The Relationship(s) Among Health & Social Problems vs. Income and Income Inequality: CS 521 (Fall 2018)

Warren D. (Hoss) Craft
12/13/2018

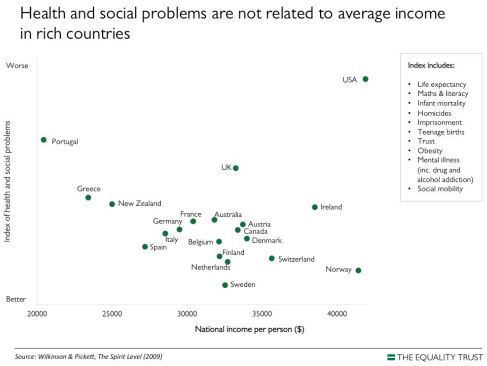
1. Introduction

Wilkinson & Pickett (2009) argued rather forcefully that the physical and mental health of a society, and societal problems such as violent crime, infant mortality, and even mistrust felt among citizens toward each other, are more clearly related to income *inequality* within the society than actual income *levels* within a society. Their basic results are captured in *Figure 1* below, which shows two of their own widely-disseminated figures purporting to show that an index of health and social problems is related more strongly to income than to absolute levels of income, at least for relatively prosperous countries such as those in the OECD. The data, and their conceptual argument about the logic of the relationship they are proposing, can be compelling, and the concept of income inequality has received increasing attention from researchers and policy-makers alike around the world.

In this project, I sought to replicate Wilkinson & Pickett's general results, both at the international level and at the level of States within the U.S., and begin an exploration of related concepts at the U.S. county level, all using more current data if available (much of their original work involved census and census-like data from around 2000). The basic problem posed is this: *is societal health and welfare related more strongly to measures of income, or to levels of income inequality?* The basic question is coming into more and more prominence and gaining more and more attention as social, economic, and political factors around the globe contribute to ever increasing levels of income inequality.

"What matters," Wilkinson & Pickett (2014) write, "is where we stand in relation to others in our own society."

> Inequality, not surprisingly, is a powerful social divider … The importance of community, social cohesion, and solidarity to human well-being has been demonstrated repeatedly in research showing how beneficial friendship and involvement in community life are to health. Equality comes into the picture as a precondition for getting the other two right. Not only do large inequalities produce problems associated with social differences and the divisive class prejudices that go with them, but they also weaken community life, reduce trust, and increase violence.



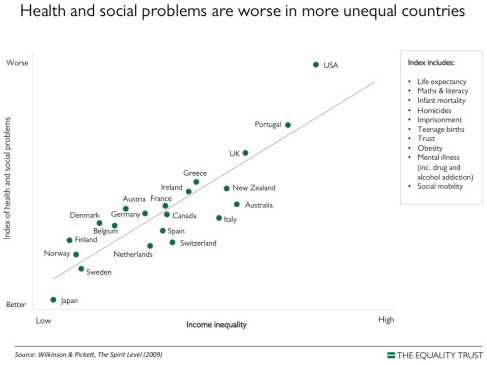


Figure 1. From Wilkinson & Pickett (2009): an index of health and social problems (HSPI) vs. per-capita income (upper panel) or income inequality (lower panel) for 20 OECD countries. Wilkinson & Pickett argue that income inequality rather than absolute income level is a better predictor of health and social problems.

2. Data

The data for this project come largely in relatively consolidated form from large aggregator institutions such as the U.S. Census Bureau, the Organization for Economic Cooperation and Development (OECD), the World Bank, the World Health Organization

(WHO), and United Nations Office on Drugs and Crime (UNODC), *etc*. The primary challenges from a data-mining perspective involved not the size of the data collections but the need to track down, extract, and combine multiple diverse data sets; the derivation of an aggregate index of health and social problems, despite corrupted, missing, or unavailable data; and (eventually) the initial exploration of important subsets of features in the search for possible clustering of community health-related vectors at the county level.

Tracking down the right data, or suitable substitute data was surprisingly time-consuming — in part because the data is often not available in easily downloadable form. A particularly unfortunate example of this was the enticing data offered by the *General Social Survey*, administered by NORC at the University of Chicago, and covering an amazing breadth of questions administered across the country. General downloads of the raw data, however, are only available in SPSS and STATA format. A more fortunate example is that of the OECD's *Excel*-friendly data detailing the past 20 years' or more of results for the internationally administered Program for International Student Assessment (PISA), giving detailed results on all components of the tests for all age groups, every year. Another fortunate example, used in this project for the beginning of a cluster analysis on health-related data at the U.S. county-level is the Robert Wood Johnson Foundation's County Health Rankings (CRR) National Data. The dataset includes dozens of detailed health-related measures aggregated at the county level for over 3,000 counties nationwide.

