Measuring Wealth in Emerging Economies: Assessing the Reliability of Wealth Indices in Malaysia

Emir Izat Abdul Rashid[1*]

[1*]Center for Technology, Strategy, and Sustainability, Asia School of Business, Jalan Dato Onn, Kuala Lumpur, 50480, Kuala Lumpur, Malaysia.

Corresponding author(s). E-mail(s): emir.izat@asb.edu.my;

Abstract

Measuring wealth and documenting its distributional impacts have become a first-order concern in the social sciences. Most studies on wealth took place in highly-developed countries due to reliable data sources from tax records; there is not much we can say about wealth ownership in countries where these sources are not available. Much of the work on micro-level wealth measures in emerging economies are increasingly reliant on remote-sensing data such as nightlight or machine-learning tools to unveil spatial wealth distribution. To assess the reliability of these wealth measurement tools, I used property transaction data in Malaysia. The reason is because housing assets are the largest source of wealth for households in the country. Thus, house prices can serve as an informative source of spatial wealth distribution akin to wealth indices. I found that the Meta Relative Wealth Index is correlated with housing prices although this relationship is weak. Further, they do not predict house prices as well in comparison to median income data from official statistics with higher aggregation. This suggests that the Meta Relative Wealth Index can provide some descriptive information, but cannot replace the precision provided by official statistics – even the ones that are highly aggregated.

Keywords: wealth measures, inequality, economic methodology, big data

1 Introduction

The growth of research in wealth ownership and its effects on well-being and social stratification has grown immensely in various social science fields[1][2][3][4]. For many developed countries with high data capacities, income surveys and tax records were actively utilized to measure wealth[5][6][7]. By leveraging on other types of well-being surveys such as health, education, wages and labor, mental health, and consumption, matching wealth and other well-being measures have become second-nature to many social scientists[8][9]. Indeed, policymakers and academics in the developed world who wish to improve the well-being of others are actively interested in wealth because of its ability to transform itself into purchasing power and how to replicate that ability for those who are not so wealthy – access to better medical care, better foods, better education, and better standards of living[10][11]. However, the same cannot be said for developing countries which often have low data capacities. We do not know much about wealth in developing countries and a large reason for this intellectual gap is due to the availability of reliable and replicable dataset on wealth in developing countries.

To potentially fix the data availability problem in developing countries, many studies have attempted to create data based on typical physical characteristics of wealth ownership such as taller buildings, night-lights, and better roads. One strand of this data creation process is by utilizing remote-sensing data[12][13][14][15]. Many papers took advantage of characteristics such as luminosity during night-time[16][17][18][19], built environment, and land use[20]. Remote-sensing data based on physical characteristics has allowed estimation of micro-level wealth ownership with relatively high anonymization. Since data creation is done without identification of individuals who own the wealth, those who use these tools to target aid or study spatial distribution of wealth may use these datasets without concerns of data security. However, one main problem that arises from these tools is reverse causality. Because amenities, physical buildings, and lights are used as data inputs to predict wealth, those who study wealth effects on well-being cannot use these tools. This is because well-beings are also predicated on the access of the amenities that became inputs to remote-sensing generated wealth data. Thus, physical characteristics predicts wealth and well-being but wealth also predict well-being. As a result, we cannot disentangle whether wealth is the variable that causes well-being improvements or whether these physical characteristics drive well-being.

To alleviate this issue, another strand of data creation process focuses on using machine learning algorithm to create wealth indices[21][22][23][24]. Wealth index is an output of a wealth measure with inputs that take in many non-survey parameters such as remote-sensing data and mobile phone usage calibrated by ground-truth survey such as the Demographic Health Survey (DHS); the most notable development of this method was the Meta Relative Wealth Index (MRWI)[25]. The creation of the MRWI has enabled micro-level estimation of spatial wealth to 135 countries. However, one main criticism of the MRWI is the difficulty in interpreting the wealth variable that they produce i.e. machine learning algorithms tend to produce predictions without understanding the underlying mechanism that lead to the current spatial wealth distribution. Another criticism is that the calibration exercise that utilized DHS only includes 56 countries which requires extrapolation for other countries. Although they acknowledge the variation in the margin of error in-between countries, no study thus far has since attempted to corroborate the wealth predictions produced by the MRWI on real-world data on wealth ownership.

To corroborate the predictions yielded by remote-sensing data and machine-learning generated wealth indices, I used the property transaction data in Malaysia as a wealth measure to show the corroboration between the MRWI and an implication of wealth, namely house prices. Why do I use property transaction data as a measure of wealth in a neighborhood? This is because household assets make up the largest portion of wealth ownership in Malaysia – 96% of all wealth owned by a household is tied to real-estate property instead of stocks, bonds, or savings[26]. Thus, the underlying data that generates the wealth variable is a tangible asset with a price which captures ownership of an individual or household of an asset[27]. The property transaction data is also highly granular, with locations that captures a whole neighborhood smaller than a pixel in remote-sensing data or wealth index. A caveat of this data exercise if it is replicated elsewhere would require the same level of data capacity as Malaysia.

I found that the MRWI and house prices are correlated. I included a rich set of controls for the robustness of the correlation and the correlation estimates still remain the same. To see further statistical power of the MRWI, I trained a gradient-boosting model to test for the capacity of the data to predict house prices. With the same covariates and the same training algorithm, MRWI performed poorly in comparison to official census data that have higher aggregation to the MRWI. This suggests that the MRWI may have been accurate in determining different house prices but not precise to determine the price better than aggregated official statistics data.

The following sections are as such: in section 2, I will be detailing the theory and literature for the study of wealth in developing countries; in section 3, I will be exploring the datasets to be analyzed; in section 4, I will be detailing the empirical strategy for the macro-level predictive performance of the MRWI; in section 5, I will be presenting the results for the performance of MRWI, in section 6; I will be delving into the error measurements of the MRWI to find potential mechanisms that could explain the performance of the MRWI; in section 7, I will explain the potential policy implication and the conclusion of the paper.

2 Theory and Literature

2.1 Wealth in Developing Countries

What is wealth in the developing context? Wealth as defined in the DHS is based on asset ownership, housing material, and water access – wealth is then transformed into an index from 1 to 5[28]. However, wealth is more suitably defined in this paper as net worth i.e. the total amount of assets owned minus the total amount of debt[29]. The reason why this definition is problematic in many developing context is because assets are poorly defined. Financial assets in developing countries such as stocks and bonds may not have been developed to the same level of complexity as their developed counterparts.

Another class of assets that are common in households is savings. Again, this is difficult to define in developing countries especially in a largely agrarian context;

although we know that savings in the developed world is best understood by the amount of money you have in the bank, most of the world's poor store their financial resource in ways that are hard to compute methodologically. For example, farmers who typically do not own bank accounts may store their cash in their own homes or in agricultural surplus they produce. Computing savings in developing countries remain a methodological challenge for development researchers.

One other common source of wealth in developing countries is in real estate or property. The feasibility of computing wealth in property is limited by the legal capacities of the state in enforcing private property rights. As one can tell, accounting property assets in developing countries with weak legal capacities to enforce property rights such as conflict areas is problematic. But in most developing countries who are low-middle to upper-middle income economies with no active conflicts tend to have at least sufficient property protections from the state. In these countries, accounting property wealth usually entails the estimation of property value. Although property value can only be determined with certainty upon the sale of the property, one can estimate property value through the transaction price of a property in a comparable location, size, and craftsmanship of the building[30]. We will be using this assumption to estimate how wealthy a neighborhood is.

