To Draw a Relative-Poverty Line: An Approach using Mixture Models and CART*

Sudeshna Maitra

Department of Economics

York University

4700 Keele Street

Toronto ON M3J 1P3

Canada

Email: smaitra.econ@gmail.com (preferred)

maitra@yorku.ca

April 2024

* I would like to thank Barry Smith for his invaluable insights on the ideas presented in this paper.

To Draw a Relative-Poverty Line: An Approach using Mixture Models and CART

April 2024

ABSTRACT

Prior research has shown that when household durable expenditure levels segregate into clusters, the lowest cluster can be interpreted as relatively poor households that are vulnerable to long-run poverty. A simple rule to classify these households would be invaluable both for policy as well as analytical ends.

We propose a methodology for deriving a relative-poverty "line" – more accurately, a relative-poverty *schedule* (RPS) – using a mixture model (to identify relatively poor households probabilistically as the lowest cluster of durable ownership) and a classification (CART) tree (to determine a rule for classifying these households). We derive an RPS using data from the urban subsample of Indian NSS, 1999-00. We claim that the RPS – or classification rule – that emerges from this process is driven by natural clusters and patterns in the data, and therefore has a greater grounding in economic information than existing cutoffs-based approaches (viz. relative poverty lines as pre-determined percentages of median income).

Keywords: relative poverty line/schedule, poverty measurement, classification tree, mixture model, durable goods, poverty of opportunities

1. Introduction

A common method for identifying relatively poor households is to set a relative-poverty line (RPL) that is a pre-determined percentage of median income in the overall distribution (Ray, 1998; Foster, 1998). All households with income below the RPL are classified as relatively poor. But in what sense are households defined in this manner *"poor"* or experiencing a state of *deprivation*? Why are households below 60% of median income deserving of our (or policy) attention?

Maitra (2016, 2017, 2023) proposes an approach that uses economic theory, simulations and empirical analysis to define and measure relative poverty. The goal of the approach is to understand the *source* of deprivation of relatively poor households, viz. those who find themselves left behind others in the society they reside in. The approach uses durable ownership to measure living standards, primarily because it is easier to measure than expenditure or income, and has been used in previous research to measure poverty. However, durable ownership is also a *choice* made by households and thus depends on preferences as well as income constraints (the driver of poverty). It is critical, therefore, to understand the fundamental process of household decision-making that connects household incomes to durable choice.

Maitra (2023) constructs a dynamic overlapping-generations model of income generation and durable choice, and simulates the relationship between household income and durable expenditure in steady state. They demonstrate that in steady state, the (optimal) durable ownership of households segregates into clusters, and that the lowest cluster is interpretable as households in relative poverty. Moreover, relative poverty is synonymous with a *poverty of opportunities in the long run* – the households in the lowest cluster are "left behind" because they fail, systematically, to access the channels of income generation (the labour market and the marriage market) in the model economy.

In addition, Maitra (2016, 2017) shows that a mixture model may be used to *empirically identify* the clusters in durable ownership predicted by the model above (Maitra, 2023). Using urban data from the Indian National Sample Survey (1993-4, 1999-00, 2004-05), Maitra (2016, 2017) estimates the size and durable ownership density of the lowest class – the cluster of households in relative poverty in urban India in the 1990s.

The mixture model uses natural clusters in durable ownership in the *data* to identify the lowest class, which we know – from theory – can be interepreted as households that are vulnerable to a long run poverty of opportunities. Thus, the mixture estimation process delivers an objective, *data-driven* criterion for identifying relative poverty, with an interpretation backed by our theoretical understanding of the underlying economic process of income generation and durables choice.

While the mixture model enables us to identify the lowest class of households as a whole, the estimates only yield class membership *probabilities* of individual households, conditional on their durable ownership. To classify specific individual households as relatively poor (or not), we need a well-defined rule that places households into one of the two groups with *certainty* – much like a relative-poverty line of household income or expenditure. This is the objective of the current paper: (1) to propose an "Opportunities Measure" of Relative Poverty (OMRP) defined by the lowest class of durable ownership, and (2) to derive a simple classification rule (in terms of household durable ownership) to empirically identify households per the OMRP.

We propose an approach that uses a classification tree (CART) to derive such a mixture-consistent relative-poverty line (or rule). We claim that households that meet the criteria established by this rule are vulnerable to long-run poverty as they are unable to systematically access the channels of income generation available in the economy. We then demonstrate the approach in the urban subsample of the Indian National Sample Survey (NSS, 1999-00). We estimate a relative-poverty *schedule* (or RPS) – rather than a single RPL – of per capita monthly expenditure conditional on owning specific durables in this data. We find that the RPS has a far lower misclassification rate than a single RPL (2% vs 16%).

The spirit of the CART approach is similar to that of Gordon and Townsend (1991), who derive a poverty line in weekly disposable income using linear discriminant analysis over a range of questions about social and material deprivation in the UK (see also Fusco et al, 2010; Atkinson, 2019). Our approach has the additional advantages (1) of being rooted in a theoretical understanding of what it means to be relatively poor (a poverty of opportunities), *why* clusters exist, and *why/how* the lowest cluster of households is an appropriate measure of relative poverty; and (2) of offering a RPS that can capture non-linearities in the data.

The main contribution of this paper is that it provides a methodology for generating a simple, country-specific rule for identifying households vulnerable to relative poverty, which we argue is a poverty of opportunities in the long run. The rule (or RPS) is determined by natural, endogenously generated clusters in country-specific data – what *is true in the data* rather than any notion of who *should be* considered to be relatively poor – and its interpretation is supported by a theoretical understanding of the process of income generation and class formation. As such, the RPS is more meanigful than an RPL defined as a percentage of median income. The simplicity of the rule recommends its use in identifying relatively poor households for policy or for analytical ends.