Even such a dataset, however, required effortful manual editing to make it accessible to MATLAB machinations. In the subfolder /DATA submitted with this project, you will find most of the manually-edited results (instead of the often much larger more raw data files), which are then accessible for MATLAB loading and processing.

3. Algorithms & Other Processing

As mentioned above, much of the effortful work was accomplished at the "front-end" portion of the data-mining process: the search, aggregation, cleaning, and eventual combining of a large number of diverse groups of datasets, in particular for the eventual combining into a collection of measures at both the international and U.S. State levels to allow the computation of Wilkinson & Pickett's Health and Social Problems Index (HSPI, as described in great detail in their 2009 book, but in helpful abbreviated version at https://www.equalitytrust.org.uk/notes-statistical-sources-and-methods).

The HSPI was interesting and educational to compute, involving not only the aggregation/combination of 9–10 dimensions of disparate data, but then the z-score normalization of the values, and finally an averaging over the z-scores for each country/State. This was an exciting and interesting application of related methods from early on in the semester in the Data Mining course (the computational details can be seen in the trio of related .mlx MATLAB files dedicated to the HSPI computations).

This painful aggregation process, and the eventual combination of data sources into a functional index of Health and Social Problems can be seen developed in the succession of MATLAB files:

```
CS521Project_HSPIData.mlx
CS521Project_HSPICalculation.mlx
CS521Project_HSPIvsIncome.mlx
```

Then the county-level data were re-organized from the raw data files available from the County Health Rankings National Data to produce a 20-dimensional, roughly 3,000-county data set, which was then subjected to a Principal Component Analysis reduction to 3 primary components, and the resulting 3D dataset subjected to a clustering analysis using the Partitioning Around Medoids (PAM) approach. This represented just an initial exploration of the possible use of such methods for characterizing the multidimensional county-level health-related data. The details of the PCA reduction process and PAM analysis can be seen in the MATLAB file:

```
CS521Project_CountyData.mlx
```

4. Results

Health & Social Problems Index (HSPI) vs. Income or vs. Income Inequality

Some of the primary results are shown in Figures 2 and 3 below, plotting our version of the HSPI vs. Income or vs. Income Inequality for 21 OECD countries. For the income-levels relationship, we see a result very similar to the earlier result from Wilkinson & Pickett (2009). This is satisfying in the sense that our efforts to

recreate the HSPI using more current data seems successful, but also satisfying in that it allows us to also more confidently criticize the claim that there is a weak relationship here. Both the U.S. (pt 21) and Japan (pt 12) stand out as clear outliers against a pretty clear inverse relationship holding for the other countries. Indeed, when the outliers are removed, the best-fit straight line r^2 value pops up from a dismal 0.055 to a respectable 0.34. So we have managed to replicate the original data, but also have the basis for concluding that the original claim of a weak-to-non-existent relationship between HSPI and absolute income level was a mischaracterization.

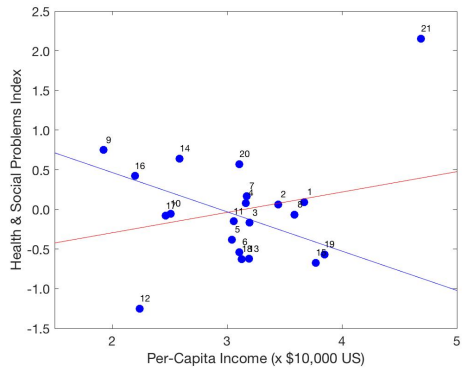


Figure 2. Index of Health and Social Problems vs. Income levels for 21 OECD countries. Red line: best–fit straight line using all data points ($r2 \approx 0.055$). Blue line: best-fit straight line omitting outliers 12 (Japan) and 21 (U.S.), giving $r2 \approx 0.343$.

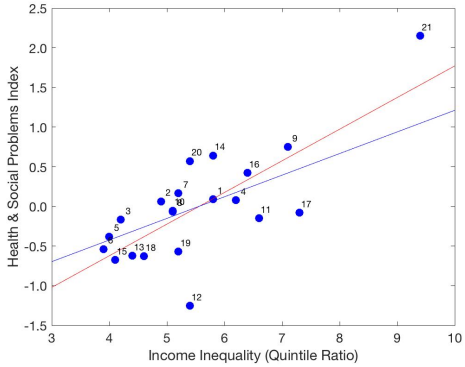


Figure 3. HSPI vs. Income Inequality (quintile ratios) for 21 OECD countries. Red line: best-fit straight line using all points ($r^2 \approx 0.57$). Blue line: best-fit straight line omitting the outlier 21 (U.S.), giving $r2 \approx 0.408$.

