PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Do alternative weighting approaches for an Index of Multiple Deprivation change the association with mortality? A sensitivity analysis from Germany.
AUTHORS	Schederecker, Florian; Kurz, Christoph; Fairburn, Jon; Maier, Werner

VERSION 1 - REVIEW

REVIEWER	Pablo Cabrera-Barona IAEN / USFQ ECUADOR
REVIEW RETURNED	17-Dec-2018

GENERAL COMMENTS	Please clearly describe in the manuscript the statistical results of your evaluation of normality of residuals, homoscedasticity of residuals, and multicollinearity

REVIEWER	Gary Abel University of Exeter
REVIEW RETURNED	08-Mar-2019

GENERAL COMMENTS	Thank you for asking me to review this paper. It presents a piece of work considering different approaches to weighting domains of multiple deprivation scores, specifically applied to the German Index of Multiple Deprivation (GIMD). Whilst this paper is generally clearly written I do find it a little lacking in motivation for why different weighting methods would be needed. It does make the case that there is no clear consensus on the best way to do things, but the introduction makes little attempt to consider conceptually what the ideal weighting would do. Without understanding what the best weighting would look like I find the paper more of a technical exercise than being truly revealing. Implicit in the method used to assess weightings is that predicting mortality is a sign of a good set of weights. Unsurprisingly the method used which optimises the relationship showed the strongest association between the index and mortality. By the time one reaches the conclusion one realises that the paper is in fact presented as investigating the stability of the IMD to different weighting choices. Or in other words does the conceptual framework behind weighting of domains matter. If this really is the aim of the paper, presenting this upfront would really improve the readability of the paper, framing it for the reader from the off. In any case this question could have been answered simply

	by investigating the sensitivity of the index to systematic changes in weightings. Indeed I do wonder if the whole exercise could have been avoided by simply noting that there was a high correlation between individual domains and thus any weighting scheme would likely give highly correlated results. For example our recent work with UK IMDs showed that 94% of the variance in the English IMD could be explained by the income and employment domains alone even though they only had weight of 22.5% each in the overall index [Abel et al, 2016, BMJOpen]. The proportion of variance explained was the same or higher in the other UK countries. The implication of this finding being that even if the weights for the other domains had been zero, there would have been very little impact on the overall index. In table 1 there are some comments on the advantages and disadvantages of different methods. However, these tend to focus on technical aspects rather than conceptual ones. For example, what PCA, EFA and CFA all have in common is that they all assume the existence of an unmeasured unifying concept, but make no prior judgements as to what that is. If we contrast that with predictive approaches (e.g. regression or the greedy algorithm used) then some judgement is made that whatever is being predicted is, in some way, the ultimate arbiter of the concept of deprivation trying to be constructed. Normative approaches, instead use a theoretical framework to decide what the concept is. I think this could be explored some more to give some real meaning to the paper. Alternatively you could pose the question, what do we learn about the stability of IMDs to weighting choices, simply by considering different methods compared to just changing each weight in turn by 50% and seeing what happens. One big problem I have with the paper is the conflation of using a greedy algorithm vs linear regression and the measure used as an outcome (mortality vs living space). One is left unable to judge which is making the most difference. I personally would be surprised if the choice of algorithm mattered much here. Unfortunately no details are given for the greedy algorithm and an unpublished paper is cited. It is described as yielding “weights for the domains close to the maximum possible correlation between the GIMD 2010 and mortality”. In the circumstance where only a simple sum of weighted domains are used I am at a loss as to how this differs at all from linear regression which works to minimise the sum of squares of residuals, or in other words maximise the correlation. Fancy algorithms and AI are currently very popular, but I fear often they are promoted as being superior on the basis of unfair comparisons and I would not like to see that done here. I would recommend deriving weights for either the greedy algorithm with living space as the outcome or linear regression with mortality as the outcome (or indeed both) to separate the analytic technique from the choice of outcome data. Minor points In methods when describing the Equal weighting approach the paper states “For this approach, an equal effect of all deprivation indicators is assumed”. Rather than “equal effect” I wonder if this should be “equal importance”? The derivation/source of the living space variable is not described in the data section. Method 5 is described as an exploratory factor analysis. However,