2. Background: Identifying relative poverty

2.1 Why durable ownership?

Durable ownership has been used in prior studies of poverty measurement (Filmer and Pritchett, 2001), primarily because it is easy to observe and measure. Maitra (2016, 2017) use durable ownership in their study to facilitate the comparison of poverty estimates in India over the 1990s. The Indian National Sample Survey reduced the recall periods for reporting expenditures in the 1999-00 round, leading to concerns that this may have artificially increased expenditures – and lowered expenditure-based poverty estimates – reported in 1999-00 compared with previous rounds (Deaton and Kozel, 2005). Durable ownership data was drawn from the question – "is good X in use in the household at the time of the survey?" – and is immune to the change in recall periods in NSS.

But durable ownership is a *choice* made by households, and hence dependent on preferences as well as constraints (the primary indicator of poverty). How do we distinguish between low durable ownership due to an income constraint (indicative of poverty) and that stemming from a dis-preference for durable goods? To answer this question, Maitra (2023) constructs a dynamic, overlapping generations model of long run income generation and household choice. There are two channels of income generation in the model – the labour market and the marriage market. Households choose to invest in the education of their offspring (that enhances the probability of higher future labor-market income) or to purchase durable goods (that have utility value but also signal the social status of the household). Higher social status signalled by (observably) more durables leads to a better match in the marriage market, thereby enhancing future income earning potential. Maitra (2023) shows that in steady state, the optimal durable expenditure levels of households segregates into clusters owing to a "wealth-begets-wealth" mechanism. Moreover, the lowest cluster of households are unable to access either channel of income generation in the long run. In this sense, the lowest cluster of households – or, households in relative poverty – are vulnerable to long-run poverty, as they fall behind due to lack of access to existing income-generation opportunities.

A mixture model is a tool for the identification of latent classes from data on an overall distribution (McLachlan and Peel, 2001). The "criterion" for class identification inherent in the mixture approach is driven by natural clusters in the variable of interest (here durable ownership). Since our theoretical model indicates that the lowest cluster of durable expenditure represents relative poverty, a mixture model is a natural tool for determining a criterion to identify relative poverty.

2.2 A Criterion for Relative Poverty: The Mixture Model

Let Y be the total number (out of 8) of durable goods owned by a household, $Y = 0, 1, 2, ..., 8$.[1] A three component mixture model (McLachlan and Peel, 2001) postulates that there are 3 latent classes in the population, where the density of Y in class$-j$ households is given by $\phi_j(Y;\theta_j)$ (θ_j are the parameters of the density function; $j = 1, 2, 3$). The density of Y in the population is then given by:

$$P(Y) = \pi_1\phi_1(Y;\theta_1) + \pi_2\phi_2(Y;\theta_2) + (1 - \pi_1 - \pi_2)\phi_3(Y;\theta_3) \tag{1}$$

where π_j denotes the proportion of households in the population that belongs to class j ($j = 1, 2$). In keeping with Maitra (2016, 2017), we assume that ϕ_j denotes a binomial function with parameters $(8, p_j)$ where p_j is the probability with which a class$-j$ household owns a durable good. Rewriting (1) with the binomial parameters, we get:

$$P(Y) = \pi_1\phi_1(Y;p_1) + \pi_2\phi_2(Y;p_2) + (1 - \pi_1 - \pi_2)\phi_3(Y;p_3) \tag{2}$$

The mixture parameters $(\pi_1, \pi_2, p_1, p_2, p_3)$ combined with Bayes' Rule imply conditional probabilities $\gamma_j(Y)$ of class membership, viz. that a household belongs to any class j, conditional on durable ownership:

[1] We use 8 goods in keeping with Maitra, 2017. These are radio, tv, electric fan, ac, fridge, bike, motor bike and car.

$$\gamma_j(Y) = P(j/Y) = \frac{P(j,Y)}{P(Y)} = \frac{\pi_j \phi_j(Y;p_j)}{\pi_1\phi_1(Y;p_1) + \pi_2\phi_2(Y;p_2) + (1-\pi_1-\pi_2)\phi_3(Y;p_3)} \tag{3}$$

$j = 1, 2, 3.$

An Expectations-Maximization (EM) algorithm (McLachlan and Krishnan, 1996) can be used to obtain estimates of the parameters – the sizes of each class ($\widehat{\pi}_1$,$\widehat{\pi}_2, 1-\widehat{\pi}_1-\widehat{\pi}_2$), and the densities of durable ownership of each class $\phi_1(Y;\widehat{p}_1), \phi_2(Y;\widehat{p}_2), \phi_3(Y;\widehat{p}_3)$. The estimated conditional probabilities $\hat{\gamma}_j(Y)$ of class membership would then be given by:

$$\hat{\gamma}_j(Y) = \frac{\hat{\pi}_j \phi_j(Y;\hat{p}_j)}{\hat{\pi}_1\phi_1(Y;\hat{p}_1) + \hat{\pi}_2\phi_2(Y;\hat{p}_2) + (1-\hat{\pi}_1-\hat{\pi}_2)\phi_3(Y;\hat{p}_3)} \tag{4}$$

$j = 1, 2, 3.$

An appealing feature of the mixture model is that households are placed in classes with *probabilities* $\hat{\gamma}_j(x)$ instead of with certainty.[2] Table 1 presents the mixture model estimates reported by Maitra (2017) using urban data from the Indian National Sample Survey, 1999-00. Notice, for example, that a household with 0 durables belongs to the lower class with 87% probability. This probabilistic assignment matches well with predictions from the theoretical model (Maitra, 2023) that captures the fact that durable ownership is a choice variable. A household with 0 durables could represent a household in relative poverty (with a binding income constraint) or a household in the middle (or upper) class that *chooses not to* purchase any durables. The mixture model uses representative data from the entire population to estimate the relative prevalence of households of each type (viz. constraint-driven vs. preference-driven), given their durable ownership. These proportions are reflected in the estimates $\hat{\gamma}_j$.