Moreover, we can see from *Figure 3* that the relationship between HSPI and income inequality is only about on par with the relationship documented between HSPI and absolute income. Especially when we remove the obvious outlier pt 21 (again the U.S.), the best-fit straight-line relationship has an $r^2 \approx 0.41$, not much different from the r^2 for the previous relationship.

It is certainly interesting and important, of course, to be able to ferret out such a relationship — this does indeed suggest that income inequality could have important impacts on societal health and well-being — but this initial replication attempt also suggests that the impact of inequality may be very similar to the impact of absolute income level itself. And of course, with two variables seen to have similar impacts, there could also very well be some correlated 3rd-variable problem here, perhaps something as straightforward as health care, which would certainly be reflected in our society by income level, and could also be a source of demoralizing experiences for those of low SES vs. high SES.

At the U.S. state level, we find similar results, but with an even weaker relationship evident for HSPI vs. income inequality, as shown in *Figure 4* and *Figure 5*. The interesting outlier pt #9, by the way, is the District of Columbia, and without that outlier, the relationship of HSPI to income level is even stronger than the shown $r^2 \approx 36$, and the relationship of HSPI to inequality is even weaker than the shown $r^2 \approx 0.125$. Again we find good reason to be skeptical of Wilkinson & Pickett's intuitively appealing, but weakly-supported, explanatory narrative involving the social ills of income inequality.

Principal Component Analysis (PCA), Dimension-Reduction, and Clustering Analysis via PAM

After cleaning, reorganizing, and then reducing the 2018 County Health Rankings & Roadmaps dataset (see http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation),

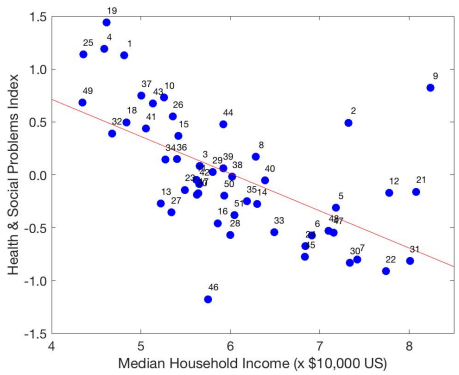


Figure 4. HSPI vs. median household income for the 50 U.S. States and District of Columbia (numbered in alphabetical order). Best-fit straight line shown in red, with $r^2 \approx 0.36$.

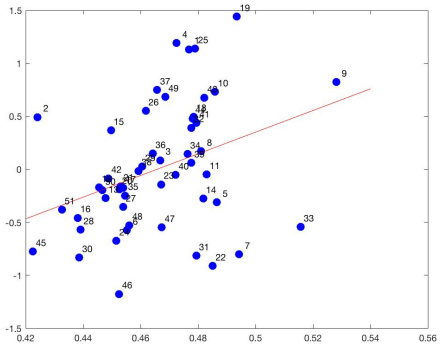


Figure 5. HSPI vs. income inequality (measured by the Gini Index) for the 50 U.S. states and District of Columbia. Best-fit line shown in red with $r^2 \approx 0.125$.

the resulting 20-dimensional set with over 3,000 county data elements was subject to further cleaning to eliminate missing data cells, and then subject to a principal component analysis (PCA) and reduced to a 3D data set with 2,794 county entries (*i.e.* a 2794×3 data matrix). The resulting 3D dataset is illustrated below in *Figure 6*.

It's a little difficult to see much in the 2D image of the 3D data, but within MATLAB you can manually spin it around a bit to give it a more 3D feel and there appears to be some interesting internal structure in the 3D data — certainly enough to entice us to practice some clustering analysis.

One difficult thing about cluster analysis, of course, is common need to decide how many clusters you should work on. I looked at $k = 2$, 3, and 4 for this report, all of which yielded moderately interesting results. Here I report on the $k = 4$ results (which can be viewed in the MATLAB file `CS521Project_CountyData.mlx`.

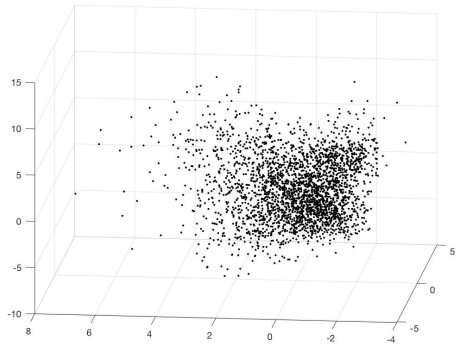


Figure 6. The 3D county-level health dataset resulting from PCA reduction. Some interesting internal structure is suggested.

The PAM approach was time-consuming, of course, and for this dataset did not settle quickly/easily into a definitive non-changing set of medoids even after 10 (rather long) full iterations. But by 10 iterations, it was clear that the medoids were changing very little, finally producing results like that shown below in *Figure 7*.

