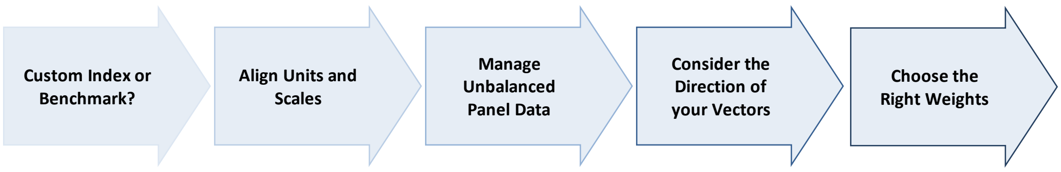
Despite a very large and diversified universe, it may happen that you don't find the index that meets all your expectations and requirements, constraining you to model a custom one. Within this paper, one highlights a few key technical and strategic aspects to consider when modelling such an index, including managing unbalanced panel data and aligning units and scales.

1. Introduction

It is very frequent using reference series or indices to present your arguments into a research paper; if you cannot find an index that meets all your expectations, it is very likely that you need to build a custom index by combining several series all together. Building such an index can be tedious and requires paying attention to several strategic and technical aspects.

The purpose of this document is stressing out a few important aspects to consider when building a custom index. This paper is organised around the below key elements:



From a semantic perspective - and avoid any confusion and misunderstanding - one defines the underlying series composing the custom index as vectors.

2. Custom Index or Benchmark?

It is vital to clarify at the beginning of the modelling process if you need a custom index, *i.e.* a standalone index that you monitor without any comparison with a portfolio, or a benchmark, *i.e.* an index dedicated to monitor the relative performances of an investment portfolio. A custom index offers you a total flexibility in the modelling process and minimizes the number of restrictions to consider; building a benchmark is more complex because it requires to translate the performance objectives into allocations, assess the Risks and the costs of the allocations, and potentially rebalance the benchmark. Modern literature flourishes about building benchmarks, within this paper we only focus on custom indices, without any performance monitoring considerations.

3. Align Units and Scales

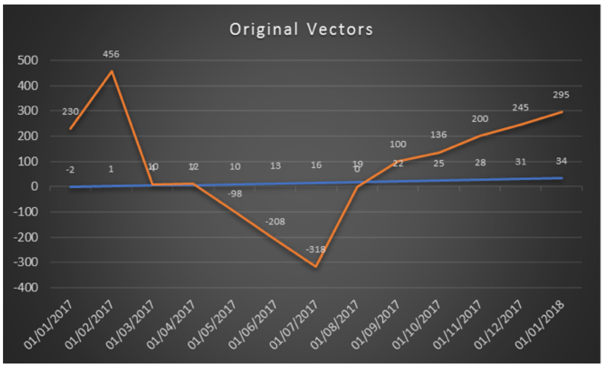
It might be tempting just to combine the underlying vectors all together in order to obtain your final index, paying little attention to the main features of those series such as the units of measurement or the scale; but the higher the number of underlying constituents you use to build your custom index, the more likely the units of those series will be different (*i.e.* constituents can be priced in currency, or in points, can be in levels or in variations, can be expressed on different scales, *etc.*) and the more likely you will have to normalize your vectors.

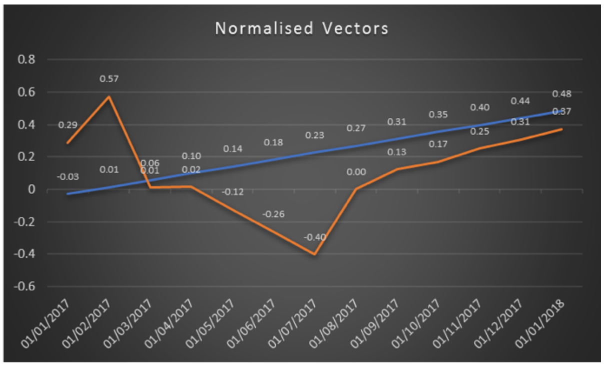
3.1. Normalizing vectors

According to Dodge [2003], "*Normalization means adjusting values measured on different scales to a notionally common scale*". Very often, normalization consists of considering the underlying series as vectors into a multidimensional space and transforming those vectors into a common vector whose norm is equal to an agreed value (usually one).