The next step on the road to defining a relative poverty "line" is to look at the expenditures of households mixture-identified as relatively poor using durable ownership, and determine an expenditure cutoff such that all households below the cutoff can be classified as relatively poor. This means, however, that we wish to translate the information contained in the mixture-estimated *probabilities* $\hat{\gamma}_j$ of being relatively poor, to an expenditure cutoff that classifies households as relatively poor *with*

[2] Maitra (2016, 2017) show also that the per capita expenditure (PCE) ranges of the lower, middle and upper classes are overlapping. This is counter to the idea of a poverty "line" expenditure that separates the poor and non-poor with certainty.

certainty. To permit this switch in interpretation, we introduce an additional assumption – the "Assignment Assumption" – explained in the next section.

3. Deriving the Opportunities Measure of Relative Poverty (OMRP) and a Relative Poverty Rule

3.1 The Assignment Assumption (AA)

In order to derive a level of expenditure below which all households are in the lowest class, we need to translate the information contained in the class membership probability $\hat{\gamma}_1(x)$ $(0 \leq \hat{\gamma}_1(x) \leq 1)$,into class assignments, say ω $(\omega \in \{0, 1\})$. Suppose class 1 represents the lowest class. Estimates of $\hat{\gamma}_1(x)$ indicate the proportion of households with durables x $(= 0, 1, ..., 8)$ that belong to the lowest class, and are hence relatively poor. However, there are infinitely many ways of *assigning* households with x variables to the lower class, so as to be consistent with the estimated value of $\hat{\gamma}_1(x)$. Which of these assignments should we choose to examine in order to define the relative poverty line?

We propose that for every level of durable ownership x, households in the lowest $\hat{\gamma}_1(x)^{th}$ percentile of per capital monthly expenditure be assigned to the lower class (viz. classified as relatively poor). We call this the "Assignment Assumption". The Assignment Assumption (henceforth referred to as AA) combined with the mixture-estimated class-membership probabilities constitutes our proposed measure of relative poverty, viz., the Opportunities Measure of Relative Poverty (OMRP). OMRP (Table 1(b)) specifies a unique assignment of households as relatively poor, consistent with the mixture estimates.

Why use AA? A primary motivation for deriving a relative poverty line is to specifically identify households who are most vulnerable to long-run poverty; presumably so they can receive policy assistance. The mixture estimates reflect long-run vulnerability but these estimates do not point to a unique assignment of households to the lower class (relative poverty). The AA assumption uses per capita monthly expenditure – a widely accepted measure of current economic well-being – to choose a unique assignment. AA formalizes our premise, that of all the relative poverty assignments consistent with the mixture estimates, the one that includes households in the lowest $\hat{\gamma}_1(x)^{th}$ percentile of expenditure by x, represents households that are most likely to be vulnerable to long-run poverty and in need of policy assistance.

Notice that the OMRP is, *itself,* a rule for classifying households as relatively poor! This rule is a *schedule* of PCE-cutoffs – one for each level of total durable goods owned – below which households are in the lowest class. Using this rule, however, requires that we possess data on the PCE levels of households as well as the *total* number of the 8 durables (viz. fan, TV, radio, fridge, AC, bike motor bike, car) the household owns. This makes the rule hard to interpret (*which* durables does a household with x goods own?) and requires that we include a very large number of questions in the questionnaire. Is it possible to derive a *simple* rule – that is (1) easier to interpret, and (2) requires fewer and more easily verifiable questions – that could classify households as relatively poor? Classification Trees (CART) provide a method for doing this.

3.2 Classification Trees (CART)

A Classification Tree approach (Breiman et al, 1984) is useful for generating easily interpretable "rules" to classify (or predict) outcomes. The approach builds on a process of recursive partitioning, which is an iterative process of splitting the data into partitions; and then splitting the partitions further on each of the branches. At each stage, the partition (and the rule for partition) is such that the the "purity" of outcomes in the partitioned group (or "node") is maximized.

In our application, the outcome of the classification tree process is a binary variable denoting membership in the lower class. Our goal is to find a rule, expressed as a cutoff in per capita monthly expenditure, that will partition the data into two groups – one group concentrated ("pure") in "lower-class" households, and the other group that is pure in non-lower classes. The rule provided by the tree will be interpreted as the relative poverty line (or schedule).

The classification tree approach is appealing as it provides a non-parametric and non-linear way to understand how the outcome variable is related to the predictors of interest. The classification rule is expressed as a series of sequential questions that outline a path for classifying the relative poverty status of specific households.

We run classification trees on the entire urban subsample of Indian NSS, 1999-00. The results are presented and discussed in the next section.[3]

[3] We choose to use the entire sample of households to run the tree – and not a "training sample" of a random subset of households in the data – as our objective is expository; the goal here is to use as much information as possible in deriving

4. Results and Discussion

Tables 2-4 present three different relative-poverty rules predicted by classification trees T1-T3.

Tree T1 (Table 2) uses only PCE as a predictor of relative poverty status. The rule presented by T1 indicates an urban relative poverty line – or "cutoff" – of Rs. 409 per capita per month in 1999-00. Interestingly, this estimate (Rs. 409) is quite close to the official absolute poverty line of Rs. 465 for urban India in 1999-00. This provides empirical validation for the durables-based mixture analysis that forms the backbone of identification of relatively poor households (the dependent variable in the classification tree).

