Applying an Unsupervised Machine Learning Approach to Analyze the Non-Income Poverty Indicators Used in the *Listahanan 2*

Dominique Ysobel P. Ting

Abstract

Poverty targeting has been used in developing countries as a means to provide social protection programs and services directly to the poor. As information on households' welfare is unavailable or difficult and costly to acquire in the developing world, the proxy means test (PMT), which uses proxy variables to estimate an unobservable welfare variable such as household income or consumption, has become a commonly used method for targeting. This study uses an unsupervised machine learning approach on the set of household- and individual-specific non-income poverty indicators used to estimate household income in the PMT models for the Philippines' National Household Targeting System for Poverty Reduction or *Listahanan, in order* to examine whether differences between households across these indicators reflect differences in their income. Applying the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm onto the *Listahanan 2* indicators shows that households naturally cluster into three to four groups. However, these clusters seem to be unrelated to income and expenditure. The richest and poorest households appear to be alike and cannot be differentiated on the basis of the non-income poverty indicators considered. This suggests that these indicators alone may not be sufficient for the PMT models to accurately target the poor. However, this study is a preliminary analysis on the limited data available. A more comprehensive analysis is required to produce conclusive results.

I. Introduction

Policy interventions specifically targeting the poor have been used in developing countries since the 1980s for poverty alleviation. These were initially developed due to concerns regarding social welfare as well as limited government resources and fiscal constraints (Weiss 2004; Weiss 2005; Lavallée et al. 2010). Since then, the focus on global poverty eradication and social protection has greatly intensified. According to a 2018 report from the World Bank, social safety nets or social assistance programs, which target the poor and vulnerable, have covered 18% of the poorest quintile in low-income countries and 43% in lower-middle-income countries. Despite progress from past years, these remain far from 76% coverage of the poorest quintile in high-income countries (The World Bank 2018, 35). With such programs being crucial in helping people escape poverty, it is necessary to further increase coverage in the developing world.

As part of its efforts to improve its social protection services, the Philippines adopted the National Household Targeting System for Poverty Reduction, more commonly known as *Listahanan*, in 2010. The *Listahanan* is a Proxy Means Test (PMT)-based targeting system that enables identifying the poor and allows for the creation of a registry of poor Filipino households (Velarde 2018). The PMT is among the most popular and widely-used targeting methods utilized to ensure that social assistance programs are reaching the target population in developing countries. It involves producing a score to estimate household welfare, usually in terms of income or consumption, using household characteristics that are highly correlated with poverty, are easily observable and measurable, and are difficult to manipulate (World Bank, n.d.; Lavallée et al. 2010; Coady, Grosh, and Hoddinott 2004). The results from PMT models determine who is identified as poor in the *Listahanan* database, which is then used for several programs, including the Pantawid

Pamilyang Pilipino Program (4Ps), the Philhealth Indigent Program, Sustainable Livelihood Program, and the Social Pension for Indigent Senior Citizens Program (DSWD, n.d.).

As the largest social protection programs in the country are based on the *Listahanan*, it is important that the targeting system accurately identifies the poor. Hence, it is crucial to evaluate the system and the PMT models used. However, in contrast to numerous research and impact evaluation studies conducted on programs that utilize the *Listahanan*, particularly the 4Ps, there have been limited studies evaluating the targeting system itself. There is extensive literature on PMT models, but the results of these studies are not necessarily generalizable to all PMT models due to differences in country contexts and given that model specifications vary depending on the available data. That is, the construction of a PMT model relies on existing datasets with information on household characteristics and welfare, so the variables to be included in each model differ based on the data source. Thus, this paper aims to contribute to the literature on assessing the *Listahanan* by examining the input of the PMT models used for the targeting system.

The use of a PMT model assumes that some non-income indicators can be used as proxies for estimating income. As such, this study analyzes the indicators used to identify poor households in the *Listahanan* 2, which resulted from the second round of assessments that were completed in 2016, using an unsupervised machine learning approach. This method allows us to explore natural groupings of households in the data and observe whether patterns are related to other relevant indicators that are not included in our unsupervised learning model. In this paper, we are specifically interested in determining whether there are patterns of households grouping according to their income based on non-income poverty indicators used in the PMT models. Since these models are used to estimate income to identify poor households, we may expect that households that are more similar across the variables in the models have closer values of income. Furthermore, households with the highest income should be clearly separated from households with the lowest incomes according to the non-income poverty indicators.

Ensuring that indicators used in the PMT models can be used to differentiate poor from non-poor households is necessary for the targeting system to be effective. This is especially crucial as the *Listahanan* is updated only every four years, which means that the specifications of the PMT models are based on data from a couple of years prior and are not updated until the next round of assessment. For instance, the PMT models for the first *Listahanan*, which was released in 2011, were based on the merged 2003 Family Income and Expenditure Survey and Labor Force Survey and used variables from these household surveys. Examining the same variables from more regularly conducted national household surveys can provide more timely insights on how well the non-income poverty indicators used in the PMT models reflect differences in income.

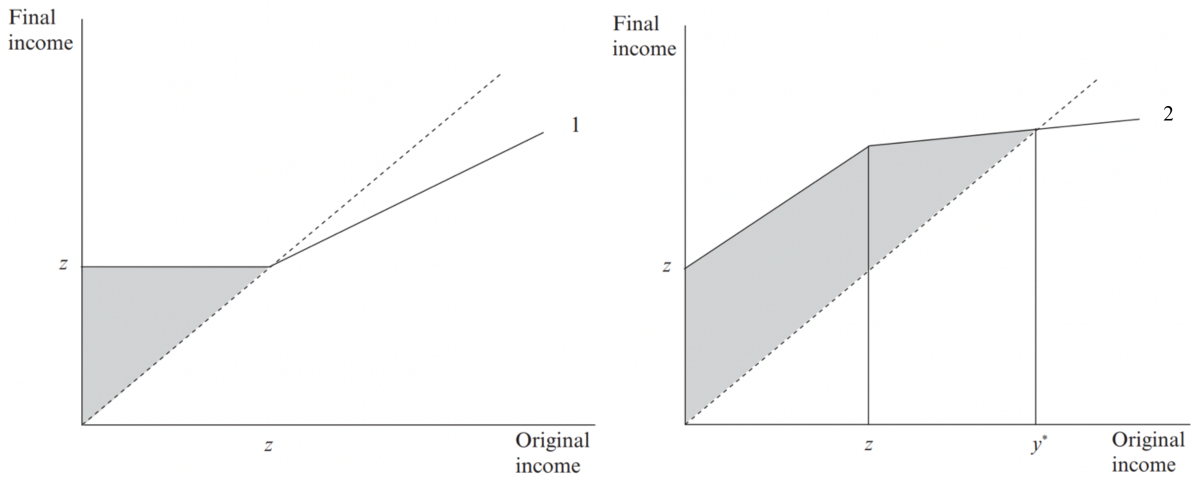
II. Background and Literature Review

Poverty targeting is defined by Weiss (2004) as “the use of policy instruments to channel resources to a target group identified below an agreed national poverty line.” This targeting approach has been an attractive strategy for developing countries facing budget constraints. Factors taken into account in targeting can be understood by considering the framework presented by Besley and Kanbur (1990), which compares poverty alleviation approaches at two extremes: an ideal solution of “perfect targeting,” wherein everyone below the poverty line is provided a subsidy equal to the difference between the poverty line and their income, and a universalistic scheme, wherein everyone is provided a transfer whether or not they are below the poverty line.