A related but distinct concept to wealth that is easily conflated by many people is income. The reason why is because many assume that higher income individuals should be wealthy as well hence should have higher levels of well-being. In many cases, this is true. Specifically, high-income individuals can purchase a greater share of assets than low-income individuals and it is easier for them to reduce their liability translating these economic facts to well-being. However, the statement of higher income implies higher wealth can only be true if the high-income individual decides to purchase wealth instead of consumption. In developing countries, especially in weaker states where the government cannot protect private property, high-income individuals may not choose to be wealthy to avoid certain undesirable outcomes such as getting their wealth taken indiscriminately thus choosing to avoid wealth accumulation. On another note, wealthy individuals also may not be the ones with high-income. In many developing countries, wealthy individuals tend to be those that are politically-linked or those that inherit wealth. Although the link between income and wealth can be established in developed countries, this link is more ambiguous in the developing world. As a result, there is still a gap in the literature on the relationship between wealth, income, and well-being in the developing world.

In Malaysia, the source of wealth for most households is in property assets. Some estimates suggest that 96% of wealth in the country is derived from property ownership[31]. Another reason why property assets is a valuable indicator of wealth is the relatively large informal labor market. The most recent estimates suggest over 20% of the labor force is employed outside of the formal sector[32]. Some estimates of other administrative data from the Employee Provident Fund (EPF) and Social Security seemed to suggest that 40% of workers are not covered[33]. The result of this large chunk of informal sector in the labor market contributes to the lack of reliable information on individual income, savings, and financial assets. Housing market does not depend on the formal employment of the buyer. The reason is because one can receive and apply for housing loans with alternative documentation to prove income status. After all, it would be a bad policy to deny housing access to 40% of the population. Thus, with the difficulty in estimating wealth of a neighborhood via financial assets and savings, I chose to use house prices as a measure of wealth.

As one can notice in the map in figure 1 and 2, house prices and income centers around the Greater Kuala Lumpur area. Income and house price is highly correlated – where people have higher income, house prices also is higher. One can clearly see the concentration of wealth in the city.

2.2 Contribution to Literature

The paper adds to the literature in two main aspects: by testing the reliability of the dataset and through adding a potential source of data to study wealth in Malaysia. As far as this study is concerned, no papers except for Sartirano et. al[34] have engaged with the use of the MRWI critically where they test the reliability of the dataset on Indonesia – one of the countries covered under the DHS dataset. In that paper, they estimated that if they have used the MRWI as a source of information to redistribute developmental aid as opposed to the government-issued Social Protection Card (KPS), the exclusion error would be 32.82%. This paper further adds towards critically advocating for a more responsible use of alternative datasets to government-sponsored ones in policy formulation. Another contribution that the paper seeks to add is by introducing another microdata source to measure wealth in absence of official statistics of individual wealth and in presence of large informal labor markets i.e. house prices. Although housing and wealth may not be perfect reflections of one another, we know that high income households tend to agglomerate and select the neighborhoods they choose to stay which drives up the prices in more premium neighborhoods[35].

3 Data

3.1 Meta Relative Wealth Index

Another method of estimating micro-level wealth levels is through utilizing machine-learning models that takes in parameters such as mobile phone use[23], geography, and remote sensing data to predict wealth in a sub-national area[36][37]. The most recent development of this method is the Meta Relative Wealth Index (MRWI) which combines the input data of the machine learning model to output the wealth index of the DHS survey[25]. As a result, the dataset has provided a predicted wealth data on a sub-national level of 135 low to middle-income countries. Summary statistics will be located in section 4.3

The dataset is consisted of the relative wealth index, the coordinate of the tiles, and the expected error of measurement in each tile. I then used QGIS to derive the tiles where I match each township to a tile. In this way, I am able to derive the wealth index measurement for each township which I then use for the subsequent analysis.

The main issue with the MRWI is that it is trained on the wealth index for the DHS data which is only consisted of 56 countries. Additionally, the 1 to 5 index rating provided by the DHS may lose a lot of information on the wealth estimate of a unit of