However, the recursive partitioning process identified by T1 partitions households in three steps – the first split of households occurs at a PCE of 481, then at 313 and finally at 409. Put together, the different splits indicate that all households with PCE less than 409 per month should be classified as relatively poor. However, the three-step classification suggests that there may be a concentration of similar outcomes around more than one value of PCE, of which 409 may be an "average".

Indeed, Rule T1 misclassifies 16.24% of the households – 11.92% of poor households are (mis)classified as non-poor and 4.32% of non-poor households are (mis)classified as poor. One of the factors driving the misclassification rate is embedded in the very nature of this intellectual exercise – attempting to generate a certain (i.e. probability = 1) classification of households from probabilities (between 0 and 1) of class membership.[4] The misclassification error is due, in part, to the goal of deriving the relative poverty line as a cutoff.

But this suggests a different question: what about a classification rule that does not depend on a *single* PCE-cutoff but a *schedule* of PCE-cutoffs? We explore this idea in trees T2 and T3 described below (Tables 3-4).

Tree T2 (Table 3) uses the ownership of *individual* durable goods to predict relative poverty status. The rule that emerges classifies any household that does not own a tv, a fan and a radio as relatively

the classification rule for relatively poor households. The analysis is easily extended to using training and testing samples as well. We use the package 'rpart' (Therneau et al, 2023) in R to conduct the recursive partitioning process described above.

[4]As mentioned in footnote 2, the mixture model, which generates the class membership probabilities, predicts *overlapping ranges* of PCE among households in different classes (Maitra (2016, 2017)). The classification tree attempts to extract a single PCE cutoff that *separates* classes. This may be a significant source of misclassification, when using CART to classify outcomes using PCE alone.

poor. Rule T2 misclassifies 7.8% of households (3.6% of poor households misclassified as non-poor; 4.2% of non-poor households misclassified as poor) – much reduced from the misclassification rates of T1.[5]

Tree T3 now uses PCE *and* individual durable ownership as predictors of relative poverty (Table 4). Misclassification reduces considerably to only 2.5% households overall (2% of poor households misclassified as non-poor and 0.5% of non-poor households classified as poor). More interestingly, a nuanced picture emerges of *who* constitutes households in relative poverty, which enables us to define a relative-poverty schedule (RPS) of PCE cutoffs. Here are the salient features of the rule that emerges from T3, and the RPS it implies (Table 5):

1. Only 4 durables – fan, tv, radio and bike – matter in the determination of who is relatively poor. The survey questionnaire for identifying poor households need ask respondents about their ownership only of these 4 specific (observable) durable goods.

2. A *different* PCE-cutoff is obtained depending on the specific durables owned by households. This finding is the basis of our derivation of a relative-poverty *schedule* instead of a single relative-poverty line for identifying households vulnerable to long run poverty. The relative-poverty schedule (henceforth called RPS) implied by T3 is presented in Table 5.

3. The PCE-cutoff is lower for higher total ownership of durables x. This follows from the fact that relative poverty (the outcome variable) is defined using mixture-estimated $\hat{\gamma}_1(x)$ and assumption AA. It is interesting, to note that the PCE-cutoff for households with no durables (Rs. 1310) is 3-4 times the other positive PCE-cutoffs (Table 5). This suggests a large dispersion of PCE reported by households with no durables. What could be driving this? Are these households exposed to greater income volatility or large transitory expenditures (frequent migration, policy assistance during natural disasters/elections etc)? Or could they be more prone (relative to higher-durable households) to the reporting bias from changing recall periods in NSS, 1999-00?

[5] Note that the mixture model uses the *total* number (of 8) durable goods to estimate the size and characteristics of classes (see Section 2.2 and footnote 1). The predictors in T2 are 8 binary variables indicating whether the *individual* goods comprising the total – viz. radio, fan tv, ac, fridge, bike, motor bike, car – are owned by the household.

Whatever the reason for the large PCE-dispersion, our results suggest that households with no durables – even those who lie above the official poverty line of 465 (but below 1310) – might be vulnerable to long-run poverty. At the same time, we obtain a sense of the upper limit of PCE (viz. 1310) beyond which we may cease to worry about 0-durable households being vulnerable to poverty.

4. The RPS in Table 5 represents a set of criteria that is based on *what is true in the data*, as opposed to *what should be true based on exogenous notions* of basic needs/poverty (e.g. income cutoffs). This is a subtle but key distinction between the CART approach and the cutoffs-based approach that must be borne in mind when interpreting the findings. Consider, for example, 3 households A, B and C each with a PCE of 310, where A owns a fan only, B owns a fan and a tv and C owns a tv only. The RPS indicates that the PCE-cutoff for both A and B is 311. Thus, A and B are both classified as relatively poor, despite the fact that B owns an additional good – a tv. But the RPS also indicates that C – who has the same PCE as A and B, and only a tv should nevertheless be classified as non-poor! These patterns appear to be counter-intuitive.