Visualizations for the two extreme approaches are shown in Panels A and B of Figure 1 from Weiss (2005), which follows the framework of Besley and Kanbur. For both graphs, original income is on the x-axis, final income (i.e., post transfer) is on the y-axis, z represents the poverty line (i.e., below z indicates poverty), and the dashed 45-degree line represent cases in which original income and final income are equal (i.e., no transfers). Points above the 45-degree line represent a subsidy, while points under the line represent a tax. Under perfect targeting, the assumption is that income is perfectly observable at no cost. The government provides subsidies to everyone below the poverty line such that their income y reaches z and these are financed by taxes on everyone above the poverty line. The resulting distribution of income is shown by line 1 in Panel A. The cost of this approach is the sum of all z-y transfers as depicted by the shaded area between line 1 and the 45-degree line.

In contrast, under a universal scheme, everyone is provided a transfer with an amount equal to z. The transfers are financed by taxing everyone above the poverty line. Thus, the non-poor receive transfers equivalent to z minus taxes. At some level of income y^*, taxes exceed subsidies resulting in peoples' final income being lower than their original income. The resulting distribution of income is shown by line 2 in Panel B. The cost of this approach is equal to z multiplied by the size of the population, which is indicated by the shaded area in the graph. In comparison to the perfect targeting approach, the cost is much higher for the universal scheme. Moreover, there is leakage to people above the poverty line who have income between z and y^*. Considering higher costs and leakage, the perfect targeting approach is preferable to the universal scheme (Weiss 2005).

Figure 1. Perfect Targeting and a Universal Scheme



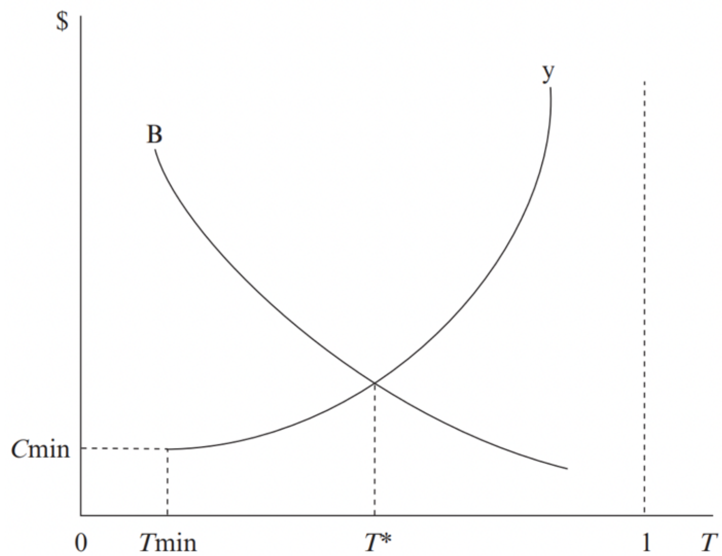
A. Perfect Targeting B. Universal Scheme

Source: Weiss (2005)

Beyond these theoretical conditions, Besley and Kanbur (1990) outline other considerations that need to be made in real-world situations. First, there are administrative costs involved in carrying out poverty alleviation programs, including a minimum cost associated with the operationalization of a program and the additional costs of verifying income when using a targeting approach. The authors breakdown revenue into a combination of three categories: administrative costs, transfers to the non-poor (i.e., leakages), and transfers to the poor. They surmise that the administrative costs as a proportion of revenue rise with the fineness of targeting, which is the ratio of transfers to the poor and non-administrative costs (i.e., sum of transfers to the poor and non-poor). They also state that a minimum level of targeting will always be achieved as some benefits, even in the opposite extreme of the universal scheme, will be received by some of the poor population.

Weiss (2005) shows the optimal degree of targeting considering these costs in Figure 2. The fineness or degree of targeting, which is defined by the share of benefits directed to the poor, is on the y-axis, while the monetary value of costs and benefits received by the poor is on the x-axis. T_{min} is the minimum level of targeting that is always attained and C_{min} is the minimum cost of operationalization. Line *y* represents that positive relationship between the degree of targeting and administrative costs as a proportion of revenue. On the other hand, line *B* represents the marginal social benefit of an extra dollar directed to the poor, which is assumed to be positive but declining with a higher degree of targeting. Considering these costs and benefits, there should be an optimal level of targeting, as shown by the intersection of lines *B* and *y* at *T**.

Figure 2. Optimal Targeting



Source: Weiss (2005)

Still, in addition to costs to government, there are also factors to be considered on the part of potential participants. These include costs incurred in being subject to assessments as well as psychic costs of social stigma that may dissuade them from participating in a targeted program. Besley (1990) theorizes that if individuals have some cost *c* associated with the targeted program, then those who have income higher than *z-c* (i.e., the difference between the poverty line and cost), will not participate in the program and will remain below the poverty line. Furthermore, under perfect targeting, individuals below the poverty line are disincentivized to work and earn more income since they are provided a subsidy equal to how much they fall below the poverty line. Finally, there are also considerations in terms of political economy. There may not be enough political power to support perfect targeting, which is rationally only favored by those below the poverty line. In contrast, there may be more support for the universal scheme, which is beneficial to the "middle class" whose income is between z and $y*$ who may have more political influence (Besley and Kanbur 1990; Weiss 2005).

Although perfect targeting may be the "ideal" solution, several considerations suggest that this approach may not be feasible. Instead, an optimal targeting approach between the two extremes may be effective for poverty alleviation. In this context, there are a range of different methods for poverty targeting. Generally, targeting is carried out through either "broad targeting," in which activities or sectors that benefit the poor the most are targeted (e.g., universal primary health care and education), or through identifying the poor and delivering resources directly and exclusively to them (Lavallée et al. 2010; Weiss 2004; Coady, Grosh, and Hoddinott 2004). Considering restraints in funding, the latter approach of focusing resources on the poor is beneficial as a means to maximize the impact of programs given a limited budget or to achieve a certain amount of impact with minimal cost (Coady et al. 2004).