	given the structure of the model was pre-determined (i.e. a single factor with loadings from all domains onto that factor) this strikes me to really be a confirmatory factor analysis. Related to this I wonder if any of the examples labelled in Table 1 as EFA allowed for any structure other than a unidimensional one? The methods for calculating SMR are presented in a supplement (which is fine), however, this is only introduced in the results, whereas it should have been introduced in the methods. No methodology for the review is given. I don't think there is a need for this to be a systematic review (nor conform to such reporting) but some idea of the scope would be useful to answer questions such as how complete is the review? And are there likely to be other techniques out there? I disagree with some of the assertions made in Table 1. For example to say that logistic regression derives weights directly from the data is misleading as it depends on the outcome variable used. The disadvantage with Basian factor analysis that "Derived coefficient dependent on the data quality" must surely apply to all empirical approaches. Also does not the observation that "Reduction to one factor does not consider the multidimensionality of deprivation" apply to all PCA, and factor analysis approaches? And even if it does, the ultimate aim of reducing deprivation to a single number in all approaches suffer from this, i.e. it is an issue with the idea of a single index rather than a particular approach. I also struggle with revealed preferences being described as an empirical method. Essentially it is a normative one whereby the importance as decided by government of the different domains is implied by spending. In the discussion, in the sentence "The other deprivation domains showed a dispersion of at least 14 percentage points" should replace the word "dispersion" with "range" as there are many different measures of dispersion. Also security deprivation has a range of only 13 points. I don't see why both standard deviation and variance are shown in Table 3 as one is a simple transformation of the other I think the authors should be more upfront about the fit of their factor analysis model. An RMSEA of 0.32 with tight confidence intervals shows pretty strong evidence it is not a good fit to the data. Equally I've never seen a TLI of 0.5 described as moderate. Having said that does it matter? I'm not sure it does for this paper, but it should not be made out to be better than it is. Supplement 5 contains results and yet is not referred to in the results. In the discussion the paper notes that when comparing correlation coefficients "even small non-relevant differences could have produced significant results" and suggested that using significance alone seemed inappropriate. I agree with this statement and wonder if it is at all worthwhile. What certainly seems overkill is to present 3 different significance testing analyses. In the discussion equal weighting is considered obsolete based simply on theoretical grounds. In this case I do wonder why it features so heavily in the paper?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 – Comments:		Authors' responses:
Please clearly describe in the manuscript the statistical results of your evaluation of normality of residuals, homoscedasticity of residuals, and multicollinearity	1	For the assumptions of the linear regression, we discussed normality of residuals, homoscedasticity, multicollinearity, autocorrelation, and nonlinearity for the domains. We calculated robust standard errors since some heteroscedasticity was present. Please find the respective results in Supplement 4.
Reviewer 2 – Comments:		Authors' responses:
Thank you for asking me to review this paper. It presents a piece of work considering different approaches to weighting	1	Thank you very much for addressing this point. What is in general the aim of a weighting approach? Firstly, we want to see the potential influence of every
domains of multiple deprivation scores, specifically applied to the German Index of Multiple Deprivation (GIMD). Whilst this paper is generally clearly written I do find it a little lacking in motivation for why different weighting methods would be needed. It does make the case that there is no clear consensus on the best way to do things, but the introduction makes little attempt to consider conceptually what the ideal weighting would do. Without understanding what the best weighting would look like I find the paper more of a technical exercise than being truly revealing.		variable or domain on the output variable, in our case mortality. Secondly, how can we derive weights reflecting the public importance of the different domains? Weighting approaches for IMDs have already been discussed in the literature, however there seems to be no gold standard and additional research is still required in this field (cf. Bradshaw 2003; Dibben et al., 2004; Watson et al., 2008). We have already pointed this in the introduction section (see page 4f).
Implicit in the method used to assess weightings is that predicting mortality is a sign of a good set of weights. Unsurprisingly the method used which optimises the relationship showed the strongest association between the index and mortality.	2	The aim of including the maximization algorithm was to assess the maximum possible correlation and the algorithm does not necessarily present the ideal weighting method.
By the time one reaches the conclusion one realises that the paper is in fact presented as investigating the stability of the IMD to different weighting choices. Or in other words does the conceptual framework behind weighting of domains matter. If this really is the aim of the paper, presenting this upfront would really improve the readability of the paper, framing it for the	3	We have integrated in the scope of our analysis in the introduction section (page 5) the following paragraph: "As the GIMD was weighted by experts following the model of the British IMDs, we conducted a sensitivity analysis for the domain weighting of the GIMD following the example of Dibben et al. [13]. The aim of our study was to test the stability of the GIMD to different weighting approaches by conducting