The conundrum arises because our "intuition" about what *should be* true stems from *exogenously* determined conceptualizations – of basic needs/poverty (viz., income cutoffs) and ("well-behaved") preferences. The rule derived by the classification tree is based on *endogenous* patterns in the observed data. Our theoretical model of class formation provides the critical foundation that permits us to interpret the (endogenously determined) lowest class as households in relative poverty.[6]

That the RPS represents what is true in the durable ownership data – and not what should be true per some exogenously given notion of poverty – makes it especially useful for comparing the "empirical" poverty status of any two households, given their PCE level and durable ownership profile (of radio, tv, fan, bike). Consider another 2 households M and N. M owns 1 durable good (of the 8) and reports a PCE of 445 (just below the absolute poverty line of 454). N owns 0 durables (fewer than A) but reports a PCE of 490 (just above the absolute poverty line). Which household is "poorer"? At an

[6]Household preferences (which are accounted for in the theoretical model) or the differential price/quality of goods owned by different households are unlikely to matter for our identification. See the discussion in Maitra (2016) on the natural weighting of goods by their price/quality in the variable that represents their sum. E.g., households with better tv's are also likely to own a higher total number of goods than households with tv's of poorer quality. It is sufficient to look at the total number of durables owned in the mixture estimation, without attempting to ascertain the differential quality of goods across households.

absolute level (represented by the poverty line of 454) M is poor, and N is not. But if we use our RPS, the answer to which household is "more relatively poor" depends on (1) which durable M owns, and (2) whether each household's PCE is below or above the RPS-defined PCE for their durable profile. Per our (data-driven) RPS, N – with 0 durables and a PCE of 490 (less than 1310) – belongs to the lower class and is hence classified as relatively poor. Household M would be classified as relatively poor – or not – depending on which of the 4 durables (in the RPS) they own! If the single durable owned by M is a tv or a fan, then the RPS places M in the middle/upper class; in this event, N is "more relatively poor" than M. But if M owns only a radio or only a bike, the RPS would place it in the lower class, the same class as N.[7] Thus, even though M is (absolutely) poorer than N in this example, N is at least as (if not more) vulnerable to a long-run poverty of opportunities as M – a characteristic that is reflected in its durable ownership profile, and the occurrence – in the data – of households of similar profiles within different clusters/classes.

The above example also demonstrates the importance of non-linearities in the relationship between individual durable ownership and PCE, and of how this *observed* relationship varies across households. Our bottom-up approach to RPS-determination – (1) a theoretical model of income-generation/durable choice, (2) a mixture model and (3) a classification tree – provides a means of estimating as well as interpreting these relationships and how they influence long-run poverty, in a country-specific setting. We hope this will be a useful tool for researchers and policy makers who care about long-run poverty outcomes.[8]

5. Conclusion

The lowest class of durable ownership – or relatively poor households – has been shown in prior

[7]Most importantly, the fact that the RPS treats ownership of a tv or a fan differently than a radio or a bike is not one driven by an exogenous "judgement" about what these goods mean about household preferences/constraints, but, simply, by what the data reveals to be true of the lowest vs. upper classes. (These findings provide justification, in fact, for using tv ownership as an indicator of household well-being in India – a frequently employed premise in empirical studies based on Indian household data.)

[8]Note that the classification tree approach may also be applied to outcome variables obtained using alternative criteria for relative poverty (instead of OMRP). For instance, we repeat the CART analysis using an outcome variable derived from the standard definition of relative poverty, viz. households that lie below 60% of the median PCE in the sample. The individual durable goods that emerge in the CART-generated rules are similar in both cases (viz. fan, tv, radio, bike).

research (Maitra, 2023) to be vulnerable to long run poverty. A simple rule to classify these households would be invaluable both for policy as well as analytical ends.

We propose a methodology for deriving a relative-poverty "line" – more accurately, a relative-poverty *schedule* (RPS) – using a mixture model (to identify relatively poor households probabilistically as the lowest cluster of durable ownership) and a classification tree (to determine a rule for classifying these households). We derive an RPS using data from the urban subsample of Indian NSS, 1999-00. We claim that the RPS – or classification rule – that emerges from this process is driven by natural clusters and patterns in the data, and therefore has a greater grounding in economic information than existing cutoffs-based approaches (viz. percentages of median income as relative-poverty lines).

References

[1] Atkinson, A. B. (2019). Measuring poverty around the world. In Measuring Poverty around the World. Princeton University Press.

[2] Filmer, D., Pritchett, E. (2001). Estimating Wealth Effects Without Expenditure Data – or Tears: An Application to Educational Enrolment in States of India. *Demography*, 38(1): 115-132.

[3] Foster, J. E. (1998). Absolute vs. Relative Poverty. *American Economic Review*, 88(2), pp. 335-341.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., Stone C. J. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

[5] Deaton, A., Kozel, V. (2005). Data and Dogman: The Great Indian Poverty Debate. *The World Bank Research Observer*, 20, pp. 177-199.

[6] Fusco, A., Guio, A. C., Marlier, E. (2010) "Characterising the income poor and the materially deprived in European countries." In A. B. Atkinson and E. Marlier (eds)., Income and living conditions in Europe. Eurostat, Luxembourg.

[7] Gordon, D., Townsend, P. (1991). “What is enough? New evidence on poverty allowing the definition of a minimum benefit.” In M. Adler, C. Bell, J. Clasen, A. Sinfied (eds)., The Sociology of Social Security. Edinburgh University Press, Edinburgh.

[8] Maitra, S. (2016). The poor get poorer: Tracking relative poverty in India using a durables-based mixture model. *Journal of Development Economics*, 119, pp.110-120. Available: https://www.sudeshnamaitra.com/Maitra-rel-poverty-mixture-Jun2015.pdf

[9] Maitra, S. (2017). Class matters: tracking urban inequality in post- liberalization India using a durables-based mixture model. *Applied Economics Letters*, 24(17), pp. 1203-1207. Available: https://www.sudeshnamaitra.com/Maitra-mixture-ineq-Nov2016.pdf

[10] Maitra, S. (2023). On the Theory and Measurement of Relative Poverty using Durable Ownership Data. Working Paper, September 2023. Available: https://www.sudeshnamaitra.com/Maitra_relative_poverty_Sep2023.pdf

[11] McLachlan, G. J., Krishnan, T. (1996). The EM Algorithm and Extensions. Wiley Series in Probability and Statistics.