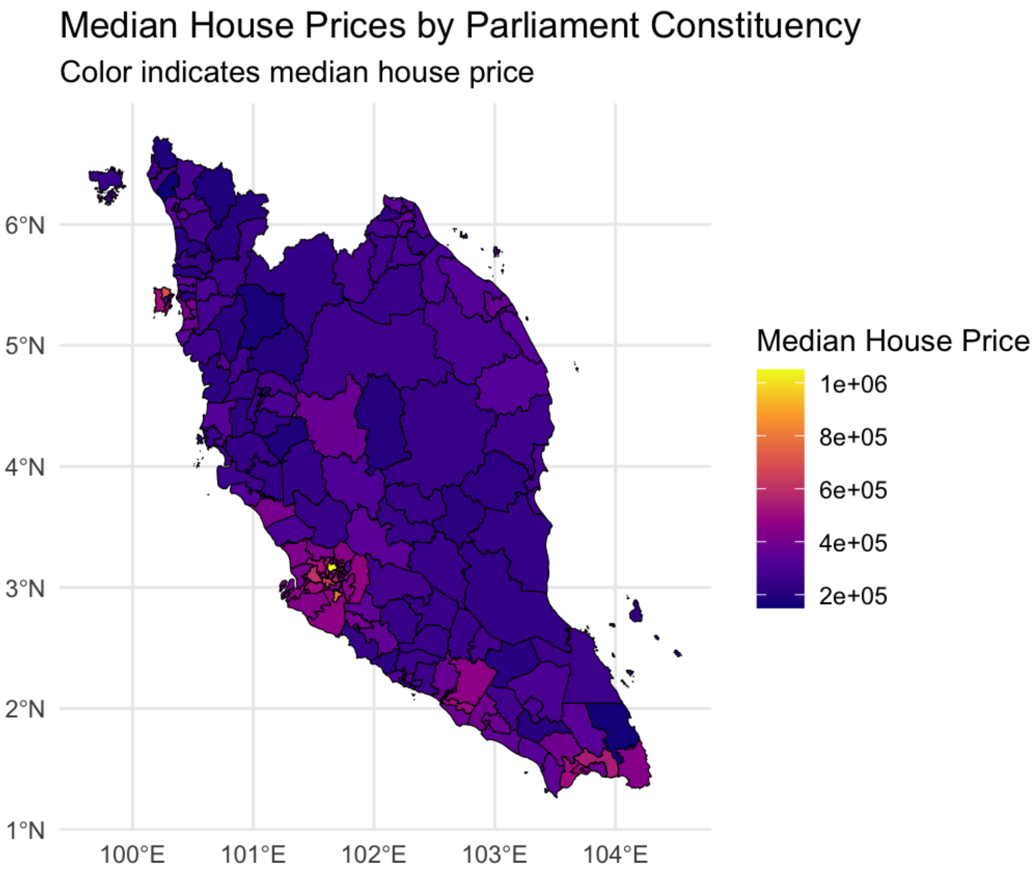

Fig. 1 House Price Distribution in Peninsula Malaysia

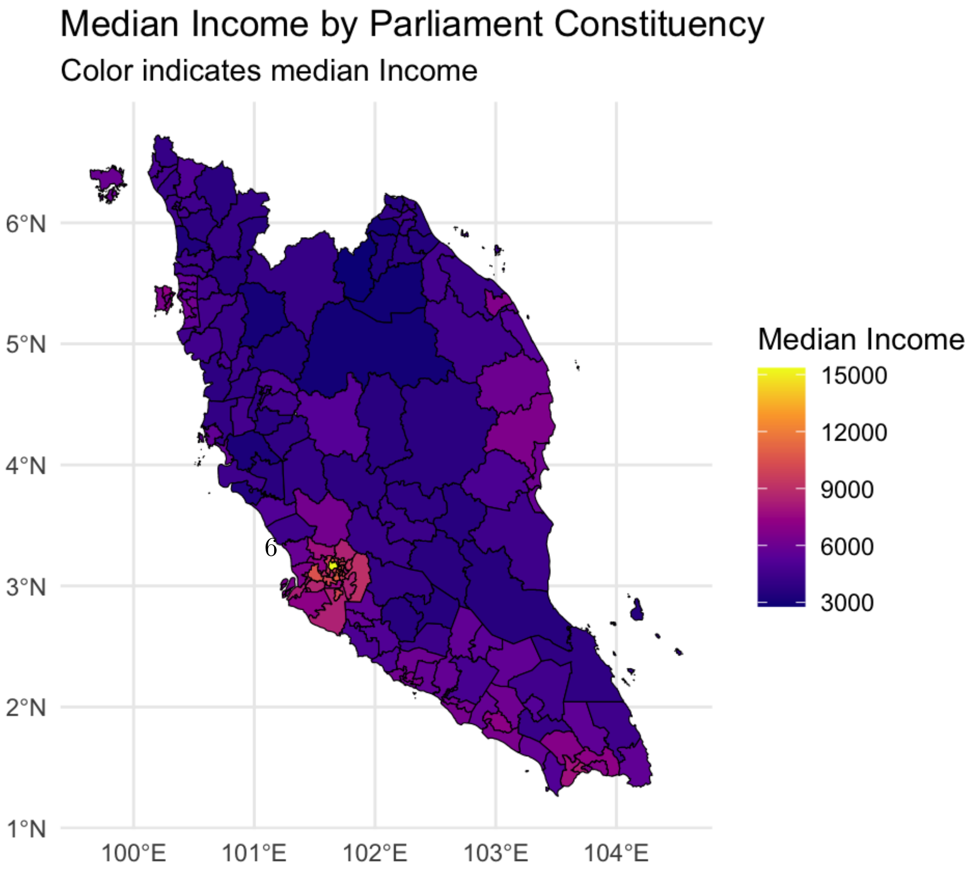

Fig. 2 Income Distribution in Peninsula Malaysia

observation. The usage of the index rating may be calibrated depending on the sample countries but not countries not covered by the survey. The authors address these by including error measures. Thus, Malaysia which is not covered under the DHS survey may not be informative as Indonesia which is covered.

Another issue with the MRWI is the lack of transparency of the wealth prediction[22]. The most common argument against machine learning models is the lack of transparency in the parameters allocated by the model. The models may have computed the parameter values given the information it received but does not understand the mechanism behind why they assign such a value. The reason why it is problematic is because certain parameter estimate may have been assigned due to institutional biases. For example, low mobile phone use is predicted of having low wealth in an area but this could be due to the denial of governments to address infrastructural problems or simply the government has no funds to pay for these projects. The machine predicts that the area has low wealth but could not determine whether an area is intentionally underserved or structurally underserved which will impact what policies are appropriate to address them.

The data generation process includes training the model with other countries. This means that in order to attain wealth measures of Malaysia, it involves conflating other countries into the algorithm which may not be the most accurate characterization of each country's wealth distribution. The reason why this may not be the best method in generating the wealth indices is because each country has a different explanation as to why wealth is distributed or arranged in a certain way[38]. There has been a wide tradition in economics and political science on theories to explain a country's institutional development. And to predict data on a country X with data from other countries may not capture the biases that explain why wealth is distributed in a certain manner. When we do not understand how wealth is distributed, certain data generation procedures that are independent of these variables may result in inaccuracies i.e. MRWI is not country-sensitive.

MRWI also may not account for the replicability of the indices. Suppose that a policymaker from countries covered in the MRWI decided to address all the parameters use in the algorithm and provide them to maximize wealth indices whilst not improving other measures. What happens is that the MRWI predicts that the area has a lot of wealth according to the index, but well-being has not improved. Although a priori it is unclear that machine-learning generated wealth indices provide some information on wealth ownership of developing countries, to show that they are correlated to house prices would be a validation to their model.

3.2 Property Transaction Data

The property transaction data is derived from the National Property Information Center (NAPIC) which is publicly available[39]. The data covers all transaction data from residential, industrial, and commercial property between January 2021 – September 2023. For the purposes of the study, we are only interested in attaining the transaction data for residential property. The total number of property transaction that covers between that period is 220,981 transactions which totals at $RM106,893,647,000$

or around $22.4 billion; it covers 25,884 distinct townships across the country. The summary statistics will be located in section 4.3

The raw data is consisted of property prices, township, district, tenure (either lease hold or free hold), land and house area, main floor area, and type of house. I then used ArcGIS to geocode the 25,884 townships to derive the coordinates for further analysis.

There are several reasons why house prices may not be the most accurate measure of wealth. First, houses are not fungible. If a family received a negative shock to their income, houses are not easily liquidated: which means that a family who lose income i.e. not wealthy could be living in a high-priced area. To address this issue, I will be aggregating to a median house price of a specific-type on the pixel-level definition of the datasets. What this allows me to do is to take a median house of for example a one-storey terrace at the level of a neighborhood. Instead of looking at the individual houses which could suffer from the fungibility issue, the outcome is not individual wealth but *neighborhood* wealth.

Second, house prices may be determined by external factors beyond the wealth of a household e.g., asset bubbles. A high-priced area could have been driven by speculation and investment choices. Thus, a wealthy neighborhood could have been wealthy not because wealthy families live there but because wealthy people invests there i.e. non-wealthy people could live and rent but the underlying house price is high. To address this, I will be controlling for the types of houses as a way to remove the effect for the types of houses more prone to speculation. For example, single-family homes are less likely to be investment vehicles, but apartments typically are[40].

Lastly, houses incur debt which means one could have a lot of high-priced asset but at the same time have a lot of high-priced debt i.e. net worth is 0. However, a main counter-point to this argument is that large housing loans would need healthy credit and stable income. Thus, there is a selection mechanism where the ones purchasing higher-priced housing are also the ones who could pay back the loans i.e. have the prerequisite of attaining a mortgage such as high projected income growth. High-priced neighborhoods can also reflect the demographics of people who buy them i.e. either they could afford it to pay cash (wealthy) or they project future high-income growth (wealthy-anticipated).

As a result, house transaction prices is a way for me to estimate neighborhood-level wealth as opposed to individual wealth similar to the wealth index dataset. Why would neighborhood-level wealth matter more than individual wealth? This is because amenities and markets are built on aggregate demand, not individual i.e. a shop, a healthcare service, or a school will not be constructed for a very wealthy individual on a non-wealthy neighborhood. The first-order concern for many public policy experts and social scientists is when we have agglomeration of wealth in neighborhoods which lead to a convergence of amenities and markets serving these areas whilst removing access from others. One main mechanism of how wealth affects well-being is access to high-quality amenities and living standards. Many economists have documented neighborhood effects on social mobility[41] including running successful randomized controlled trial that have improved children's outcome when they move to a high-mobility neighborhood[42].

