The Economic Diversification Index

Methodology Overview

Ben Shepherd, Principal.

June 28th, 2021.

developing trade consultants
policy • research • capacity building

315 W 99th St. #7C
New York, NY 10025
Ben@Developing-Trade.Com

1 INTRODUCTION

This report presents the methodology behind the Economic Diversification Index (EDI). It first addresses general issues in developing this kind of index, then presents the specific solution adopted. Finally, it provides an overview of how the sub-indices and overall index are computed.

2 CONCEPTUALIZING AND REALIZING THE EDI

Economic diversification is not a concept captured in a single data point. Rather, as the main report makes clear, it covers a wide range of indicators that currently do not have expression as a single aggregate index. The starting point is therefore a high dimensional dataset, i.e. observations on a set of indicators for a sample of countries over a given time period. The desired end point is a set of three sub-indices capturing the key dimensions of diversification as set out in the main report—Output, Revenue, and Trade—and an overall index bringing together the three sub-indices. Selection of the detailed indicators is based on the analysis in the main text, i.e. a review of the literature as well as analytical priors.

Conceptually, the problem is one of dimensionality reduction: for the set of indicators relevant to each sub-index and the overall index, the objective is to reduce the number of dimensions in the dataset from the number of indicators to just one. Two general approaches are available to solve this kind of problem: data compression; and prediction. The first set of approaches reduces the dimensionality of a dataset by uncovering the key components of variation across indicators and using a purely mathematical approach to summarize them according to a pre-defined criterion. The second set of approaches uses a given function of the indicator set to predict a variable of interest that should be strongly correlated with economic diversification.

With these two general approaches in mind, exploratory analysis of the EDI dataset examined the following potential methodologies:

- Data compression:
  - Principal component analysis (PCA).
  - Principal factor analysis (PFA).
- Prediction:
  - Bayesian model averaging (BMA).
  - Artificial neural network (ANN).

The two data compression techniques are well established in the economics literature, and have been used by international organizations such as the World Bank (Logistics Performance Index). The two prediction methodologies are much newer, and have not been widely used to produce indices in this way. Examination of their performance was therefore more speculative, with the objective of ascertaining whether or not it was possible to improve on classical techniques. A key limitation of the prediction techniques is that economic diversification—the variable the model should predict—is not observable, as noted above. The models therefore rely on observable proxies, in this case measures of GDP volatility from IMF quarterly GDP data, namely the standard deviation and coefficient of variation of GDP, as well as predicted volatility from an autoregressive conditional heteroskedasticity (ARCH) time series model of quarterly GDP.

3 DATA PRE-TREATMENT

In developing an index like the EDI, a key requirement is that scores be comparable across countries and through time. As such, each EDI observation must be based on the same underlying indicators.

While many statistical techniques can deal easily with missing values for one of a set of indicators, the case of a multi-indicator index is different. To take a simple example, consider an index based on two indicators, A and B, which are aggregated by taking the arithmetic (simple) mean. If B is missing for one country, then the mean is simply A. If A is missing for another country, then the mean is simply B. If both series are observed for a third country, then the mean is (A+B)/2. So the three index scores in this case are not comparable, even if all variables are measured on the same scale: each observation is based on different information sets.

In the context of the EDI, this requirement would mean that the index could only be calculated for those country and year pairs where all component indicators are observed. This constraint is a major one, which would significantly reduce coverage in both the country and time dimensions.

In an effort to ensure the broadest coverage possible, the dataset is therefore pre-treated using linear interpolation and extrapolation to fill in missing observations to the extent possible. The output is a complete input dataset for 89 countries for the 2000-2019 period.

The only other pre-treatment applied to the data is standardization. To eliminate any potential impact of different variable scales, all input data are converted to series with mean zero and unit standard deviation.

4 Assessment of Candidate Methodologies

4.1 Overview of Methodologies

PFA is a standard dimensionality-reduction technique. It starts from a modified correlation matrix of a set of indicators. The principal diagonal of that matrix (all entries equal to unity) is replaced with the R2 from a multiple regression of the variable in question on all other variables in the set, and therefore is strictly less than unity. This step essentially separates out common variation in the set of variables, and idiosyncratic variation in individual variables. The modified correlation matrix is then decomposed into its eigenvectors and eigenvalues. Each eigenvector is a linear combination of the variables in the set, with given weights ("loadings"). The eigenvector that corresponds to the largest eigenvalue (the principal eigenvector) is the one that accounts for the largest possible proportion of the common variation in the indicators.

In applying PFA to develop an index from a set of indicators, it is standard to use the principal eigenvector as the output index. The interpretation is then that the index is the linear combination of the underlying indicators that accounts for the maximum possible proportion of their common variance.

PCA is conceptually very similar to PFA. The only difference is that it starts from a standard correlation matrix of the indicator variables, not the modified one used for PFA. As such, it assumes that the indicator variables only have common variation. An indicator produced using PCA is therefore the linear combination of the indicators that accounts for the maximum possible proportion of the total variance in the set of underlying indicators.