[12] McLachlan, G. J., Peel, D. (2001). Finite Mixture Models. Wiley Series in Probability and Statistics.

[13] Ray, D. (1998). Development Economics. Princeton University Press.

[14] Therneau, T. M., Atkinson, E. (2023). An Introduction to Recursive Partitioning Using the RPART Routines. The Comprehensive R Archive Network. Available at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (last accessed: April 26, 2024).

Table 1: Mixture estimates from Maitra, 2017

Indian National Sample Survey (NSS), 55th Round, 1999-00, Urban sub-sample

1A: Mixture Model Parameter Estimates

Class Proportions			Binomial Probability		
Lower	Middle	Upper	Lower	Middle	Upper
π_1	π_2	π_3	p_1	p_2	p_3
0.200	0.621	0.179	0.035	0.341	0.590

1B: Mixture-Estimated Probability of belonging to the lower class (being relatively poor), γ_1 (y)

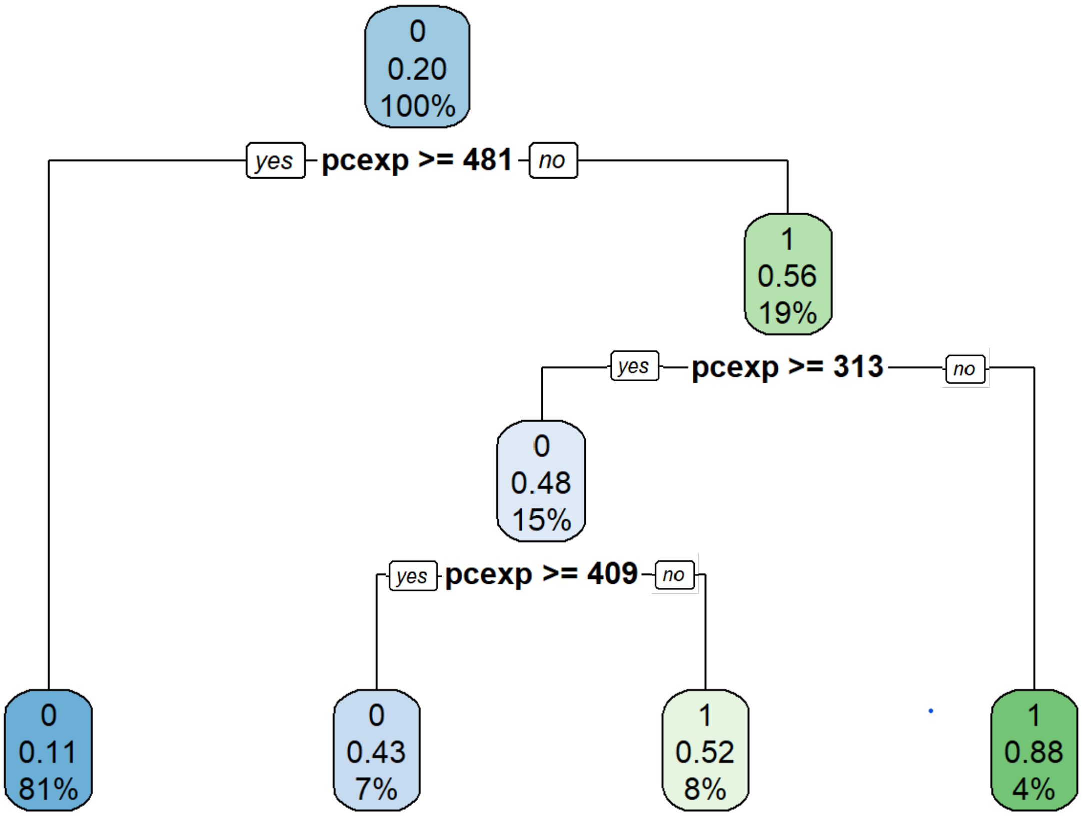
(The Opportunities Measure of Relative Poverty (OMRP) = All households that own y durables and lie below the $\gamma(y)^{th}$ percentile of PCE)

Total durables Owned (Y)	γ_1 (y)	γ_1 $(y)^{th}$ percentile of per capita monthly household expenditure*
0	0.871	1310
1	0.320	480
2	0.031	310
3	0.002	232
4	0.000	147
5	0.000	266
6	0.000	329
7	0.000	417
8	0.000	740

*Data Appendix: Indian NSS, 55th Round, 1999-00, Urban sub-sample (N=48921)

Total durables Owned (Y)	Obs.	PC Monthly Household Expenditure (Sample Mean: 1018.72, SD: 1535.32)								
		Percentile					Mean	SD	Min	Max
		25	50	75	90	99				
0	8417	424	637	1000	1414	2800.74	793.29	720.32	17	40562
1	6692	433	629	1012	1503.7	3252.19	838.82	1335.95	55	84424
2	8743	496	691	1034	1529	3218.56	862.85	598.32	49	10838
3	9600	547	756	1097	1646.9	3301.93	953.01	2188.08	139	205987
4	7592	657	932	1382.75	1952	4130.24	1179.92	2236.77	147	151196
5	4539	842	1180	1673	2321	4761.8	1393.58	917.23	266	20212
6	2413	946.5	1336	1874	2671.2	5805.26	1585.55	1087.66	329	15144
7	806	1074.75	1491.5	2091.25	3138.8	8084.47	1838.39	1455.09	417	15904
8	119	1378	1784	2490	3734	14717.2	2207.65	1712.211	740	16834

Table 2: Classification Tree (T1)
Outcome: If relatively poor
Predictor: PCE



KEY:

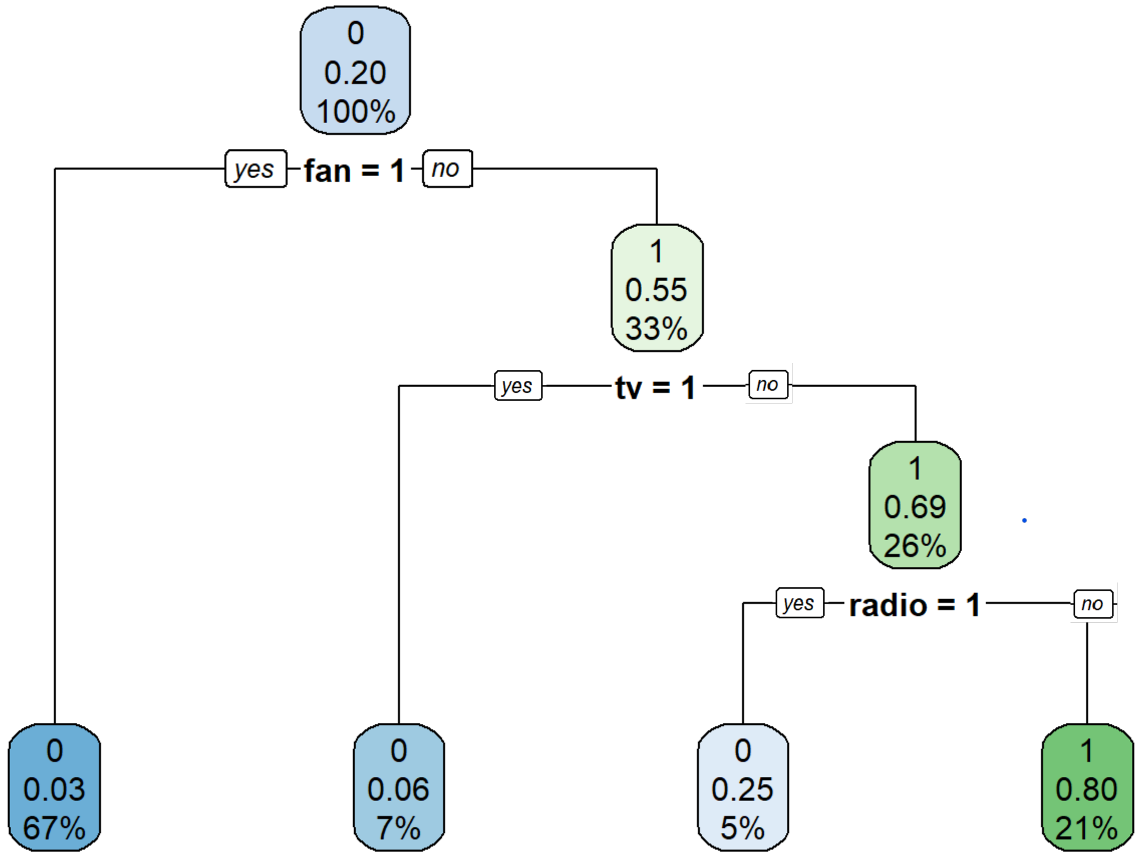
- Each blue or green box is a node of the tree.
- All households in blue nodes are classified '0' (non-poor).
- All households in green nodes are classified '1' (i.e. poor)

- Each node has 3 numbers: x, y, z. E.g.

x	---------->	households in this node are classified x
y		the proportion of 1's in this node is y
z		the proportion of the entire sample in this node: z

- The darker the shade (of blue or green) of each node, the higher the concentration of 'x' households in this node.

Table 3: Classification Tree (T2)
Outcome: If relatively poor
Predictors: If owns individual durables (radio, tv, fan, ac, fridge, bike, motor bike, car)



KEY:

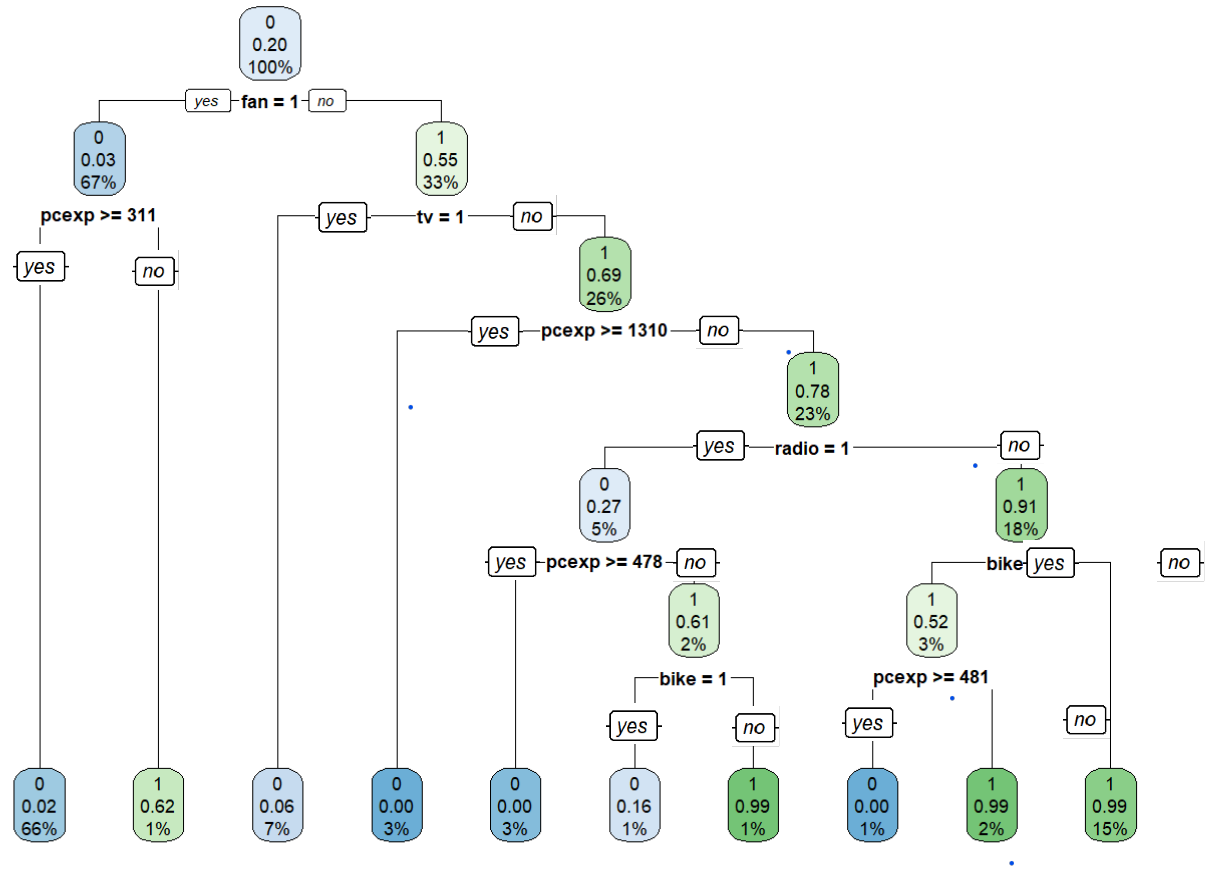
- Each blue or green box is a node of the tree.
- All households in blue nodes are classified '0' (non-poor).
- All households in green nodes are classified '1' (i.e. poor)