4 Measuring Data Correspondence and Predictive Power

4.1 Data Correspondence

Data correspondence is the measurement in how well the correlation is between the datasets we pick and the house price. The empirical strategy needed to study data correspondence must take into account the type of houses that is being studied. The reason is because high-rise apartments may have a different pricing mechanism than a single family home because of its ability to serve as an investment vehicle. As argued before, prices of single family homes should reflect wealth levels better than prices of high-rise apartments because a family is more likely to stay in them. Thus, we will separate the measurements for different types of houses. The model we will use to measure the correlation between wealth index on house prices is a simple OLS model. Although the OLS may not reflect the underlying functional form of the wealth datasets we are utilizing, it is used for ease of interpretation. The model is as follows, for a tile i for a house j for the wealth index:

$$ln(HousePrice_{ij}) = \beta_0 + \beta_1 WealthIndex_i + \epsilon_i \tag{1}$$

Tile i is a unit of observation of a pixel in the dataset we are interested in. For example, the MRWI has a 2.4km resolution. I will also include a rich set of controls to ensure robustness of the estimate (which includes housing characteristics such as tenure, housing size). $HousePrice_{ij}$ is a house j located in tile i.

4.2 Predictive Ability

Simply showing correlation between two variables is insufficient to convince that the MRWI is a robust tool to predict wealth data. Another source of interest that some may be interested in is the predictive power of the relative wealth index in determining house prices. If there is limited correspondence between house prices and wealth index in the OLS, one could argue that these tools could at least predict house prices as a way to find where wealthy neighborhoods may be.

To measure the performance of the index in predicting house prices, I will be utilizing a boosting algorithm LightGBM[43] that takes in a defined set of covariates such as property type, month of transaction, land tenure, and area size. Then, I will split the house price datasets into a training and test sets. I will train the housing data by using the MRWI as an input. To benchmark the performance of the MRWI, I will be using the median income from the 2020 census as an input with the same covariates as the wealth indices trained on the same algorithm. If the error rates for the wealth indices is more or less similar to the 2020 census, we know that wealth indices couldn't do any better than official statistics. Also note that the census data is less granular than the MRWI. This is due to the fact that the census data is on the level of parliamentary constituency which is the most granular boundary level that is publicly and readily available at present.

4.3 Summary Statistics

The summary statistics for the MRWI and house prices is as follows:

Data	Median	Mean	Max	Min	*N*
House Prices	360,000	483,723.20	30,000,000	5,000	220,981
Relative Wealth Index	-0.045	0	1.757	-1.404	18,147

Table 1 Summary Statistics of House Prices and Wealth Index

Already in our summary statistics it is clear that there is higher variation in house prices than they are in the relative wealth index. This could cause some problems in our OLS estimates as this level of variation asymmetry could cause our findings to be significant given the number of observations but the effect would be really small as the distinction between wealthy and poor neighborhoods by the wealth index is more or less similar.

5 Results

5.1 Data Correspondence to Wealth Index

The result of the simple OLS is as follows:

2-3	*Dependent variable:* $ln(HousePrice)$	
	(1)	(2)
Meta Relative Wealth Index	0.289*** (0.004)	0.243*** (0.004)
Observations	220,979	221,041
Controls	✓	

Note: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Controls include district, month of transaction, tenure type (leasehold or freehold), and area of the house. There exist duplicate observations due to one house being in the boundary between two tiles. Removing these observations do not change the estimate nor the inference.

Table 2 Correlation between Meta Relative Wealth Index and House Price

As one can observed, the MRWI is correlated with house price. However, simply showing the correlation between two variables is insufficient. The coefficient is measured to be 0.268 which means that a one unit increase in the wealth index corresponds to a 0.268 increase in log price units which roughly translates to RM1.30. This increase is very small considering that the range of MRWI is between -1.4 to 1.76. The fact that the increase is relatively small indicate that the lack of variation in the MRWI cannot account for the variation in house prices. We have highlighted a big weakness in the MRWI – the lack of variation in the index measures led to the loss of information to distinguish between property prices in wealthy and poor neighborhoods.

5.2 Which is Better at Predicting House Prices?

To measure whether the MRWI can predict where expensive houses may be, we will be using the 2020 census for income measurement aggregated at the level of parliamentary constituency. This is the most granular geographical level that is publicly available which includes the Federal Territories of Kuala Lumpur, Putrajaya, and Labuan.

We will be measuring performance of each dataset on the root mean square error (RMSE), the mean absolute error (MAE), and the R^2 of the validation set. A better dataset should have a lower RMSE and MAE, combined with a higher R^2.

Measure	MRWI	Median Income
RMSE	0.501	0.384
MAE	0.383	0.283
R^2	0.505	0.708

Table 3 Performance of MRWI Benchmarked with Official Statistics

As one can observed, the MRWI performed worse than an aggregated data of median income for official statistics in all measures. This shows that although the results of the simple OLS showed that MRWI and house prices are correlated, the official statistics (despite being aggregated) is better at predicting house prices.

6 Sources of Data Deviation

To investigate where MRWI may have not performed at its best, I will compare the errors generated by the RWI measurement and the errors generated by the correlation between house prices and the MRWI. The reason for including this section of this analysis is to see where the source of errors come from – namely whether errors of the estimates are driven by the error during the MRWI data creation process or by the biased relationship between wealth and house price. This has potential implications for the discussion surrounding the creation of indices generally because if the estimated errors of the data creation process are shown to correctly correspond to the OLS errors we have produced then we know that wealth and house prices are correlated and the errors in MRWI measurements are to be attributed to data deviation. If there is a biased relationship between wealth measures and house price, then the errors should be equally distributed across all the house price distribution.

To test for the two types of errors mentioned above, we have to assume that one is true; if we assume that house prices perfectly reflects wealth, then we should expect that the errors generated by the OLS estimates to be correlated to the expected errors generated during the data production process and if we assume that the wealth index perfectly reflects the wealth, then we should expect that data deviates equally across all house prices to establish that the relationship between house prices and wealth index is unbiased.

6.1 Are MRWI Errors Correlated with OLS Residuals?

To provide researchers who utilize the MRWI dataset with enough information with regards to expected performance of the MRWI in predicting wealth, expected errors were generated for each tiles of the MRWI observation[1]. The question is whether the OLS model in section 3 produce similar error measurements than the expected errors produce by the author? The reason why this matters is because if there is strong correlation between the errors generated by our OLS estimates and the predicted errors generated by the authors, then we can attribute the deviation of the wealth index from the house prices on the predicted errors of the MRWI given that house prices perfectly reflects wealth. If there is no correlation between the predicted errors given by the authors and the errors generated by the OLS model, then we know there is some form of inaccuracies in the production of the wealth index.