reader from the off. In any case this question could have been answered simply by investigating the sensitivity of the index to systematic changes in weightings. Indeed I do wonder if the whole exercise could have been avoided by simply noting that there was a high correlation between individual domains and thus any weighting scheme would likely give highly correlated results. For example our recent work with UK IMDs showed that 94% of the variance in the English IMD could be explained by the income and employment domains alone even though they only had weight of 22.5% each in the overall index [Abel et al, 2016, BMJOpen]. The proportion of variance explained was the same or higher in the other UK countries. The implication of this finding being that even if the weights for the other domains had been zero, there would have been very little impact on the overall index.		correlation analyses with mortality as a key health outcome." We added to the limitations in the discussion section the following paragraph (pages 24f): "We are aware that the stability of the GIMD could have also been tested by applying systematic changes to the weighting of the GIMD domains without using a framework of different weighting approaches. The correlation between some deprivation domains (e. g., income or employment) is relatively high and thus any weighting scheme would likely give highly correlated results with mortality. A recent study from the UK showed that 94% of the variance in the English IMD could be explained by the income and employment domains alone, even though they had weights of 22.5% each in the overall index. The authors stated that even if the weights for the other domains had been zero, there would have been very little impact on the overall index [47]. Nevertheless, the aim of our study was to provide a conceptual framework of weighting approaches (normative and empirical) for an index of multiple deprivation and to combine the results of the literature search with a sensitivity analysis based on the GIMD."
In table 1 there are some comments on the advantages and disadvantages of different methods. However, these tend to focus on technical aspects rather than conceptual ones. For example, what PCA, EFA and CFA all have in common is that	4	We have found different predicitive and non-predictive methods in the existing literature. We agree on your distinction of predictive and non-predictive algorithms and made this more clear in our manuscript having added the following paragraph in the discussion section (page 21):

they all assume the existence of an unmeasured unifying concept, but make no prior judgements as to what that is. If we contrast that with predictive approaches (e.g. regression or the greedy algorithm used) then some judgement is made that whatever is being predicted is, in some way, the ultimate arbiter of the concept of deprivation trying to be constructed. Normative approaches, instead use a theoretical framework to decide what the concept is. I think this could be explored some more to give some real meaning to the paper. Alternatively you could pose the question, what do we learn about the stability of IMDs to weighting choices, simply by considering different methods compared to just changing each weight in turn by 50% and seeing what happens.		"A further distinction of the methods can be made regarding their conceptual aspects. Factor analysis and PCA are unsupervised methods that require no prior judgements and construct deprivation solely based on the domain knowledge. On the other hand, linear regression and the maximization algorithm are supervised or predictive methods considering deprivation based on a specific proxy and assuming a relationship between this proxy and deprivation."
One big problem I have with the paper is the conflation of using a greedy algorithm vs linear regression and the measure used as an outcome (mortality vs living space). One is left unable to judge which is making the most difference. I personally would be surprised if the choice of algorithm mattered much here. Unfortunately no details are given for the greedy algorithm and an unpublished paper is cited. It is	5	The reason why we included the greedy maximization algorithm was to get domain weights that would maximize the correlation to mortality. A linear regression model minimizes the sum of squared residuals, but not directly correlation. We agree that this difference might be subtle, but in our opinion including this algorithm presents an interesting additional weighting scheme. As suggested by the reviewer, we conducted an additional analysis for the linear regression with mortality and premature mortality as outcomes and it