BMA takes a different approach to creating an index. The problem conceptually is again to aggregate an underlying set of indicators into a single index. However, BMA is a technique for prediction and inference, rather than dimensionality reduction. It is akin to a regression model, but accounts systematically for model uncertainty, for instance in relation to prior expectations on parameters, or the set of variables being used. By estimating a potentially large number of models, BMA makes it possible to derive a set of parameter estimates that can be used to construct an index that is a linear

combination of the underlying indicators, based on a weighted average of estimated parameters from the set of models. The interpretation is that the index is a "good" predictor of some output variable of interest, across the range of models estimated.

Many machine learning techniques are available for prediction problems. One popular one is an ANN. It is designed to make predictions using input variables based on complex optimization procedures that feed outputs potentially through a number of layers of analysis, in an analogous way to how information is processed in the human brain. Whereas the techniques discussed above all involve linear predictions, ANNs can capture more complex, nonlinear patterns of variation. The interpretation of an ANN-based EDI is again that it is a "good" predictor of some outcome of interest, with the additional specification that it accounts for nonlinear effects.

4.2 Selection

In pre-analysis, candidate indices were produced using the four methodologies discussed above. The two prediction methodologies yielded similar results, but they were not intuitive. The reason is that the output variable used to test prediction accuracy—GDP volatility—is not perfectly correlated with economic diversification, and so resulted in the introduction of significant noise into the model. The two data compression methodologies produced much more intuitive results. Given the similarity in the two methodologies, results only differed slightly. PCA was therefore preferred because it is the simpler of the two approaches, which aids transparency and replicability in other contexts. The final EDI and its sub-indices were therefore produced using PCA.

5 PCA Output

The strategy for applying PCA to the detailed indicators relied on two steps. The first was to use PCA to produce the three sub-indices: output, revenue, and trade.[1] The second was then to aggregate the three sub-indices into an overall EDI by taking the arithmetic (simple) mean. The rationale for using the simple mean in the second stage is that it is the simplest and most transparent approach, and there is no a priori reason for believing that any one of the three sub-indices is more important to the overall measurement of economic diversification than the others.

The output sub-index takes the following data series as inputs:

- Real GDP.
- Agriculture as a percentage of GDP.
- Gross fixed capital formation as a percentage of GDP.
- Industry as a percentage of GDP.
- Manufacturing as a percentage of GDP.
- Resource rents as a percentage of GDP.
- Services as a percentage of GDP.
- Medium and high technology manufacturing as a percentage of GDP.
- Manufacturing value added per capita.

Table 1 shows the factor loadings produced by PCA. The principal eigenvector accounts for 33.5% of the observed variation in the underlying series. The loadings show that real GDP, services as a percentage of GDP, medium and high technology manufacturing as a percentage of GDP, and

[1] Indices are produced using the standard sum of squares approach, and are converted from variables with mean zero and unit standard deviation to variables with mean 100 and standard deviation 10.

manufacturing value added per capita correlate positively with the EDI output sub-index, while the remaining variables correlate negatively. This finding is intuitive in most cases, but the contrast between industry and services shows that the data tend to support the importance of the services sector as a determinant of output diversification. Resource rents exhibit a strong negative correlation, which means that resource dependent economies tend to score lower on this sub-index. This fact perhaps explains the result for industry, which includes extractive industries.

Table 1: PCA loadings for the EDI output sub-index.

Variable	Loading
Real GDP.	0.246
Agriculture as a percentage of GDP.	-0.332
Gross fixed capital formation as a percentage of GDP.	-0.022
Industry as a percentage of GDP.	-0.305
Manufacturing as a percentage of GDP.	-0.004
Resource rents as a percentage of GDP.	-0.394
Services as a percentage of GDP.	0.510
Medium and high technology manufacturing as a percentage of GDP.	0.389
Manufacturing value added per capita.	0.410

The revenue sub-index includes the following variables:

- Excise tax revenue as a percentage of GDP.
- Income tax revenue as a percentage of GDP.
- Goods and services tax revenue as a percentage of GDP.
- Tax revenue as a percentage of GDP.
- Total revenue as a percentage of GDP.
- Trade revenue as a percentage of GDP.

Table 2 shows PCA loadings for this sub-index. The principal eigenvector accounts for 51.9% of the observed variation in the individual indicators. The Table shows that all variables except trade revenue are positively correlated with the EDI revenue sub-index. This interpretation is intuitive: higher proportions of revenue from different sources in GDP should indeed be indicative of greater diversification. But reliance on revenues from trade (tariffs) is usually associated with underdevelopment of the tax system in general, in particular income and consumption taxes; so the finding for this last variable is also intuitive, as it suggests that revenues tend to be less diversified if there is high reliance on trade taxes to raise revenue.