The implementation of a narrow poverty targeting approach, however, is less straightforward than a broad targeting approach. A crucial factor in such approaches is determining who belongs to the target group. In developed countries, income can be used as a measure to target the poor as it is reported through the tax system. In contrast, a large part of the population does not pay taxes in developing countries; hence, governments need to use alternative methods to identify the poor (Banerjee et al. 2020; Hanna and Olken 2018). When information on income is not available, other means of targeting can include: targeting by indicator, in which indicators correlated with income are used; targeting by location, in which area of residence is used; and targeting by self-selection, in which programs are specifically made to appeal to only the poor (Weiss 2004).

Proxy Means Test

The Proxy Means Test (PMT) is a commonly used method of targeting by indicator. As mentioned earlier, the PMT estimates a score to measure household welfare using household characteristics that are highly correlated with poverty, are easily observable and measurable, and are difficult to manipulate. PMT models are typically developed using existing data sources, such as household income and expenditure surveys, with information on household income or consumption (i.e., to indicate welfare) and household characteristics. Some indicators that are usually included in the PMT are a household's geographic location, housing quality, occupancy status, ownership of durable goods, demographic structure, labor force status, occupation or sector of work, and educational attainment. A statistical analysis is performed on the chosen indicators to determine weights to be assigned for each variable. Potential members of the target population are then surveyed to collect their information on the indicators and estimate their score based on the specifications of the statistical model. The resulting score is then used to determine whether the household qualifies as a beneficiary for the targeted program (Coady, Grosh, and Hoddinott 2004; Lavallée et al. 2010; World Bank, n.d.).

A range of studies have assessed the use of PMT models in various developing countries and have produced mixed results. A 2011 study by the Australian Agency for International Development examined the PMT in Bangladesh, Indonesia, Rwanda, and Sri Lanka and found high in-built errors, particularly at low levels of coverage of less than or equal to 20% of the population. According to the study, the PMT selects beneficiaries arbitrarily due to imperfect correlation between proxy variables and consumption, untimely and inaccurate representations of reality, and errors in surveys and assumptions. The paper argues that schemes that do not directly target the poor may have better performance. In contrast, Grosh and Baker (1995) assert that although proxy systems have significant undercoverage errors, they reduce leakage substantially such that imperfect targeting has a larger impact on reducing poverty than using no targeting at all. Similarly, using evidence from Indonesia and Peru, Hanna and Olken (2018) show that targeted programs provide much larger welfare gains to the poor than universal programs.

With respect to targeting performance in comparison to other targeting methods, the literature generally indicates that the PMT performs just as well as or only somewhat better (i.e., usually for poorer households) than other methods (Coady and Parker 2009; Alatas et al. 2012; Karlan and Thuysbaert 2019; Premand and Schnitzer 2021). In fact, a core finding by Coady, Grosh, and Hoddinott (2004), based on their analysis of their exhaustive database of targeted programs for the poor, is that "there is no clearly preferred method for all types of programs or all country contexts." They noted that 80% of the variability they observed were within, rather than across, targeting methods. Moreover, some of the variability was related to country context; countries with higher income, had more government accountability, and had greater inequality had better targeting performance (Coady, Grosh, and Hoddinott 2004).

Targeting in the Philippines

In March 2010, through Executive Order No. 867, the Philippine government adopted the National Household Targeting System for Poverty Reduction or *Listahanan* as its main system for identifying poor households and mandated that all national government agencies must use the system for their social protection programs and services. The *Listahanan* consists of a database with information on families who are classified by the Department of Social Welfare and Development (DSWD) as poor through proxy means testing. The department constructs the PMT model to estimate household income using variables from official surveys conducted by the Philippine Statistics Authority, such as the Family Income and Expenditure Survey (FIES), the Labor Force Survey (LFS), and the Census of Population and Housing (CPH). The model uses observable and verifiable indicators of household characteristics such as households' housing construction materials, access to water and electricity, and ownership of some specific assets. Data collected from households using a Household Assessment Form are processed using the PMT model to estimate income. These estimates are then compared to official poverty thresholds at the provincial level to determine whether a household is poor. Households falling below the threshold are considered poor, while those above are considered non-poor (Velarde 2018; Department of Social Welfare and Development, n.d.a).

The *Listahanan* is updated every four years, allowing for the enhancement of the PMT models based on a review of the model accounting for more recent data. The development of the first PMT models began in 2007 and was led by the World Bank alongside local academics. The models included household- and individual-level indicators and used the 2003 FIES and LFS datasets as reference data. The first round of assessments to create the first *Listahanan* database was completed in 2011. Following this, a review of the first PMT models and the development of the Second PMT models began in 2012 led by local academics with inputs and guidance from the World Bank. The models improved upon the first models, e.g. by increasing the number of correlates, including community-level indicators. The reference data used was also updated to the 2009 FIES-LFS dataset and included the 2007 CPH as well. Unlike the first models, which used urban and rural areas for its sub-models, the second models used sub-models for the National

Capital Region and the rest of the Philippines (Velarde 2018; Department of Social Welfare and Development, n.d.b). A comparison of the PMT models for the first two *Listahanan* are summarized in Table 1. A third round of assessment to update the *Listahanan*, which was delayed due to the COVID-19 pandemic, was aimed to be completed by the last quarter of 2021 (Department of Social Welfare and Development, n.d.c). However, it should be noted that the *Listahanan 3* had not been available for some regions in the first few months of 2022 (Saavedra 2022; Petinglay 2022).

Table 1. Comparison of PMT Models

	***Listahanan* 1**	***Listahanan* 2**
Explanatory variables	Household-level variables from the Labor Force Survey (LFS) and Family Income and Expenditure Survey (FIES); Aggregate occupations used in the model based on 2-digit occupational codes	Household-level variables from the LFS and FIES + community-level variables from the Census of Population; More detailed occupations used based on 4-digit occupational codes
Reference data	2003 FIES-LFS	2009 FIES-LFS; 2007 CPH
Sub-models	1 Model for Urban areas 1 Model for Rural areas	1 Model for NCR 1 Model for the Rest of the Philippines (ROP)
Layers	1 layer for both Urban and Rural models to predict per capita income of households and balance the exclusion and inclusion errors	2 layers for both NCR and ROP models: Layer 1 – predicts per capita income of households and minimizes exclusion error; Layer 2 – predicts misclassification of real non-poor households as poor to minimizes inclusion error
Reference population to estimate the PMT	All poor households in the official household surveys (LFS and FIES)	Bottom 40 population in the official household surveys (LFS and FIES)
Basis for identifying poor households	Point estimate of the predicted per capita income versus the official poverty threshold	Lower bound of the 95% predicted interval of per capita income versus the official poverty threshold

Source: Velarde (2018)

There have been few studies evaluating poverty targeting and, more specifically, the use of the PMT in the Philippines. A comprehensive analysis of different targeting methods used by the government for various programs prior to the adoption of the *Listahanan* are provided by Balisacan and Edillon (Weiss 2005, 227-243). With respect to proxy means testing, Velarde (2018) reports on the development of the *Listahanan* targeting system and cites a number of studies providing assessments of the PMT models for the system. However, these reports are internal documents of the DSWD; hence, are not publicly accessible. External studies on the *Listahanan* are very limited, likely due to the fact that detailed information on the PMT models is kept confidential. That is, as stipulated in Memorandum Circular 12, Series of 2017, sharing the formula of the PMT is not permitted. Nevertheless, there is one recent study evaluating targeting in the country, specifically its use in the Pantawid Pamilyang Pilipino Program (4Ps) and *Listahanan*, by Dadap-Cantal, Fischer, and Ramos (2021). The authors use extensive document analysis to examine the Philippines' targeting system and argue that, despite the system receiving wide recognition for positive poverty outcomes, it has not been able to properly identify the poor and provide them social protection. This was primarily attributed to an outdated social registry (Dadap-Cantal, Fischer, and Ramos 2021).