$$\textit{Consider the original vector } y = \begin{bmatrix} -2 \\ 1 \\ 4 \end{bmatrix}, \textit{where the norm vector is } \|y\| = \sqrt{-2^2 + 1^2 + 4^2} = 4.58$$

$$\textit{The normalised vector } \tilde{y} = \begin{bmatrix} \frac{-2}{4.58} \\ \frac{1}{4.58} \\ \frac{4}{4.58} \end{bmatrix} = \begin{bmatrix} -0.44 \\ 0.22 \\ 0.87 \end{bmatrix}, \textit{where the norm vector is } \|\tilde{y}\| = \sqrt{-0.44^2 + 0.22^2 + 0.87^2} \approx 1$$





3.2. Standardizing vectors

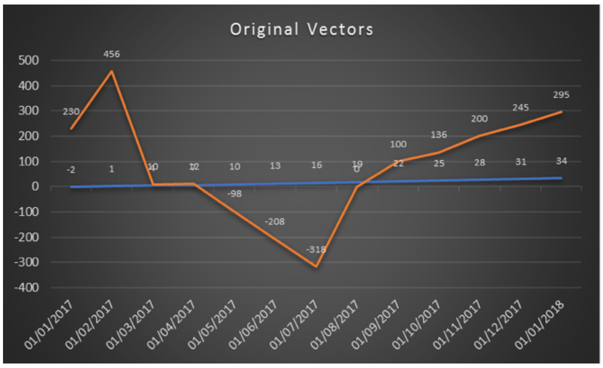
Standardization is a specific normalization type, where the unit of measurement is eliminated, and data is transformed into a new score, called Z-Score, with a mean of 0 and a standard deviation of 1.

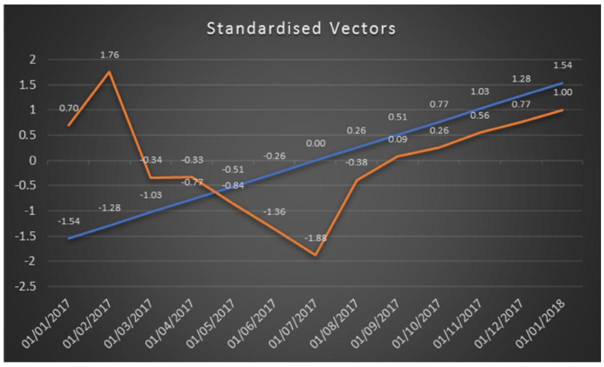
$$Z\ Score = \frac{y_n - \bar{y}}{\hat{\sigma}_n}$$

Let's consider the same example as in the previous sub-section, the mean of the vector "y" is 1, and the standard deviation is 3, then:

$$\textit{The standardized vector } \tilde{y} = \begin{bmatrix} \frac{-2-1}{3} \\ \frac{1-1}{3} \\ \frac{4-1}{3} \end{bmatrix}$$

$$= \begin{bmatrix} -1.00 \\ 0.00 \\ 1.00 \end{bmatrix}, \textit{and it is easy to demonstrate that the mean is } 0 \textit{ and standard deviation } 1$$



