P., Rai, S. C., and Sarup, S. (1991). Statistical evaluation on development on socio-economic front. *Journal of the Indian Society of Agricultural Statistics*, **43**, 329–345.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2004). Estimation of socio-economic development in hilly States.*Journal of the Indian Society of Agricultural Statistics*, **58**, 126–135.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2005). Dimension of socio-economic development in Jammu and Kashmir. *Journal of the Indian Society of Agricultural Statistics*, **59**, 243–250.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2003). Evaluation of economic development at micro level in Karnataka. *Journal of the Indian Society of Agricultural Statistics*, **56**, 52–63.

Rai, A., Sharma, S. D., Sahoo, P. M., and Malhotra, P. K. (2008). Development of livelihood index for different agro-climatic zones of India. *Agricultural Economics Research Review*, **21**, 173–182.

Senna, L. D., Maia, A. G., and Medeiros, J. D. F. (2019). The use of principal component analysis for the construction of the Water Poverty Index. *Brazilian Journal of Water Resources*, **24**, 1–14.

Sharpe, A. (1999). A Survey of Indicators of Economic and Social Well-being. *CSLS Research Reports*.

Skinner, C. J., Holmes, D. J., and Smith, T. M. F. (1986). The Effect of Sample Design on Principal Component Analysis. *Journal of the American Statistical Association*, **81**, 789–798.

Smith, T. M. F. and Holmes, D. J. (1989). *Multivariate Analysis.* In Analysis of Complex Surveys (Eds: Skinner, C. J., Holt, D. and Smith, T. M. F.), 165–190, New York: John Wiley and Sons.