III. Data and Methodology

For this study, we are interested in assessing the *Listahanan 2*, which is the most recently released registry as of the first half of 2022.[1] The PMT models for the *Listahanan 2* are based on the 2009 merged Family Incomes and Expenditure Survey (FIES) and Labor Force Survey (LFS) and the 2007 Census of Population and Housing (CPH). We focus our analysis only on the household- and individual-specific variables that are based on the FIES-LFS, rather than the community-specific variables that are based on the census. This is because geographic locations of households at the community level are not included in the FIES-LFS public use file that is provided upon request by researchers due to the sensitive nature of the data. Hence, even with the

[1] The *Listahanan* 3 is in the Validation and Finalization Phase as of July 2022. The DSWD expects to release the *Listahanan* 3 database in the third quarter of the year (Department of Social Welfare and Development, eFOI request, July 5, 2022).

available CPH microdata, it is not possible to match households from the FIES-LFS to community-specific information in the CPH. Although this will not provide a complete picture of the structure of the data, we still expect this to be a close approximation.

The first dataset we will use for our analysis is the reference data used for the second PMT models, the 2009 FIES-LFS dataset. Following this, we also analyze the data for 2016 and 2017 to explore possible changing patterns over time. Since the FIES is conducted only every three years, the data is not available for 2016 and 2017. Instead, we use the Annual Poverty Indicators Survey (APIS) to examine the data for these two years. The APIS is an annual nationwide survey conducted by the PSA to measure the socioeconomic profile and living conditions of Filipinos (Philippine Statistics Authority, n.d.). The survey includes most of the non-income poverty variables measured in the FIES and LFS; hence, a similar analysis can be applied to the data.

As the specific variables used in the PMT models are kept confidential, we rely on previous studies as well as the Household Assessment Form (HAF)—the questionnaire used to collect data from households—to determine the non-income poverty indicators to include in our analysis. In particular, we use all the household- and individual-specific indicators identified by Velarde (2018), along with other items that are not included in Velarde's report but are in the HAF (e.g., number of air conditioners the household owns). Some variables that need to be transformed are based on those used by Mapa and Albis (2013) in their proposed enhancement for the second PMT models. The full set of non-income poverty indicators used for the second PMT models as well as those used in this study based on the 2009 FIES-LFS and the 2016 and 2017 APIS are listed in Table 2 below.

Table 2. Non-Income Poverty Indicators

Second PMT-models	2009 Family Income and Expenditure Survey - Labor Force Survey (FIES-LFS)	2016 and 2017 Annual Poverty Indicators Survey (APIS)
Barangay-level Indicators		
Presence of town city hall/ provincial capitol in the Barangay		
Presence of high school		
Presence of street patterns		
Number of recreational establishments		
Number of commercial establishments		
Number of hotel dormitory, motel or other lodging places in the barangay		
Number of establishments offering personal services like restaurants, cafeteria, etc.		
Share of population 10 yrs old and above who are farmers, farm laborers, fishermen, loggers, and forest product gatherers (>50%)		
Number of auto repair shop, vulcanizing shop, electronic repair shop, or other repair shops		
Poblacion/ City District indicator		
Presence of cemetery		
Availability of landline telephone system or calling station		
Availability of cellular phone signal		
Number of banking institutions/ pawnshops financing and investment		
Number of recreational establishments OUTSIDE the barangay but within 2 kms		
Number of households dwelling in private land which they do not own except in danger areas		

Household-specific Indicators		
Ownership of assets		
Ownership of house and lot	Hhld Tenure Status	Tenure status of the housing unit and lot occupied by the family
Number of the following appliances owned:		
Refrigerator/s	Hhld Number of refrigerator	Number of refrigerator/freezer the family own
Washing Machine/s	Hhld Number of washing machine	Number of washing machine the family own
Telephone/s or cellphone/s	Hhld telephone	Number of cellular phone the family own; Number of landline/wireless telephone the family own
TV set/s	Hhld Number of TVs	Number of television the family own
Radio/s	Hhld Number of radios	Number of radio/radio cassette player
VTR/ VHS/ VCD/ DVD	Hhld Number of VCRs	Number of CD/DVD/DVD Player the family own
Stereo or CD player/s	Hhld Number of stereos	Number of audio component/stereo set the family own
Microwave oven/s	Hhld Number of ovens	Number of stove with oven/ gas range the family own
Sala set/s[2]	Hhld Number of sala sets	
Dining set/s	Hhld Number of dining sets	
Airconditioner/s	Hhld Number of aircons	Number of aircon the family own
Computer/s	Hhld Number of Microcomputer	Number of personal computer the family own
Housing conditions		
Number of the following vehicles owned:		
Car/jeep	Hhld Number of vehicle	Number of car, jeep, van
Motorcycle/tricycle	Hhld Number of motorcycles	Number of motorcycle, tricycle
Make of roof	Hhld House Type of Roof	Type of construction materials of the roof
Make of walls	Hhld House Type of Wall	Type of construction materials of the outer wall
Building type	Hhld House Building type	Type of building/house the family reside
Access to services		
Main source of water supply	HHld Main source of water	Family's main source of water supply
Type of toilet facility	Hhld Toilet facility	Kind of toilet facility the family use
Access to electricity	Hhld availability of electricity indicator	Presence electricity in the building/house

[2] A sala set refers to living room furniture, especially a matching set of a sofa and chairs.