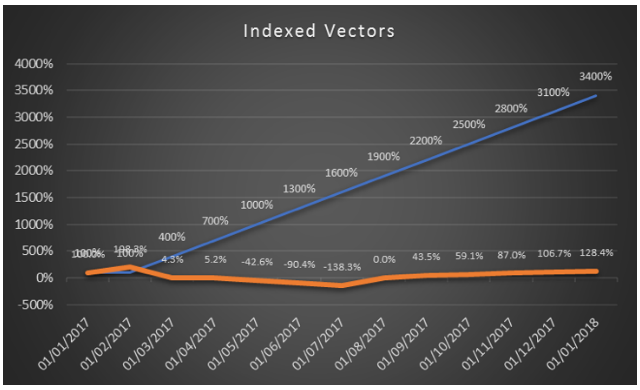


3.3. Indexing data to a shared starting point

Another approach would consist to compare the growth rates of the underlying vectors by indexing data to a common starting point; this allows to compare vectors with different amplitudes. Despite this approach is intuitive, it has several limitations, for instance if vectors have very different growth rates or alternate positive and negative observations.

$$Y_t = \left(\frac{y_t}{y_0}\right) * 100$$

Below, one can see - if we compare the usual underlying vectors - that the graph is not particularly intuitive because the blue vector growth rate is proportionally much higher than the amber one.

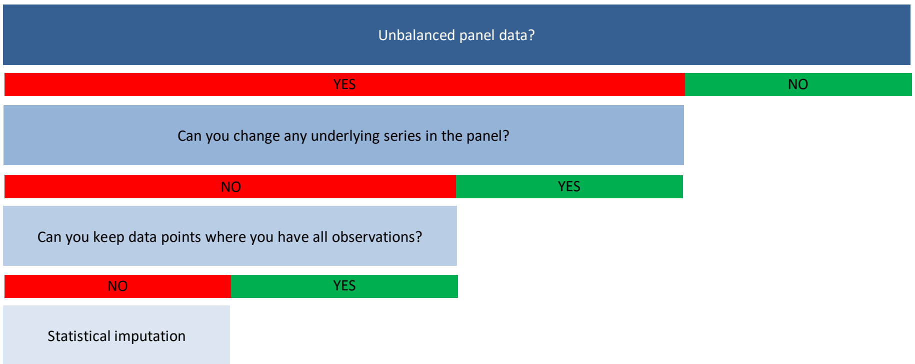


4. Manage Unbalanced Panel Data

In a way, building a custom index is very similar to a panel data analysis, *i.e.* observing several individuals or series over time. The larger the dimension of your vectors, the higher the probability to deal with unevenly spaced series (*aka* unequally or irregularly spaced series), *i.e.* to deal with series that haven't an observation at every data point.

Missing data can introduce a bias by diluting the information and make your custom index less intuitive; advanced methods can be used to resample irregular time-series onto regular matrices, such as the Compound Fourier Transform method, the Least-Squares Estimation, or the Matrix Inversion approach, but more simplistic techniques can also be used to handle unbalanced panels and provide a satisfactory and robust output.

In the below organigram, one highlights an approach that could be used to manage unbalanced panels while avoiding, as much as possible, the use of advanced resampling methods.



4.1. Can you change any underlying series in the panel?

If too many data-points are missing across your panel, a good solution could be to replace an underlying vector by a more appropriate one, for instance a vector with different time frequencies (*i.e.* from monthly to daily).

4.2. Can you keep data points where you have all observations?

This is the most objective and reliable method allowing to avoid any data interpolation/extrapolation, and that can easily be automated. It consists of keeping data points (*i.e.* days) where all your underlying vectors have an observation. A common shortfall is when too many data points are missing, in that case cleaning your sample will considerably reduce the total number of observed data-points and your custom index could become irrelevant.

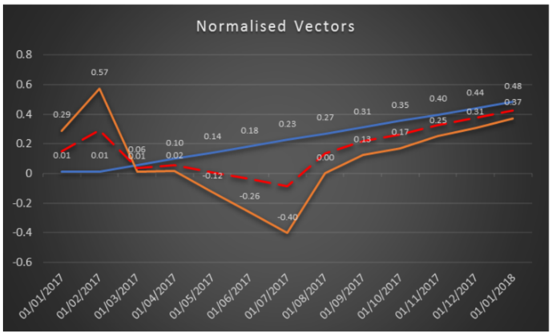
4.3. Statistical imputation and data points interpolation.