	Dependent variable:
2-2	OLS Residuals
Predicted MRWI Error	-0.830*** (0.036)
Observations	221,001

Note: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table 4 Correlation between Residuals and Predicted Errors from MRWI Measurements

As observed here, the tiles where the authors expect to find the errors are not where the areas in which there is greatest OLS residuals between house prices and wealth index. To illustrate this in a graphical form, below is the comparison between the tiles where the authors expected to find the largest errors and where the OLS residuals are:

As observed in figure 3, the predicted error is well distributed along the wealth indices i.e. where we expect to find errors should be the same across neighborhoods of varying wealth levels. As observed in figure 4, the OLS residuals indicate that MRWI produces more errors the higher we go in wealth index. At lower levels of the wealth index, the residuals are more bounded than at higher levels of the wealth index. This indicates that if we assume that the house prices reflect the true value of wealth in a country, then the wealth index underestimate wealth i.e. as neighborhoods become more wealthy, house price costs more than their predicted value.

[1] The authors did this to ensure that users of the data know the uncertainties in their data production process for each tile.

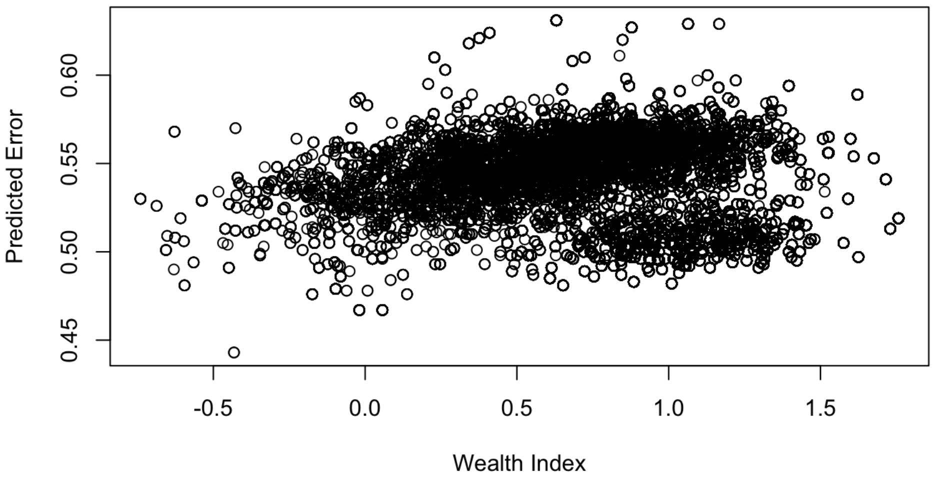

Fig. 3 Predicted Error

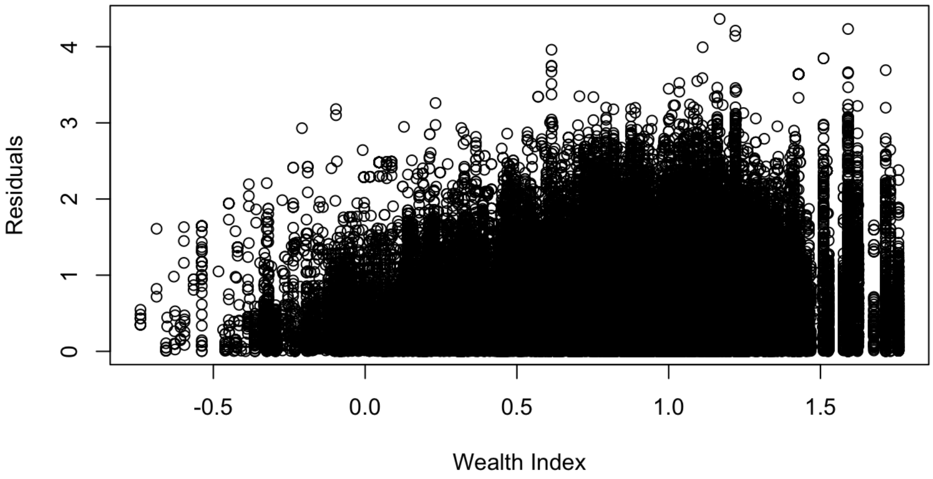

Fig. 4 OLS Residuals

6.2 Does Wealth Predict House Prices?

Another issue that may arise from our analysis is the ability of wealth to predict house prices given underlying biases, assuming MRWI perfectly observes wealth. The reason why this matters is because the observation of housing transaction is endogenous. Our dataset only displays *observed* housing transactions. And observed housing transactions may feature certain underlying features that can drive biases. For example, a prospective house seller may have not sold their homes below the price they buy thus house prices will typically biased upwards. If the MRWI performs poorly it could be that house prices are a reflection of the underlying market conditions. To test this underlying biases that house prices can exhibit unobserved by the wealth index, I use a Q-Q plot to show the violation of normality between wealth index and house prices. I also plotted the residuals against house price to see which prices produce the largest errors:

In figure 5, I show here in the Q-Q plot that at low-levels of the wealth index, houses are significantly cheaper than what should be predicted by the normal distribution. At high-levels of wealth index, houses are significantly more expensive than what should be predicted by the normal distribution. This indicates that assuming that MRWI perfectly predicts wealth, houses in wealthy neighborhoods are overpriced and in poor neighborhoods are underpriced. In figure 6, I showed that at lower house prices, the MRWI performs poorly in overpredicting the price whereas at higher prices MRWI underpredicts the house prices. The MRWI performs well for houses in the middle range. This indicates some form of underlying biases induced by the property market which explains the reason why MRWI do not perform as well in its predictive inference. Assuming that the wealth index perfectly reflects wealth in the country, house prices do not translate to wealth measures at the top and the bottom part of the house price distribution i.e. there is a heavy-tail skew in house prices. Thus, academics who would like to use this dataset should be aware of the potential skewness that the MRWI may have on other wealth measures.

6.3 Correcting Extreme Value Deviation

In order to correct for the extreme value deviation, I will now conduct a weighted regression where I assign weight equals to 0.2 for the bottom and top quantile, and assign weight equals to 1 for the second and third quantile[2]. What I did here is to put more weight on observations where the model performs well and less weight on observations where the model performs poorly. The interpretation of these weights are the magnitude in which there is an unobserved variable driving the house prices in extreme values. Of course, this is not the best method to deal with flawed data but this exercise is to illustrate how we can correct for the violation of normality in the OLS.

As observed in figure 7, the QQ plot indicates that the regression is no longer heavily-skewed. Not only that, the model generated less extreme value errors which means that this model has less deviations from their prediction. The estimate also decrease in its magnitude. The magnitude of 0.243 in the original regression now falls to a mere 0.106. The model has become less biased in its expense of explanatory power.

[2] The weights that I assign is arbitrary but one can run an optimized model to obtain the best value.