Other HH Characteristics		
Household type	Hhld type	
Number of HHS in housing unit	Number of Households in the Housing Unit	
Agricultural household	Agricultural Household indicator	
Availability of domestic help	Relationship to Household Head: domestic helper	
Regional location	Region	Region
Urban location	Urban/ Rural	Urban/ Rural*
Individual-specific Indicators		
Marital status of the HH Head	HH head Marital status	Head: Marital Status
Gender of the HH Head	HH head Sex	Head: Sex
Number of family members (family size)	Family Size	Family Size
Age of family members	Age as of last birthday	Age as of last birthday
Education of family members		
Highest grade completed	Highest grade completed	Highest grade completed
Currently attending school	Currently attending school	
Occupation of working family members:		
Worked	Did work or had a job during the past quarter	Did work or had a job or business anytime from January 1 to June 30, 2014* Did work or had a job or business anytime from January 1 to June 30, 2017**
Primary Occupation	Primary Occupation	
Class of worker	Class of worker	Class of worker
Nature of Employment	Nature of Employment	
Basis of payment	Basis of payment	
Overseas Filipino	Overseas Filipino Indicator	

** Only available in the 2016 APIS*
*** Only available in the 2017 APIS*

Ideally, a supervised learning method simulating the actual PMT models should be used to assess how well the targeting system correctly distinguishes between poor and non-poor households through estimating the models' inclusion (i.e., non-poor households classified as poor) and exclusion (i.e., poor households classified as non-poor) errors. However, to conduct such an analysis, a complete dataset and information on the PMT models' specifications, including the formula used and coefficients for each variable, are required. While it is possible to create models estimating income that may be similar to the second PMT models using the limited data and information available, the estimates resulting from these will not necessarily replicate those of the actual PMT models used. Hence, we do not attempt to estimate the errors resulting from the second PMT models. Instead, we take an alternative approach that does not rely on model specifications to assess the PMT models used in the targeting system.

Using the FIES-LFS and APIS datasets, we assess the *Listahanan 2* household- and individual-specific non-income poverty indicators to examine whether there are differences across these indicators among poor and non-poor households. Since these indicators are used in the PMT models to estimate a household's income, differences across these indicators should reflect differences across income levels. To test this, we use an unsupervised machine learning method to find natural groupings of households based on the *Listahanan 2* indicators. This approach allows us to examine the underlying structure of the data without providing labels on how the data should be classified. For our analysis, since we know that households are being classified as poor and non-poor based on the non-income poverty indicators, we may expect the data to show patterns of clustering according to household income. In particular, poorer households may appear similar to each other and different from richer households. Conversely, richer households may appear similar to each other and different from poorer households.

We apply the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm to learn the manifold of the datasets of households with the *Listahanan 2* indicators, project this into a lower dimensional space, and visualize this projection. UMAP is a non-linear dimension-reduction technique that assumes data is distributed along an n-dimensional smooth geometric shape (i.e., manifold) along which distances can be computed and represented into a lower dimension. This algorithm has a number of advantages over other dimension-reduction methods including being able to learn nonlinear patterns, more clearly separating clusters of cases, and preserving both local and global distances (Rhys 2020, 337-343). Using this algorithm allows us to better understand the structure of the dataset and observe whether there patterns of divisions between poor and non-poor households.

Following the construction of the PMT models for the *Listahanan 2*, we only consider the bottom 40% of households (i.e., households in the first four income deciles) for each of our datasets for our analysis. We also train separate models for the full dataset and a dataset including only households residing in areas outside the National Capital Region (AONCR) to emulate the use of separate models for NCR and AONCR for the *Listahanan 2*. Furthermore, as UMAP only takes numeric variables, we transform all categorical variables into numeric variables. After pre-processing the data, we train a UMAP model on each of our datasets using varying values for different hyperparameters (i.e., number of neighbors, minimum distance, and distance metric). We choose a final model for each and use these to examine patterns in more detail. We begin our analysis on the 2009 FIES-LFS dataset to first gain insight into the data the government used to develop the second PMT models. To better understand the environment when the targeting system had been implemented, we analyze the 2016 and 2017 APIS datasets. Since the APIS datasets do not use exactly the same variables as the FIES-LFS, we train different UMAP models for each.

We note that this means we cannot directly compare the results; nevertheless, this provides an examination of how patterns may have changed in the following years.

IV. Results

The first UMAP model is trained using data on 16,651 households and 83 variables (see Table 3 in the appendix) from the 2009 FIES-LFS dataset. The variables include all household- and individual- specific indicators we consider to be included in the second PMT models as discussed in the previous section. Figure 3 below shows the embeddings for the UMAP model with varying values for the number of neighbors and the minimum distance while using a Euclidean metric and 500 epochs. Examining the results below, it appears that the households can generally be grouped into four clusters. Using a Manhattan metric also shows similar results as displayed in Figure 4.

To further explore the patterns observed, the final model using a Euclidean metric, 25 neighbors, 0.1 minimum distance, and 500 epochs is presented in Figure 5. The plots in the figure are colored according to variables related to income and expenditure to assess whether these may explain the natural groupings in the data. Based on the final UMAP embeddings, the four clusters observed do not appear to be related to households' income decile, total and per capita income, and total and per capita expenditure. While households with the highest per capita income and per capita expenditure tend to be located at the lower sections of the lower two clusters (see panels e and f in Figure 5), they remain closely grouped together with other households. The plots illustrate that there is no clear separation between households of varying income and expenditure levels in the bottom 40% of the population based on the non-income poverty indicators considered in the analysis. In addition, using the same method for the 2009 dataset but only for households residing outside the National Capital Region (AONCR) produces nearly identical results (see Figures

Figure 12, Figure 13, and Figure 14 in the appendix). This is expected as households residing in NCR account for less than 2% of the data (i.e., a total of only 298 households).

Figure 3. 2009 UMAP embeddings with varying number of neighbors (rows) and minimum distance (columns)

