17 February 2025 · CC-BY 4.0

Qeios

Peer Review

Review of: "Compositional Data Analysis for Modelling and Forecasting Mortality with the α-Transformation"

Michael Greenacre[1]

[1]. Economics and Business, Universitat Pompeu Fabra, Spain

This article analyses mortality data, treated as compositional, using one of the alternative methods for analysing compositional data, the alpha-transformation. This review raises the following questions, which have not been fully explained.

1. Why are mortality data compositional?
2. How is the alpha-transformation justified to tackle the analysis of mortality data?
3. Since the approach is model-based, where are the interpretable results, apart from the overall RMSE MAE statistics?

Why compositional?

Not being accustomed to working with mortality data, I tried to understand why the authors think they are compositional. To my intuitive way of thinking, mortality concerns numbers of deaths, usually normalized (e.g., per 10000 of population), and it is the raw figures that are of interest, not their relative values. The authors say the compositional approach is "pioneered by Oeppen[6]", which is a conference paper in 2008 and available online. I looked at this paper, which is apparently not published. Its TLDR summary starts like this: "This paper abandons the conventional approach of using log mortality rates and forecasts the density of deaths". Looking at this paper, it seems that for life tables the so-called radix values are fixed and the life table decomposes this radix total across the age groups. Because of this, the data are deemed compositional. This still doesn't explain why the data need to be treated by logratios, which are the basis of compositional data analysis (CoDA). CoDA was developed for data such as geochemical data, where the elements in a sample were a subset of the possible elements that could potentially have been included, depending on the sophistication of

doi.org/10.32388/RQN1VR

measuring instruments and the research question. Such data sets, which are essentially subcompositions, are found in many fields, such as biochemistry, genetics, and microbiomics. Ratios of components remain constant irrespective of the subset of components under study. But in the case of life tables, the composition is across the full spectrum of ages – these are thus complete compositions, so the issue of subcompositional coherence in Aitchison-style CoDA is no longer relevant. It is unlikely (to my knowledge) that some researchers would use selected age groups and reproportion the data to sum to 1. Hence, in my opinion, in spite of this tendency to CoDA-fy mortality data, the regular way of modelling the counts directly, or by log-transformation as usually done, seems simpler and more appropriate; e.g., the generally accepted LC model seems quite reasonable. The fact that the proportions of this full composition sum to 1 is not a problem for data analysis. Hence, the authors have to convince me, and the readers, why the mortality data they treat should be treated as compositional data in the Aitchison sense. And there is also the Dirichlet modeling approach for complete compositional data such as these, which are not mentioned or considered.

Why alpha-transformation

Given that the data should be treated compositionally, which still has to be fully justified, the authors describe the present CoDA tools for analyzing compositional data. Here, the centered logratio transformation (clr) is described, and the problem of zeros for any logratio-type transform is mentioned as the obvious drawback. The authors then propose to use Tsagris' alpha-transformation, which is a Box-Cox style power transform designed to converge to the so-called isometric logratio (ilr) as the power tends to 0 if there are no data zeros. But the idea is to use it with a nonzero alpha selected according to a wider objective. This is now a much more complicated workflow to model what were basically a simple set of probabilities in a lifetable! Choosing this approach raises a whole lot of other questions, which have not been dealt with. What has the isometric logratio got to do with life-table probabilities? The ilrs are a set of linear transformations on the log-transformed data, and there are millions of possible sets of ilrs. The authors gloss over this and report no H matrix that defines the transformation. Notice that the life table categories are ordered: so, does the choice of H take this order into account? Is the choice important? Will the same alphas emerge from different choices of H matrices? If these questions have been answered in the literature, citations should be given. As a matter of interest, a recent paper on the so-called chiPower transformation explains a similar Box-Cox style power transformation but uses the clr, not the ilr. There is only one clr, so all these doubts about how to choose H, etc., are avoided. Another doubt emerges when the authors report alpha values

doi.org/10.32388/RQN1VR

equal to 0. This is impossible if there are zeros in the data, since we are at the convergence point (ilr in this case), which is undefined when there are data zeros. Or are the authors substituting the zeros and then applying the alpha-transformation? (which would defeat the object of its use...). The alpha values of zero are a mystery.

The alpha transformation, like the chiPower transform, is useful in other fields, where subcompositional coherence is indeed an issue (i.e., the given compositional data are in fact composed of subcompositions) and thus logratio methods are desirable, but there are lots of zeros.

The results

The plots are very difficult to read; the font size for the lettering is very small. I don't understand how it is possible to compare the alpha method with clr (e.g., Table 3), since clr needs zeros to be replaced, so it is working with different data than the alpha approach. Are the RMSEs comparable then? All the error measures are tiny; I presume they are errors on estimates of proportions, so maybe the table's values can be "eased" a bit by multiplying by 100 so they are effectively on a percentage scale. Table 4 too.

A specific example of how an alpha was estimated could be given; in the present article, the estimated alphas are simply reported.

I find this conclusion over-optimistic: "The results show that the model fitted to the alpha-transformed data has a comparable and superior performance in most of the selected countries as compared to the CLR-transformed data." All the results show an extremely similar pattern of results, just tiny differences. I'm still puzzled, however, how alpha-transform and clr can be compared: are errors in the clr-transform version evaluated on the data values that were replaced because originally they were zero?

In summary, I'm not convinced that the alpha-transformation approach, with its "mystery" ilr transform, is casting any new light on the analysis of mortality data; neither am I convinced that CoDA and the CLR are either. It may be that the existing way of analysis, which has not been used here as a comparison, is quite sufficient, and the CoDA approach is more hype than science in this application field.

Minor comments

On page 5, there is all this alternative notation for the algebra in the simplex, which is totally unnecessary in this basically applied article. It should all be removed; just the closure operator can be retained. Later on, there is no need to refer to the perturbation operator, since what is being done is clear (and in fact simple), e.g., subtracting the geometric mean, adding it back...

Page 5: I doubt readers will immediately recognize what a “Helmert sub-matrix" is! This whole issue of the H matrix and ilr is a distractor. One could just do the regular Box-Cox power transformation and see which power is optimal. Why don't you try just that and see what you get?

Early on, “fine-tuned alpha” is mentioned, and we have no idea what is meant until the application later. As mentioned earlier, an example of the process of “fine-tuning” is required.

What can be learned from Table 2?

How can the results in Table 4 be made clearer?

Page 13, line 5. At the limit (alpha=0), when the “accuracy is similar,” they should be the same; the information in the ILR and the CLR is identical.

Attachments: available at https://doi.org/10.32388/RQN1VR

Declarations

Potential competing interests: No potential competing interests to declare.