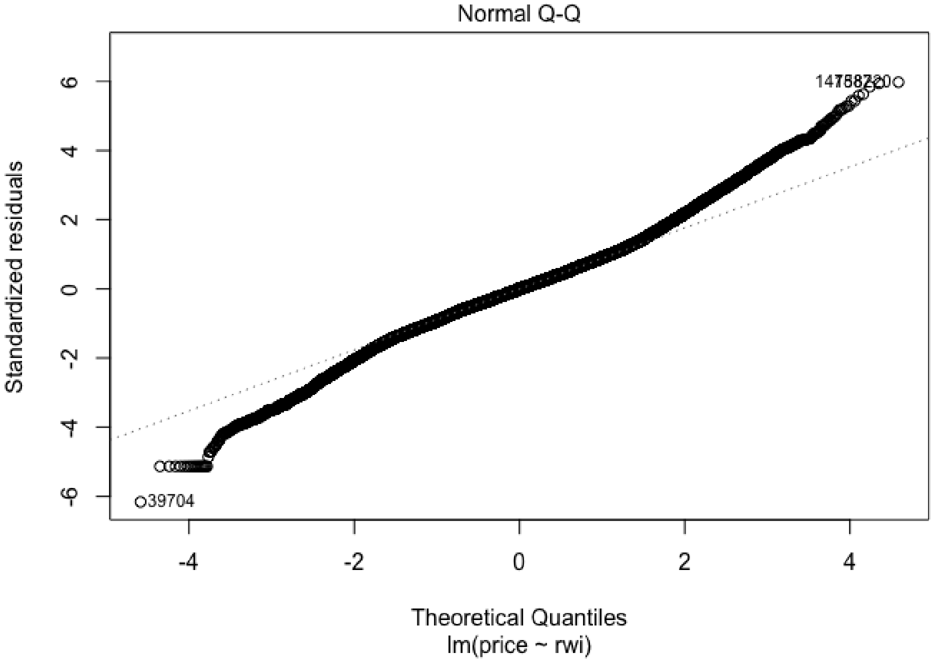

Fig. 5 QQ Plot

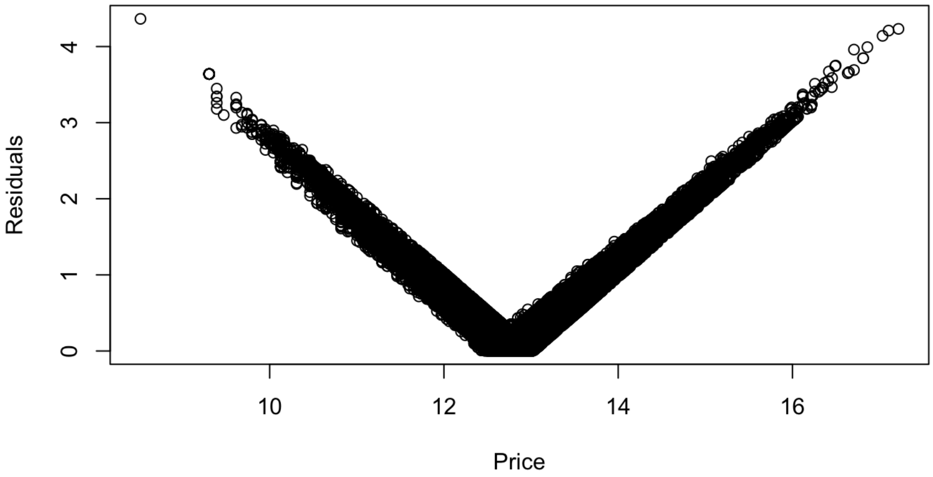

Fig. 6 Price Residual Plot

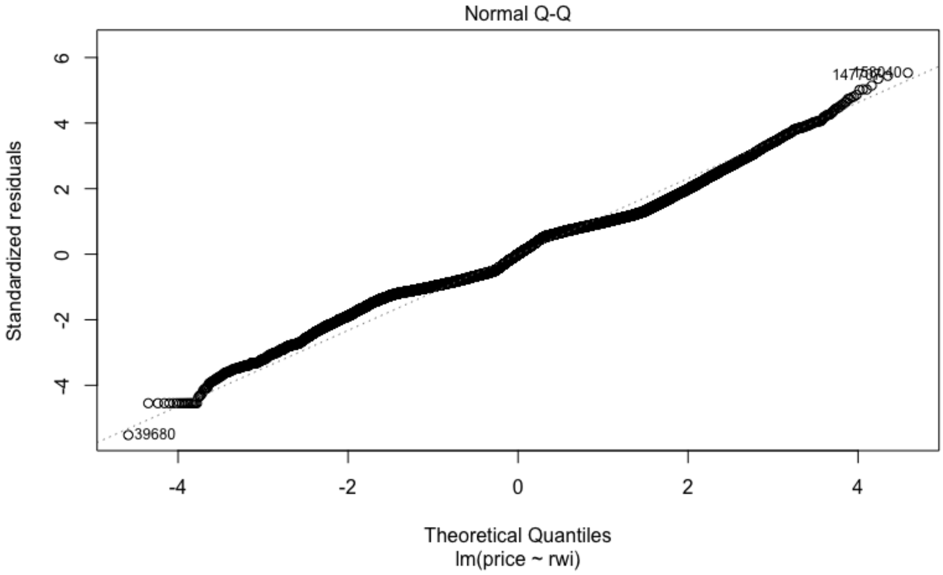

Fig. 7 QQ Plot of Weighted Regression

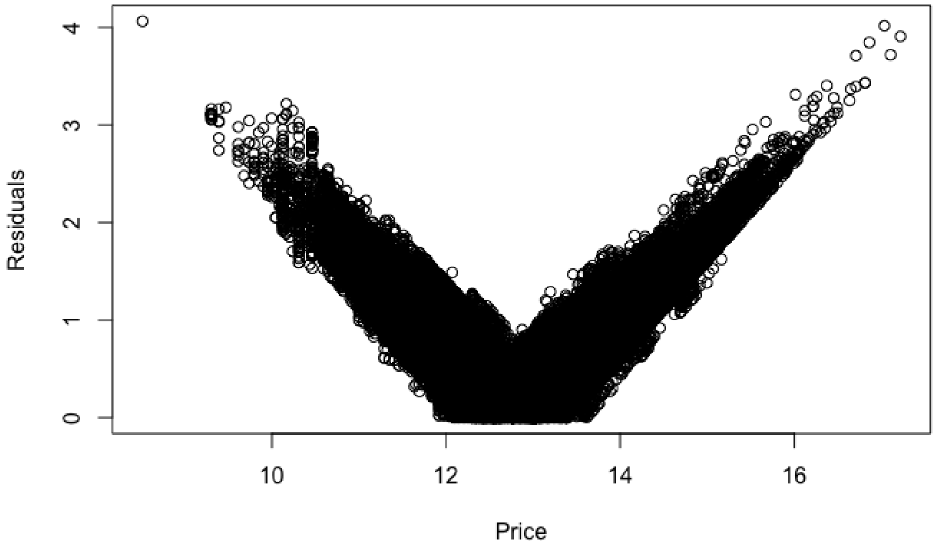

Fig. 8 Price Residual Plot of Weighted Regression

The two different interpretation of the source of errors highlight where house prices and the MRWI may fail in capturing how wealth is distributed. On the one hand, if house prices is correct in capturing what wealth is, then the MRWI should be an underestimation of wealth. If the wealth index is correct in capturing what wealth is, then house prices should exhibit a heavy-tail skew in its relationship to wealth. The two competing interpretation is the fundamental flaw in using these alternative data sources to measure wealth i.e. we need to assume one to be true to show the validity and weaknesses of another: circular reasoning. Thus, this merits the use of official statistics with reliable datasets to measure wealth.