As was the case with the 3D PCA-resulting data itself, the 2D image doesn't really do justice to the subtle 3D structure in the dataset, but we see even with this relative brief analytic effort a plausible clustering scenario with a central relatively dense cluster (in red), and three peripheral clusters sprouting outward and taking partitioning the outer "shell" of points in

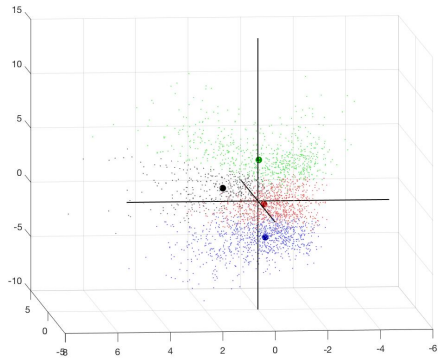


Figure 7. Result of a PAM-based clustering using $k = 4$ on the 3D PCA-reduced county health data. Medoids shown as larger colored points. The red medoid was the first one added to the medoid list.

three distinct directions from the center.

Further Discussion

Ultimately we have a small handful of interesting results, plus some excellent practice

with some mining techniques and non-glamorous data-mining activity of real-world data aggregation, cleaning, and reorganizing.

We saw that Wilkinson & Pickett's original data-agglomeration approach could be replicated, with considerable effort, but that the replication result is quite mixed: instead of the repeated narrative espoused over the past few years about the critical importance of income inequality and its effects on societal health and well-being, we see a much more nuanced result, with (at best) roughly equally weighted effects of income level and income inequality on a composite measure of life quality.

Using further data aggregation, and PCA and PAM analyses, we also made an interesting start on examining the internal structure of a complex high-dimensional dataset representing health-related societal and lifestyle characteristics at the county level for the United States.

Further useful work in this last regard would most likely entail: considering 2D and 4D PCA reduction and considering cluster analyses that could better accommodate non-convex datasets.

Selected Bibliography & List of Datasets/Locations

Behavioral Risk Factor Surveillance System (used for data on prevalence of mental health issues): https://www.cdc.gov/mmwr/volumes/67/ss/ss6709a1.htm

FBI crime database (used for data on violent crime): https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-5

European Values Study (used for data on trust/mistrust): https://europeanvaluesstudy.eu/

National Vital Statistics System (for info on teen birth rates): https://www.cdc.gov/nchs/nvss/index.htm and https://www.cdc.gov/nchs/pressroom/sosmap/teen-births/teenbirths.htm

Organization for Economic Development and Cooperation: Program for International Student Assessment (PISA) (used for data on educational attainment): http://pisadataexplorer.oecd.org/ide/idepisa/ to obtain 2015 results for Maths and Reading

Porter, E. (3/25/2014a). Income equality: A search for consequences. *The New York Times* online, 3/26/2014. Accessed 10/14/2018 at https://www.nytimes.com/2014/03/26/business/economy/making-sense-of-income-inequality.html

Porter, E. (3/25/2014b). Q&A: A sociologist on inequality. *The New York Times* online, 3/25/2014. Accessed 10/14/2018 at https://economix.blogs.nytimes.com/2014/03/25/qa-a-sociologist-on-inequality/

The Equality Trust. (n.d.) Web site last accessed 10/14/2018 at https://www.equalitytrust.org.uk/

The General Social Survey (GSS) (n.d.) Web site accessed 10/14/2018 at http://gss.norc.org/

United Nations (used for data on violent crime): https://dataunodc.un.org/crime/intentional-homicide-victims

United Nations Human Development Report (2017) (used for data on life expectancy at birth): http://www.hdr.undp.org/en/data.

US Census Bureau (for data on life expectancy): https://www.census.gov/library/publications/2011/compendia/statab/131ed/births-deaths-marriages-divorces.html

US National Center for Health Statistics (part of the CDC, used for data on infant mortality): https://www.cdc.gov/nchs/nvss/linked-birth.htm

Wilkinson, R. & Pickett, K. (2009). *The spirit level: Why greater equality makes societies stronger*. New York: Bloomsbury Press.

Wilkinson, R. & Pickett, K. (2017). The science is in: Greater equality makes societies healthier and richer. *Evonomics* (online). Accessed 10/14/2018 at http://evonomics.com/wilkinson-pickett-income-inequality-fix-economy.

World Bank (for data on infant mortality rates): https://data.worldbank.org/indicator/SH.DYN.NMRT?view=chart) (and for data on teen birth rates: https://data.worldbank.org/indicator/SP.ADO.TFRT)

WHO (for data on prevalence of mental health issues): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3039289/

World Obesity Federation (previously known as the International Obesity Task Force or IOTF) (used for information on obesity rates): https://www.worldobesitydata.org/presentation-graphics/resources/tables/

World Values Survey (used for data on trust/mistrust): http://www.worldvaluessurvey.org