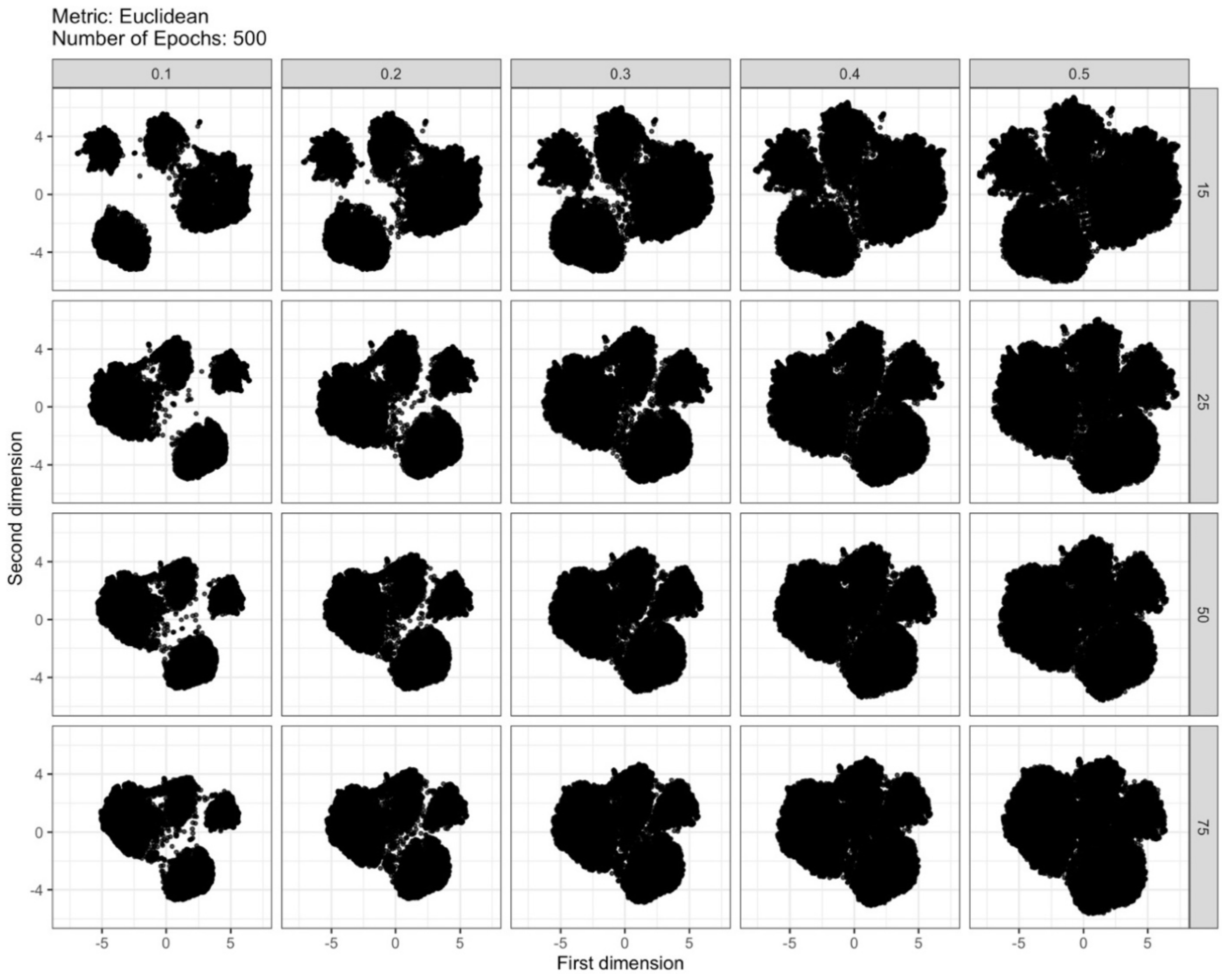


Figure 4. 2009 UMAP embeddings with different metrics (rows and varying minimum distance (columns)

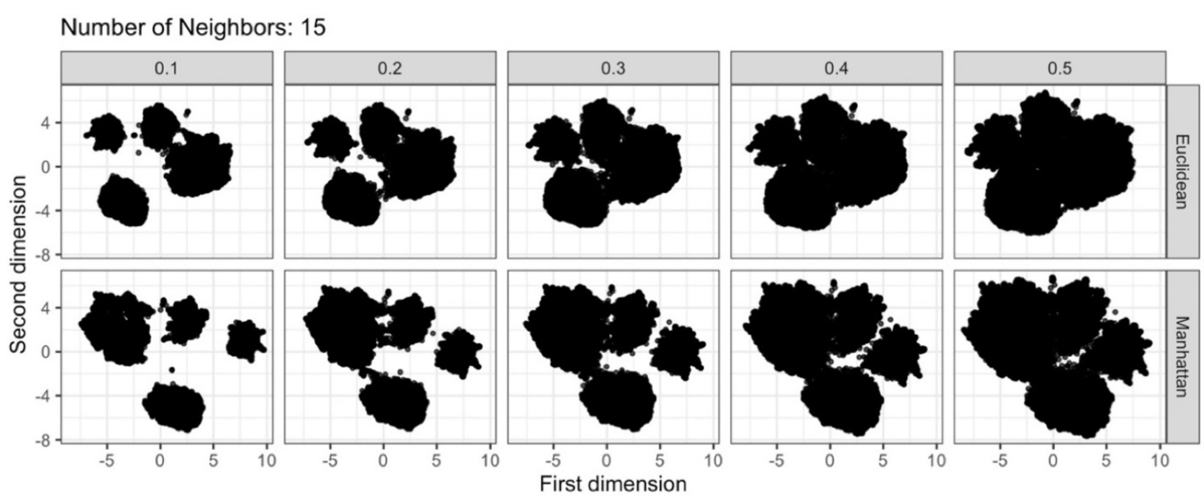
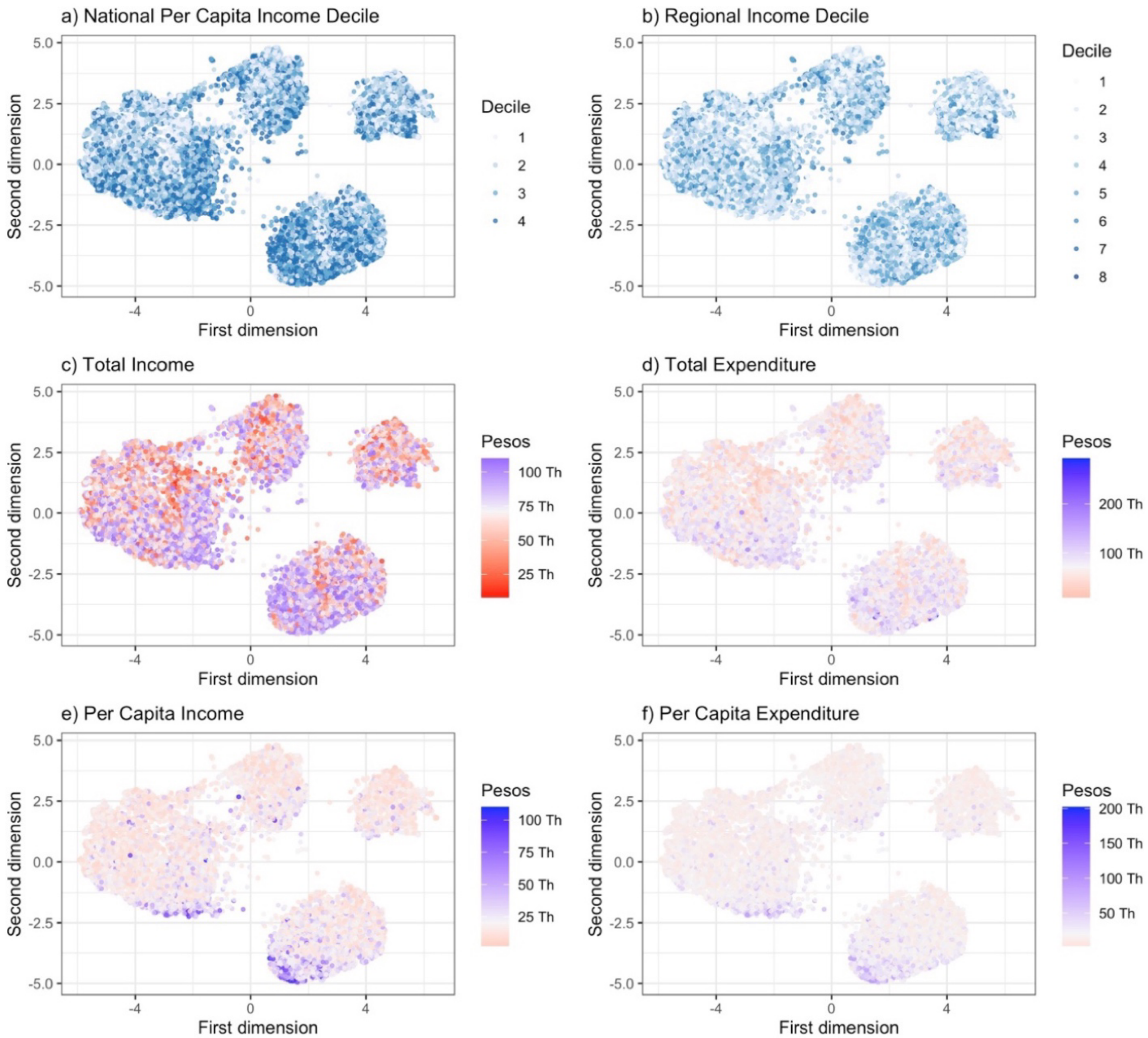