Not only that, the two different interpretations also mean that the relative wealth index is unsuitable to study wealth inequality. On the one hand, if house price reflects wealth, then MRWI suggests that wealthy neighborhoods are not as rich which reduces measure of wealth inequality. On the other hand, if MRWI perfectly reflects wealth, then it underestimates housing wealth for the very rich and overestimates housing wealth for the very poor which reduces measure of wealth inequality. Thus, policy-makers and academics should be cognizant of this flaw in order to obtain a clearer estimate of wealth inequality. Thus, in Malaysia at least, the lack of variation in the wealth index seemed to suggests that wealth inequality is low relative to other countries, but we know this is not empirically rigorous to establish.

7 Policy Implication

A major implication of this study is on assessing the reliability of data on two fronts – whether the data reflects the underlying relationship it seeks to establish and whether the data reflects the implications of what it is measuring. In this particular context of the study, the Meta Relative Wealth Index (MRWI) seeks to establish a measurement of wealth in countries where the sources of such measurement is not readily and publicly available. To assess how well the index has achieved in establishing the measure of wealth for one of the countries covered under the dataset, I use house prices as a source of data that proxies wealth in a neighborhood. I chose Malaysia as a subject of this assessment both because the underlying creation of the index do not involve Malaysia as part of the countries that contribute to its ground-truth data i.e. the Demographics Health Survey and the fact that Malaysia has a developed data infrastructure in comparison to other countries covered in the dataset.

What I found is that the MRWI presents a weak correlation to house prices. Although the relationship I found was positive, the lack of variation in the MRWI contributed to the loss of information to distinguish between poor and wealthy neighborhoods. However, the MRWI does not perform as well in terms of predicting house prices in comparison to the highly aggregated 2020 census data. The reason why MRWI may not perform as well in terms of its predictive ability lies in two potential reasons – measurement error or biased measurement. We found that predicted errors the authors produce is not correlated to the error produced by our simple OLS equation assuming that house prices perfectly reflects wealth; in this case, there should be some data inaccuracies on wealth in Malaysia during the construction of the dataset. If we assume the wealth index perfectly reflects wealth, we found that the relationship between wealth index and the house prices is heavy-tailed – which means there is an underlying bias that drives house prices independent of wealth in this case.

Due to the fact that we cannot distinguish whether house prices or wealth index as the true representative of wealth makes the problem of measuring wealth difficult. The reason is because we need to assume one measure is true to derive the validity of the other measurement i.e. circular reasoning. The lack of reliable and well-defined datasets in developing countries would need stricter assumptions if it relies on these types of alternative datasets. Further, responsible and careful use of these types of data must be stressed for researchers in search of samples to study poverty in developing countries as well as policymakers wanting to target development policies.

Another particular group implicated by the findings of this study is businesses who rely on these types of datasets in order to generate wealth-targeted advertising. Since the MRWI utilized mobile phone usage is heavily weighted to generate wealth predictions and we know that social media advertising today relies on these types of data to generate advertisements, we can infer that there may not be huge wealth variations with regards to mobile phone use if we assume the MRWI to be true. However, since the production of such data is relatively inexpensive for a company such as Meta, the cost of such services for other enterprises that use Meta's data may be small enough to justify the expected returns from a targeted advertising campaign.

This study further highlights the difficulty in inferring wealth from auxiliary sources that are not defined by official statistics. Wealth is conceptually difficult to define due to its connection to the culture or the political institution or the social system in which it operates, especially in developing countries. Yet, wealth remains an important question in the social sciences. The recent surge in interest in wealth as a first-order concern is the result of a long-run development of income inequality. In Malaysia, the top 20% of income earners make up nearly 50% of income share before taxes[44]. Of course, this figure pales in comparison to the US where the top 10% income earners make up nearly 50% of income share[45][5]. In economics for example, researchers are interested in the effects of wealth on well-being[9]; in sociology, researchers are interested in the effects of wealth in the creation of social class[3][11][4]; in political science, researchers are interested in the role of wealth in shaping political institutions[46]. Although the social science community has widely studied this phenomena in the developed world, income and wealth inequality is still understudied in the developing country. Thus, in order to properly study the inequality phenomenon in the developing world: the right data, especially official statistics, must be updated and clarified

Acknowledgements. I would like to personally thank to Pieter Stek for offering valuable advice in the production of this paper.

Declarations

The author did not receive financial support from any organization for the submitted work. Further, the author declares no conflict of interest in the production of this paper. The data and code to reproduce the results will be made available in the supplementary materials

References

[1] Zucman, G.: Global Wealth Inequality. Annual Review of Economics **11**(1), 109–138 (2019) https://doi.org/10.1146/annurev-economics-080218-025852 . _eprint: https://doi.org/10.1146/annurev-economics-080218-025852. Accessed 2024-02-11

[2] Autor, D.H., Katz, L.F., Krueger, A.B.: Computing Inequality: Have Computers Changed the Labor Market?*. The Quarterly Journal of Economics **113**(4), 1169–1213 (1998) https://doi.org/10.1162/003355398555874 . Accessed 2024-02-09

[3] Keister, L.A.: The One Percent. Annual Review of Sociology **40**(1), 347–367 (2014) https://doi.org/10.1146/annurev-soc-070513-075314 . _eprint: https://doi.org/10.1146/annurev-soc-070513-075314. Accessed 2024-02-18

[4] Killewald, A., Pfeffer, F.T., Schachner, J.N.: Wealth Inequality and Accumulation. Annual Review of Sociology **43**(1), 379–404 (2017) https://doi.org/10.1146/annurev-soc-060116-053331 . _eprint: https://doi.org/10.1146/annurev-soc-060116-053331. Accessed 2024-02-18

[5] Piketty, T., Saez, E.: Income Inequality in the United States, 1913–1998*. The Quarterly Journal of Economics **118**(1), 1–41 (2003) https://doi.org/10.1162/00335530360535135 . Accessed 2024-02-09

[6] Auten, G., Splinter, D.: Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends. Journal of Political Economy (2023) https://doi.org/10.1086/728741 . Publisher: The University of Chicago Press. Accessed 2024-02-11

[7] Autor, D.H., Katz, L.F., Kearney, M.S.: Trends in U.S. Wage Inequality: Revising the Revisionists. The Review of Economics and Statistics **90**(2), 300–323 (2008) https://doi.org/10.1162/rest.90.2.300 . Accessed 2024-02-09

[8] Deaton, A.: The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. World Bank Publications, ??? (1997). Google-Books-ID: Mfe8ukMh_v4C

[9] Deaton, A.: Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll. Journal of Economic Perspectives **22**(2), 53–72 (2008) https://doi.org/10.1257/jep.22.2.53 . Accessed 2024-02-18

[10] Kahneman, D., Deaton, A.: High income improves evaluation of life but not emotional well-being. Proceedings of the National Academy of Sciences **107**(38), 16489–16493 (2010) https://doi.org/10.1073/pnas.1011492107 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-02-18

[11] Spilerman, S.: Wealth and Stratification Processes. Annual Review of Sociology

26, 497–524 (2000). Publisher: Annual Reviews. Accessed 2024-02-18

[12] Elvidge, C.D., Sutton, P.C., Ghosh, T., Tuttle, B.T., Baugh, K.E., Bhaduri, B., Bright, E.: A global poverty map derived from satellite data. Computers & Geosciences **35**(8), 1652–1660 (2009) https://doi.org/10.1016/j.cageo.2009.01.009 . Accessed 2024-02-09