Table 2: PCA loadings for the EDI revenue sub-index.

Variable	Loading
Excise tax revenue as a percentage of GDP.	0.328
Income tax revenue as a percentage of GDP.	0.481
Goods and services tax revenue as a percentage of GDP.	0.471
Tax revenue as a percentage of GDP.	0.534

Total revenue as a percentage of GDP.	0.294
Trade revenue as a percentage of GDP.	-0.260

The EDI trade sub-index is based on the following indicators:

- Total value of exports.
- Fuel exports as a percentage of GDP.
- Export market concentration index (Hirschman-Herfindahl Index, HHI).
- Total value of imports.
- Manufactured exports as a percentage of total merchandise exports.
- Medium and high technology manufactured exports as a percentage of total merchandise exports.
- Merchandise exports as a percentage of GDP.
- Total value of services exports.
- Export product concentration index.
- Import product concentration index.

Table 3 shows PCA loadings for the trade sub-index. The principal eigenvector accounts for 37.3% of the observed variation in the individual indicators. The table shows that export market concentration, product concentration of exports and imports, and fuel exports are all negatively correlated with trade diversification, but the remaining variables are positively correlated. This result is intuitive, as the positively correlated variables all capture aspects of country performance that suggest deeper integration into the global trade system. The case of fuel exports is important, as it suggests that countries with significant reliance on that sector tend to be less diversified from a trade point of view. It therefore complements the finding on revenue diversification, where resource rents (for instance, from extractive industries) are negatively correlated with revenue diversification.

Table 3: PCA loadings for the EDI trade sub-index.

Variable	Loading
Total value of exports.	0.407
Fuel exports as a percentage of GDP.	-0.289
Export market concentration index (Hirschman-Herfindahl Index, HHI).	-0.059
Total value of imports.	0.424
Manufactured exports as a percentage of total merchandise exports.	0.366
Medium and high technology manufactured exports as a percentage of total merchandise exports.	0.369
Merchandise exports as a percentage of GDP.	0.058
Total value of services exports.	0.410
Export product concentration index.	-0.353
Import product concentration index.	-0.034

MATHEMATICAL APPENDIX

PFA

Let X be a random vector with finite variance. It can be expressed as a linear function of unobserved factors and an error term as follows:

$$X = a + bf + e$$

Where: a is a vector of means; f is the matrix of factors; b is the matrix of loadings; and e is a vector of errors.

The variance-covariance matrix of X (Σ)can be written as follows:

$$\Sigma = bb' + \Psi$$

Where: Ψ is the variance-covariance matrix of the errors, which is assumed to be diagonal. The first term in the expression is a normalization that identifies the matrix of loadings based on an assumption that the factors are uncorrelated.

PCA

PCA is an application of factor analysis in which the factors are assumed to be fixed rather than random, and the residuals are homoskedastic.

BMA

Consider a linear regression model where the matrix of explanatory variables is partitioned into subset: one that is sure to be included in the model, and a second where inclusion is uncertain.

$$Y = X_1 B_1 + X_2 B_2 + e$$

Model uncertainty means that it is possible to obtain an estimate with lower mean squared error than the unrestricted OLS estimate using all variables. There are $I = 2^{k_2}$models, where k_2 is the number of variables in X_2. Model M_i is obtained by including a subset of those k_2 variables such that $0 \leq k_{2i} \leq k_2$, so that it can be written as follows:

$$Y = X_1 B_1 + X_{2i} B_{2i} + e_i$$

A model averaging estimate of B_1 is given by:

$$\widehat{B_1} = \sum_{i=1}^{I} \lambda_i \widehat{B_{1i}}$$

Where: λ_i is a weight; and $\widehat{B_{1i}}$ is the estimate of B_1 obtained by conditioning on Model M_i.

To introduce Bayesian prior beliefs, models are weighted based on their posterior probability. Under equal prior probabilities, the weights are given by:

$$\lambda_i = p(Y|M_i) = c\left(\frac{g}{1+g}\right)^{k_{2i}/2} (Y' M_1 A_i M_1 Y)^{-(n-k_1)/2}$$

Where: c and g are constants; $M_1 = I - (X'X)^{-1}X'$; and $A_i = \frac{g}{1+g}\{M_1 - M_1 X_{2i}(X'_{2i} M_1 X_{2i})^{-1} X'_{2i} M_1\}$.

ANN

An ANN can be represented schematically as follows:

$$Y_n = f(w_{n-1}Y_{n-1})$$

Where: Y is the output; w is the weight vector; and f is a function, in the case explored for the EDI a rectified linear model for input and intermediate layers, and a simple linear function for the output layer.

Errors are back-propagated through the network:

$$E_{n-1} = w_n' E_n$$

Weights are updated at each pass with learning rate L:

$$w_n = w_n - L\frac{\delta E_{n+1}}{\delta w_n}$$

The model is run using stochastic gradient descent.