- Each node has 3 numbers: x, y, z. E.g.

x	---------->	households in this node are classified x
y		the proportion of 1's in this node is y
z		the proportion of the entire sample in this node: z

- The darker the shade (of blue or green) of each node, the higher the concentration of 'x' households in this node.

Table 4: Classification Tree (T3)
Outcome: If relatively poor
Predictors: PCE; If owns individual durables (radio, tv, fan, ac, fridge, bike, motor bike, car)



KEY:

- Each blue or green box is a node of the tree.
- All households in blue nodes are classified '0' (non-poor).
- All households in green nodes are classified '1' (i.e. poor)

- Each node has 3 numbers: x, y, z. E.g.

x y z	---------->	households in this node are classified x the proportion of 1's in this node is y the proportion of the entire sample in this node: z

- The darker the shade (of blue or green) of each node, the higher the concentration of 'x' households in this node.

Table 5: Relative-Poverty Schedule (RPS) predicted by T3
Indian (urban) NSS, 1999-00

Total Durables (Y)	Owns Good? (1: Yes, 0: No)				RPS
	Fan	TV	Radio	Bike	PCE-Cutoff
0	0	0	0	0	1310
1	1	0	0	0	311
	0	1	0	0	0
	0	0	1	0	478
	0	0	0	1	481
2	1	1	0	0	311
	1	0	1	0	311
	1	0	0	1	311
	0	1	1	0	0
	0	1	0	1	0
	0	0	1	1	0
3	1	1	1	0	311
	1	1	0	1	311
	1	0	1	1	311
	0	1	1	1	0
4	1	1	1	1	311

Data Appendix: 'rpart' Complexity Tables[1]

Tree T1:

Predicted variable Y --> Opportunities Measure of Relative Poverty
Predictor variable X1 --> per capita monthly household expenditure (PCE)

```
rpart(formula = Y ~ X1, data = dataset, method = "class")

Variables actually used in tree construction:
[1] PCE

Root node error: 9766/48921 = 0.19963

n= 48921

     CP      nsplit rel error xerror  xstd
1 0.115093      0   1.00000 1.00000 0.0090529
2 0.033586      1   0.88491 0.88501 0.0086378
3 0.017510      2   0.85132 0.85306 0.0085132
4 0.010000      3   0.83381 0.83791 0.0084526
```

Tree T2:

Predicted variable Y --> Opportunities Measure of Relative Poverty
Predictor variables X2 --> whether household owns radio, tv, fan, ac, fridge, bike, motor bike, car

```
rpart(formula = Y ~ X2, data = dataset, method = "class")

Variables actually used in tree construction:
[1] fan   radio tv

Root node error: 9766/48921 = 0.19963

n= 48921

     CP      nsplit rel error xerror  xstd
1 0.24713      0   1.00000 1.00000 0.0090529
2 0.12656      2   0.50573 0.50573 0.0068233
3 0.01000      3   0.37917 0.37917 0.0059906
```

[1] Therneau et al, 2023.

Data Appendix: 'rpart' Complexity Tables[1]

Tree T3:

Predicted variable Y --> Opportunities Measure of Relative Poverty

T3: Predictor variables X3 --> PCE, whether household owns radio, tv, fan, ac, fridge, bike, motor bike, car

```
rpart(formula = Y ~ X2, data = dataset, method = "class")

Variables actually used in tree construction:
[1] bike  fan   pcexp radio tv

Root node error: 9766/48921 = 0.19963

n= 48921

     CP      nsplit rel error xerror  xstd
1 0.247133      0   1.00000 1.00000 0.0090529
2 0.150317      2   0.50573 0.50573 0.0068233
3 0.104034      3   0.35542 0.35572 0.0058171
4 0.035890      4   0.25138 0.25169 0.0049474
5 0.026572      6   0.17960 0.17981 0.0042132
6 0.012390      8   0.12646 0.12687 0.0035583
7 0.010000      9   0.11407 0.11694 0.0034197
```

[1] Therneau et al, 2023.