Figure 5. Final 2009 UMAP embeddings using a Euclidean metric, 25 neighbors, 0.1 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.

For the 2016 model, UMAP is trained on 40 variables (see Table 4 in the appendix) and 4,346 observations from the 2016 APIS dataset. The number of variables included in this model is considerably lower than the previous model, primarily because the APIS does not have detailed information on the occupation of working household members; the survey only has data on the class of worker. The number of observations is also lower as the APIS is conducted on a much smaller sample of households since the survey is done annually. The UMAP embeddings for the 2016 data with various hyperparameters are shown in Figures Figure 6 and Figure 7 below. In contrast to the results of the 2009 UMAP model, the clusters in the 2016 UMAP model are less discernable. Nonetheless, there seem to be about three to four larger clusters.

Figure 8 presents the results of the final UMAP embeddings using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs. The plots are colored according to income- and expenditure-related variables, specifically, national per capita income decile, total income, total expenditure, per capita income, and per capita expenditure. Similar to the 2009 results, the clusters do not appear to be related to income and expenditure. Instead, richer and poorer households are mixed together in the various groups, indicating that the bottom 40% of households cannot be clearly separated according to their income or expenditure based only on the non-poverty income indicators included in the model. The results for the UMAP model applied to only households in AONCR are also very similar as only about 3.6% of households (i.e., a total of only 158 households) who live in NCR are excluded from the model. These results are presented in Figures Figure 15, Figure 16, and Figure 17 in the appendix.

Figure 6. 2016 UMAP embeddings with varying number of neighbors (rows) and minimum distance (columns)

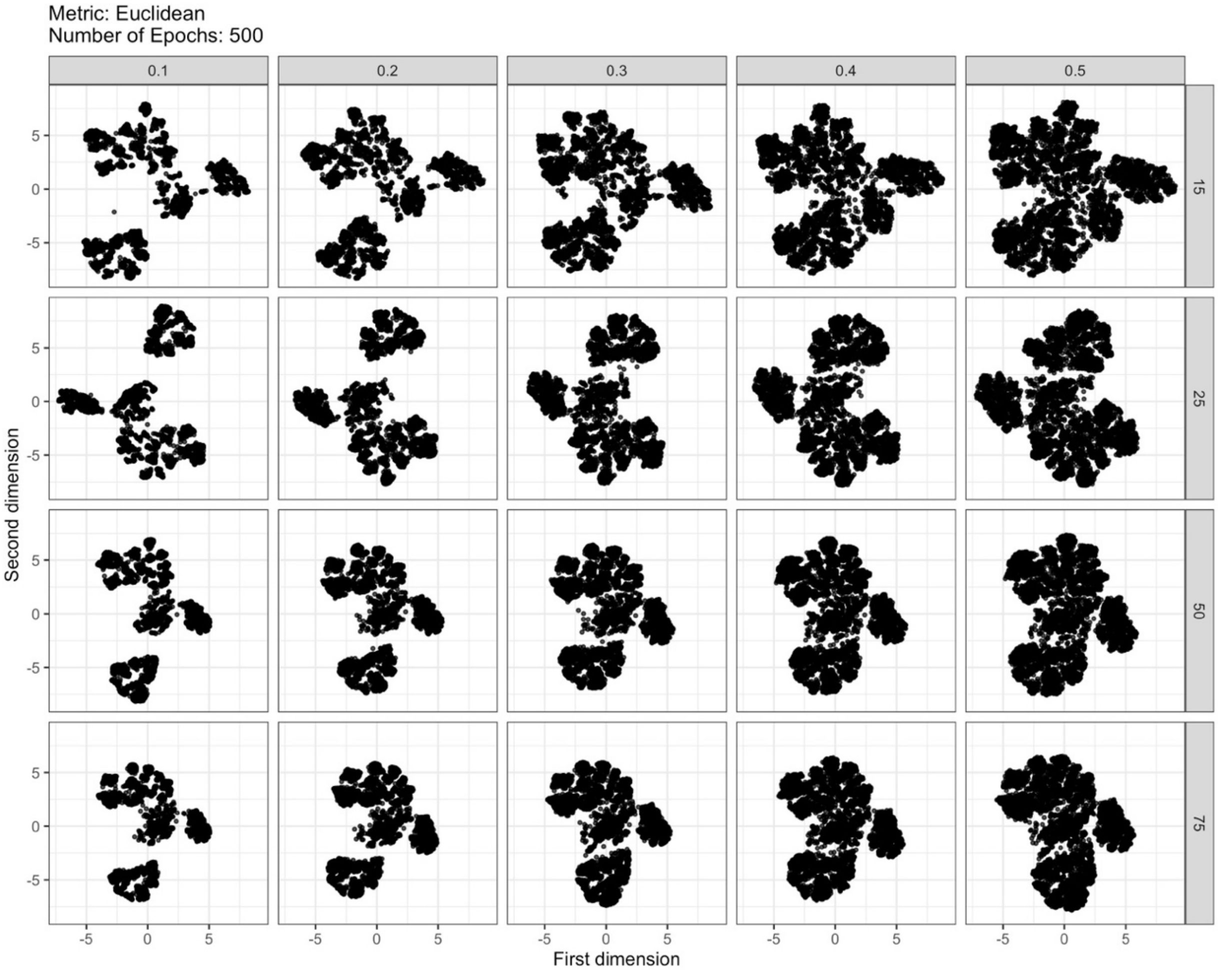


Figure 7. 2016 UMAP embeddings with different metrics (rows and varying minimum distance (columns)