described as yielding “weights for the domains close to the maximum possible correlation between the GIMD 2010 and mortality”. In the circumstance where only a simple sum of weighted domains are used I am at a loss as to how this differs at all from linear regression which works to minimise the sum of squares of residuals, or in other words maximise the correlation. Fancy algorithms and AI are currently very popular, but I fear often they are promoted as being superior on the basis of unfair comparisons and I would not like to see that done here. I would recommend deriving weights for either the greedy algorithm with living space as the outcome or linear regression with mortality as the outcome (or indeed both) to separate the analytic technique from the choice of outcome data.		has yielded different weights for the domains in comparison to the algorithm (e.g. employment: 2% vs. 21%; income: 20% vs. 18%, municipal income: 37% vs. 28%). The spearman correlation was also different in comparison to the algorithm (mortality: r=0.58 vs. 0.61 and early mortality: r= 0.75 vs. 0.83), indicating that the maximization algorithm finds the optimal correlation. When we used living space as an outcome for the maximization algorithm in an additional analysis, it showed different results in comparison to the linear regression (employment: 34% vs. 22% and education: 7% vs. 15%, income was pretty much the same) and in particular a lower correlation with mortality (r=0.35 vs. 0.56) and premature mortality (r=0.54 vs. 0.74), suggesting that a different weighting can indeed make a difference regarding the correlation of deprivation domains with mortality. We did not intend to promote the greedy maximization algorithm as a superior technique in comparison with linear regression, it was rather intended to use it as a benchmark to show the weighting for a given measure (e. g., correlation with a given outcome). We had already mentioned in our discussion section that the use of an algorithm should not be taken as a gold standard in domain weighting: “A transfer of the algorithm to other areas of interest is possible without difficulty but should be used mainly for orientation, which is possible concerning a selected measure, given a data set.”

Minor points		
In methods when describing the Equal weighting approach the paper states “For this approach, an equal effect of all deprivation indicators is assumed”. Rather than “equal effect” I wonder if this should be “equal importance”?	6	Thank you for pointing this out, we have changed the wording to „equal importance”.

The derivation/source of the living space variable is not described in the data section.	7	We have now added the source of the variable 'living space' in the data section (p. 7): "We used the variable 'available living space per inhabitant' from the German Federal Statistical Office from 2010 [18] as a proxy of deprivation. We reversed the polarity of its values and thus make it more comparable to the GIMD scores."
Method 5 is described as an exploratory factor analysis. However, given the structure of the model was pre-determined (i.e. a single factor with loadings from all domains onto that factor) this strikes me to really be a confirmatory factor analysis. Related to this I wonder if any of the examples labelled in Table 1 as EFA allowed for any structure other than a unidimensional one?	8	We agree that this can be seen as a confirmatory factor analysis, however, we leant on the examples in table 1 for the factor analysis. In other studies using EFA (Cesaroni et al. 2006, Testi et al. 2009, Alvarez-Del Arco et al 2013), the authors used EFA to construct a unidimensional factor acting as a deprivation surrogate.

The methods for calculating SMR are presented in a supplement (which is fine), however, this is only introduced in the results, whereas it should have been introduced in the methods.	9	Thank you for pointing this out, we have moved the introduction of the SMR calculations in the methods section (page 7).
No methodology for the review is given. I don't think there is a need for this to be a systematic review (nor conform to such reporting) but some idea of the scope would be useful to answer questions such as how complete is the review? And are there likely to be other techniques out there?	10	We have searched in PubMed and Embase using keywords and limits. We have now added the keywords of the literature search in the method section (page 7). "We searched relevant literature in the databases PubMed and Embase [e.g., keywords used in PubMed: (deprivation OR deprived) AND (index OR indices) AND (area* OR region* OR neighborhood OR neighbourhood), limits: English OR German OR French OR Italian OR Spanish.]"