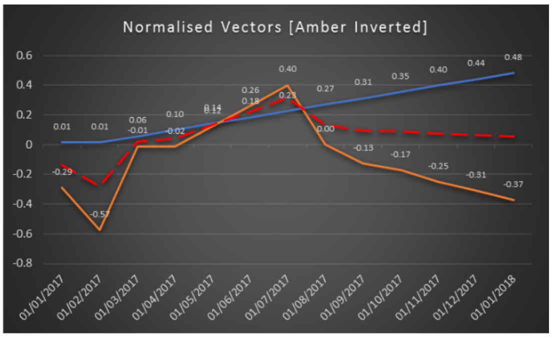
Statistical imputation consists of filling missing data with substitute values; the method can be as easy as replacing the point with a chosen value (*i.e.* average value). In the data interpolation process, R. Franke [1982] distinguishes the notion of global method from the concept of local method. Under the concept of global method, the interpolant is dependent on all data points, *i.e.* adding, deleting or changing any data point will propagate throughout the domain of definition; under the concept of local method, any of the above changes will only affect the interpolant at the nearby point, *i.e.* the interpolant will be unchanged at distances greater than some given distances.

Independently from your domain or definition, you also must decide if the interpolation will be linear or not. If only a few data-points are missing, it can be enough to use a linear interpolation, it offers as main advantages of being understandable by everyone and reducing the number of assumptions. If you nevertheless decide to use polynomial interpolations, the modern literature flourishes on this topic.

5. Consider the Direction of Your Vectors

Pay attention to the direction of your vectors when combining them all-together; *i.e.* imagine you combine two vectors (blue and amber line below) into a custom index (red line), if you want your custom index to move upwardly when both vectors increase, the red line in the first graph is the right representation of your custom index; if you want your custom index to increase when the blue vector increases and when the amber one decreases, you then have - as represented in the second graph - to inverse the amber line.





6. Choose the Right Weights

Because you don't necessarily want your underlying vectors to contribute in the same proportion to your final custom index, you often must decide the weights you will use *per* vector, it is often as easy as applying the below formula.

$$Custom\ Index = \sum_{i=1}^{n} Weight_i * Vector_i$$

You may apply quantitative or qualitative decision rules to arrive to "optimal" weights depending on your expectations.

7. Conclusion

This paper was dedicated to highlight a few key aspects to consider when modelling a custom index. More importantly, one expected to show that building an index is more complicated than assembling blocks all together and that an important data cleaning process must be performed to obtain an intuitive and reliable index.

8. References

- Abdi H., 2010, "Normalizing Data", The University of Texas at Dallas, Working Paper ;
- Adorf H-M., 1995, "Interpolation of Irregularly Sampled Data Series - A Survey", *Astronomical Data Analysis Software and Systems IV*, ASP Conference Series, Vol. 77 ;
- BlackRock Portfolio Solutions, "Building Better Benchmarks" ;
- Dodge Y., 2003, *The Oxford Dictionary of Statistical Terms* ;
- Eckner A., 2014, "A Framework for the Analysis of Unevenly Spaced Time Series", Working Paper ;
- Franke R., 1982, "Scattered Data Interpolation: Tests of Some Methods", *Mathematics of Computation*, Volume 38, Number 157 ;
- Mayer M., 2010, "Unbalanced Panel Data Models", Universität Wien, Presentation Slides ;
- Ryavec C., 2004, "A non-linear interpolation formula", *Journal of Mathematical Analysis and Applications*, Volume 87, Issue 2, pp. 467-473 ;
- Scott Armstrong J., 2001, "Extrapolation for Time-Series and Cross-Sectional Data", University of Pennsylvania, Marketing Paper.