Review - A consistent dataset for the net income distribution for 190 countries, aggregated to 32 geographical regions and the world from 1958-2015

The paper puts together a world income data set drawing on several global repositories and filling the data gaps with a regression and a Principal Component Analysis (PCA) approach. The authors obtain a world database for 190 countries and 32 geographical regions.

The paper contributes to a growing literature on world databases of incomes. The data sources selected are appropriate but insufficient. There is no mention of the work led by Branko Milanovic in the US and Thomas Piketty in France. Both these authors oversee global data efforts to measure income and income inequality that should be reviewed/acknowledged by the authors.

The central contributions of the paper are the regression-based approach to estimate net incomes from consumption data and the PCA approach used to estimate income data from national Gini coefficients.

The regression-based approach (section 2.3) is popular but there are better methods that should provide better fit including quantile regressions and random forest methods. The problem with simple OLS is that they are very poor at predicting incomes on the tails of a distribution (the predicted income distribution is always much narrower than the original income distribution). The fact that the authors run different regressions for each decile accentuates this problem by creating discontinuities between deciles. This problem can be overcome with quantile regressions or, better, with random forest. The probabilistic nature of random forest fits the tails if income distributions much better than standard OLS. I would recommend the authors to test both methods and compare results with the current ones.

The PCA method (section 3.4) is somewhat unconventional for this specific literature. This, per se, is not a critique, but it does require validation beyond what the authors offer. Here I would suggest taking the entire net income distribution for a few countries where these data are publicly available, calculate the net income deciles and the Gini coefficient, plug this Gini into equations 3) and 4) and compare the resulting estimated deciles with those calculated from the full data. Also, it is important to clarify where the coefficients in equations 3 and 4 come from. I could not find the model and the results of the "equation estimated on 1659 observations".

The revisions suggested above are substantial and I would recommend the authors to cut out of the paper the work on regions, which is important for the GCAM but a distraction from the main objective of the paper. Instead, the regional work could be the object of a separate paper. This strategy would also allow the authors to target different audiences better.