I disagree with some of the assertions made in Table 1. For example to say that logistic regression derives weights directly from the data is misleading as it depends on the outcome variable used. The disadvantage with Basian factor analysis that “Derived coefficient dependent on the data quality” must surely apply to all empirical approaches. Also does not the observation that “Reduction to one factor does not consider the multidimensionality of deprivation” apply to all PCA, and factor analysis approaches? And even if it does, the ultimate	11	Thank you very much, we agree that this must be made more clear. We changed the logistic regression description in table 1 to: “derives outcome specific weights from the data”. For Bayesian factor analysis, we changed it to: “At least weakly informative prior information is required” and as a second disadvantage: “Computationally more expensive” We also agree on your observation, that all “Reduction to one factor does not consider the multidimensionality of deprivation” applies to all factor analysis

aim of reducing deprivation to a single number in all approaches suffer from this, i.e. it is an issue with the idea of a single index rather than a particular approach. I also struggle with revealed preferences being described as an empirical method. Essentially it is a normative one whereby the importance as decided by government of the different domains is implied by spending.		approaches and we changed the description accordingly. We agree that the revealed preferences can be seen as a normative approach, however, we derived the classification as an empirical approach from Dibben et al., 2013.
In the discussion, in the sentence “The other deprivation domains showed a dispersion of at least 14 percentage points” should replace the word “dispersion” with “range” as there are many different measures of dispersion. Also security deprivation has a range of only 13 points.	12	Thank you for pointing this out, we have changed the wording accordingly to your suggestion.
I don’t see why both standard deviation and variance are shown in Table 3 as one is a simple transformation of the other	13	We now show only the standard deviation in Table 3.
I think the authors should be more upfront about the fit of their factor analysis model. An RMSEA of 0.32 with tight confidence	14	Thank you for pointing this out, we have changed the wording in the results

intervals shows pretty strong evidence it is not a good fit to the data. Equally I’ve never seen a TLI of 0.5 described as moderate. Having said that does it matter? I’m not sure it does for this paper, but it should not be made out to be better than it is.		section (page 17) as follows: “The factor analysis generally had significant explanatory power (Chi-square: 584.65, $p < 0.0001$), but showed low reliability (Tucker–Lewis index = 0.50) and a RMSEA of 0.32 with tight confidence intervals (0.30- 0.34) indicated that this one factor was not a good fit to the data (Supplement 5).”

Supplement 5 contains results and yet is not referred to in the results.	15	We now refer to this supplement (which is now supplement 7) in the results section. We have added the following statement: "When we corrected for the multiple comparison of the difference of the correlation between the GIMD versions, there was a slight difference present in the significances (Supplement 7)."
In the discussion the paper notes that when comparing correlation coefficients "even small non-relevant differences could have produced significant results" and suggested that using significance alone seemed inappropriate. I agree with this statement and wonder if it is at all worthwhile. What certainly seems overkill is to present 3 different significance testing analyses.	16	We agree with the reviewer and have removed table 5 from the main manuscript and added it as supplement 6.
In the discussion equal weighting is considered obsolete based simply on theoretical grounds. In this case I do wonder why it features so heavily in the paper?	17	Equal weighting is a widely used fundamental weighting method. Also, the first British deprivation indices used equally weighted indicators (Townsend index, Carstairs index etc.). Therefore, we found it important to integrate this approach in our sensitivity analysis.

VERSION 2 – REVIEW

REVIEWER	Pablo F. Cabrera Barona IAEN / USFQ
REVIEW RETURNED	22-Apr-2019

GENERAL COMMENTS	The authors have addressed all my comments