[13] Ghosh, T., Powell, R.L., Anderson, S., Sutton, P.C., Elvidge, C.D.: Informal Economy and Remittance Estimates of India Using Nighttime Imagery. International Journal of Ecological Economics and Statistics™ **17**(P10), 16–50 (2010). Number: P10. Accessed 2024-02-11

[14] Ghosh, T., L. Powell, R., D. Elvidge, C., E. Baugh, K., C. Sutton, P., Anderson, S.: Shedding Light on the Global Distribution of Economic Activity. The Open Geography Journal **3**(1) (2010). Accessed 2024-02-11

[15] Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., Chen, L., Wu, J.: Evaluating the Ability of NPP-VIIRS Nighttime Light Data to Estimate the Gross Domestic Product and the Electric Power Consumption of China at Multiple Scales: A Comparison with DMSP-OLS Data. Remote Sensing **6**(2), 1705–1724 (2014) https://doi.org/10.3390/rs6021705 . Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2024-02-11

[16] Chen, X., Nordhaus, W.D.: Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences **108**(21), 8589–8594 (2011) https://doi.org/10.1073/pnas.1017031108 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-02-11

[17] Henderson, J.V., Storeygard, A., Weil, D.N.: Measuring Economic Growth from Outer Space. The American Economic Review **102**(2), 994–1028 (2012). Publisher: American Economic Association. Accessed 2024-02-11

[18] Weidmann, N.B., Schutte, S.: Using night light emissions for the prediction of local wealth. Journal of Peace Research **54**(2), 125–140 (2017) https://doi.org/10.1177/0022343316630359 . Publisher: SAGE Publications Ltd. Accessed 2024-02-11

[19] Maldonado, L.: Living in darkness: rural poverty in Venezuela. Journal of Applied Economics **26**(1), 2168464 (2023) https://doi.org/10.1080/15140326.2023.2168464 . Publisher: Routledge _eprint: https://doi.org/10.1080/15140326.2023.2168464. Accessed 2024-02-09

[20] Engstrom, R., Hersh, J.S., Newhouse, D.L.: Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being, Rochester, NY (2017). https://papers.ssrn.com/abstract=3090770 Accessed 2024-02-09

[21] Aiken, E., Bellue, S., Karlan, D., Udry, C., Blumenstock, J.E.: Machine learning

and phone data can improve targeting of humanitarian aid. Nature **603**(7903), 864–870 (2022) https://doi.org/10.1038/s41586-022-04484-9 . Number: 7903 Publisher: Nature Publishing Group. Accessed 2024-02-11

[22] Athey, S.: Beyond prediction: Using big data for policy problems. Science **355**(6324), 483–485 (2017) https://doi.org/10.1126/science.aal4321 . Publisher: American Association for the Advancement of Science. Accessed 2024-02-12

[23] Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264), 1073–1076 (2015) https://doi.org/10.1126/science.aac4420 . Publisher: American Association for the Advancement of Science. Accessed 2024-02-12

[24] Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016) https://doi.org/10.1126/science.aaf7894 . Publisher: American Association for the Advancement of Science. Accessed 2024-02-09

[25] Chi, G., Fang, H., Chatterjee, S., Blumenstock, J.E.: Microestimates of wealth for all low- and middle-income countries. Proceedings of the National Academy of Sciences **119**(3), 2113658119 (2022) https://doi.org/10.1073/pnas.2113658119 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-02-09

[26] Khalid, M.A.: Household Wealth in Malaysia: Composition and Inequality among Ethnic Groups. Jurnal Ekonomi Malaysia

[27] Guerrieri, V., Hartley, D., Hurst, E.: Endogenous Gentrification and Housing Price Dynamics

[28] USAID: Demographic and Health Surveys (various)

[29] Wolff, E.N., Zacharias, A.: Household wealth and the measurement of economic well-being in the United States. The Journal of Economic Inequality **7**(2), 83–115 (2009) https://doi.org/10.1007/s10888-007-9068-6 . Accessed 2024-02-18

[30] Ismail, S., Ho, G., Kunasekaran, T., Mohamad Firouz, A.M.: Residential Settlements and Spatial Inequalities. Khazanah Research Institute, Kuala Lumpur (2022)

[31] Khalid, M.A.: Malaysia Household Wealth Distribution: Current Evidence and Future Prospects. International Journal of Management Studies **19**(1), 73–86 (2012). Number: 1. Accessed 2024-02-09

[32] Tumin, S.A.: Informality in the Malaysian Economy. Khazanah Research Institute (2021)

[33] Abdul Hamid, H., Megat Muzafar, P.M.: Social Insurance Pension: Saving for

retirement collectively. Khazanah Research Institute (2023)

[34] Sartirano, D., Kalimeri, K., Cattuto, C., Delamónica, E., Garcia-Herranz, M., Mockler, A., Paolotti, D., Schifanella, R.: Strengths and limitations of relative wealth indices derived from big data in Indonesia. Frontiers in Big Data **6** (2023). Accessed 2024-02-11

[35] DiPasquale, D., Glaeser, E.L.: Incentives and Social Capital: Are Homeowners Better Citizens? Journal of Urban Economics **45**(2), 354–384 (1999) https://doi.org/10.1006/juec.1998.2098 . Accessed 2024-02-18

[36] Sutton, P.C., Elvidge, C.D., Ghosh, T.: Estimation of Gross Domestic Product at Sub-National Scales using Nighttime Satellite Imagery **8** (2007)

[37] Nischal, K.N., Radhakrishnan, R., Mehta, S., Chandani, S.: Correlating nighttime satellite images with poverty and other census data of India and estimating future trends. In: Proceedings of the 2nd ACM IKDD Conference on Data Sciences. CODS '15, pp. 75–79. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2732587.2732597 . https://doi.org/10.1145/2732587.2732597 Accessed 2024-02-08

[38] Acemoglu, D., Johnson, S., Robinson, J.A.: The Colonial Origins of Comparative Development: An Empirical Investigation. The American Economic Review **91**(5), 1369–1401 (2001). Publisher: American Economic Association. Accessed 2024-02-24

[39] Center, N.P.I.: Open Housing Transaction Data

[40] Case, K., Shiller, R.: The Efficiency of the Market for Single-Family Homes. Technical Report w2506, National Bureau of Economic Research, Cambridge, MA (February 1988). https://doi.org/10.3386/w2506 . http://www.nber.org/papers/w2506.pdf Accessed 2024-02-18

[41] Chetty, R., Hendren, N., Kline, P., Saez, E.: Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States *. The Quarterly Journal of Economics **129**(4), 1553–1623 (2014) https://doi.org/10.1093/qje/qju022 . Accessed 2024-02-09

[42] Chetty, R., Hendren, N., Katz, L.F.: The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. American Economic Review **106**(4), 855–902 (2016) https://doi.org/10.1257/aer.20150572 . Accessed 2024-02-09

[43] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)

[44] Statistics Malaysia, D.: 2020 Malaysian Census

[45] Piketty, T., Zucman, G.: Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010 *. The Quarterly Journal of Economics **129**(3), 1255–1310 (2014) https://doi.org/10.1093/qje/qju018 . Accessed 2024-02-09

[46] Scheve, K., Stasavage, D.: Wealth Inequality and Democracy. Annual Review of Political Science **20**(1), 451–468 (2017) https://doi.org/10.1146/annurev-polisci-061014-101840 . _eprint: https://doi.org/10.1146/annurev-polisci-061014-101840. Accessed 2024-02-18