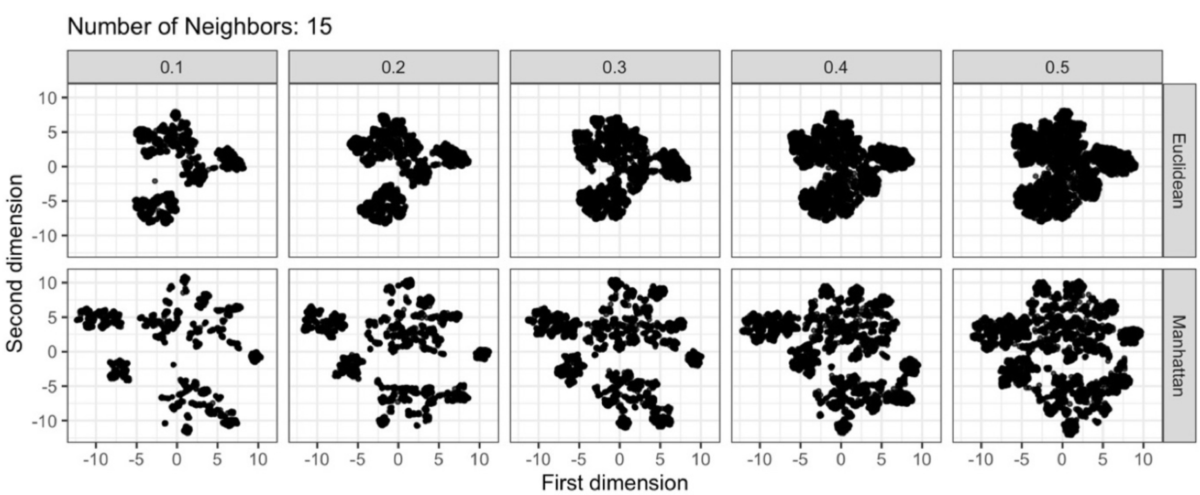
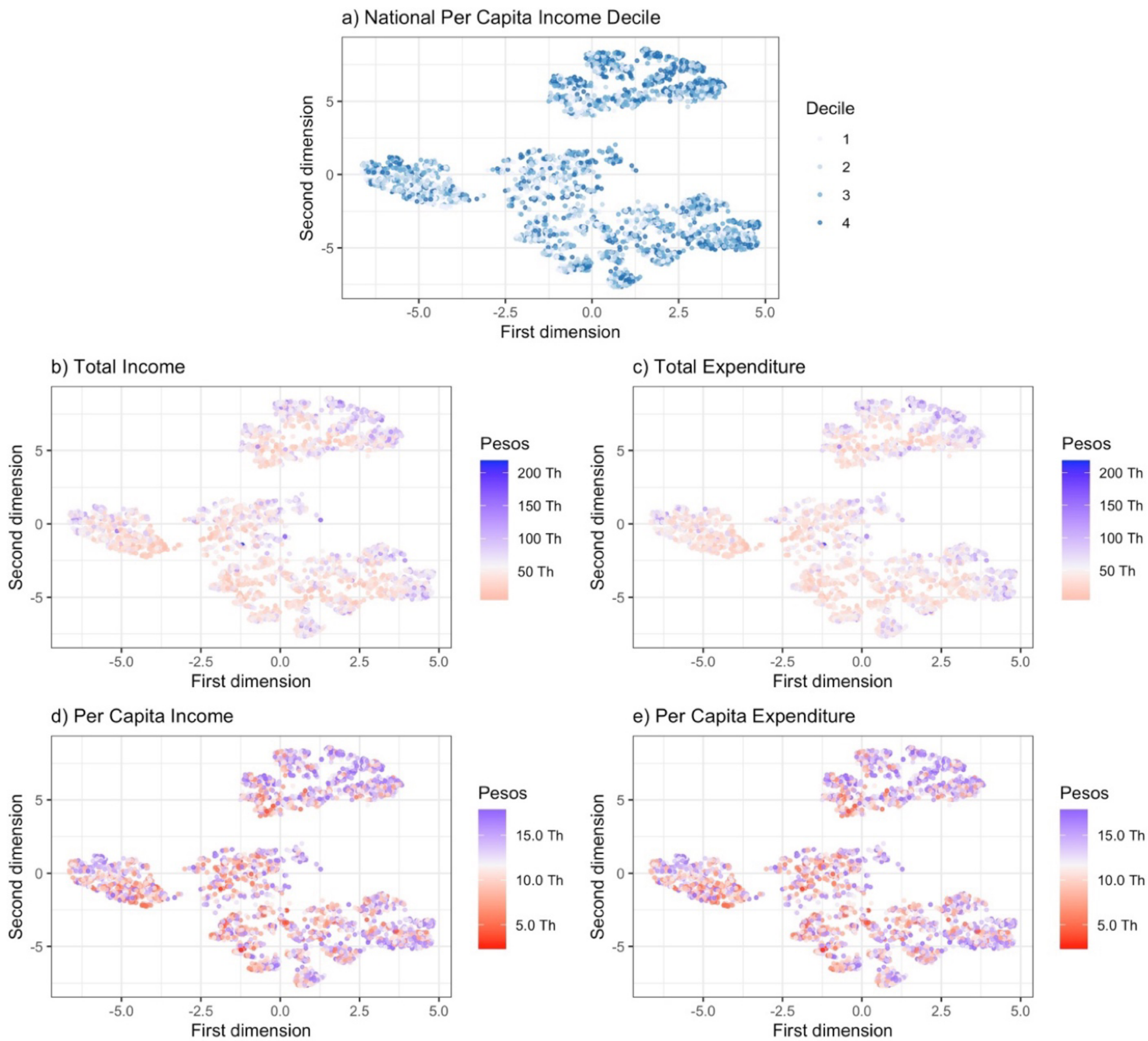


Figure 8. Final 2016 UMAP embeddings using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.

Finally, the 2017 model is trained on 41 variables (see Table 4 in the appendix) and 4,433 observations from the 2017 APIS dataset. This has one additional variable in comparison to the 2016 APIS as it includes a variable indicating whether the household resides in an urban or rural area, which is not available in the 2016 data. The resulting embeddings for the 2017 UMAP model with varying hyperparameters are shown in Figures Figure 9 and Figure 10. In the same way as the 2016 results, the groupings are not very apparent with around three or four larger clusters.

The final embeddings for the 2017 model using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs are presented in Figure 11 above. Like the previous models, each plot in the figure is colored according to an income- or expenditure-related variable. The results of this model are similar to that of the 2009 and 2017 models in that the natural groupings observed appear to be unrelated to income and expenditure. Households that earn and spend more are in the same groups as households that earn and spend less. Again, there is no evident division among richer and poorer households in bottom 40% of the population according to only the non-income poverty indicators considered. Likewise, the same patterns are observed for the UMAP model applied to the dataset without households from NCR. As anticipated, these households have little impact on the model as they only make up 4.4% (i.e., a total of only 197 households) of the whole dataset (see Figures Figure 18, Figure 19, and Figure 20 in the appendix).

Figure 9. 2017 UMAP embeddings with varying number of neighbors (rows) and minimum distance (columns)

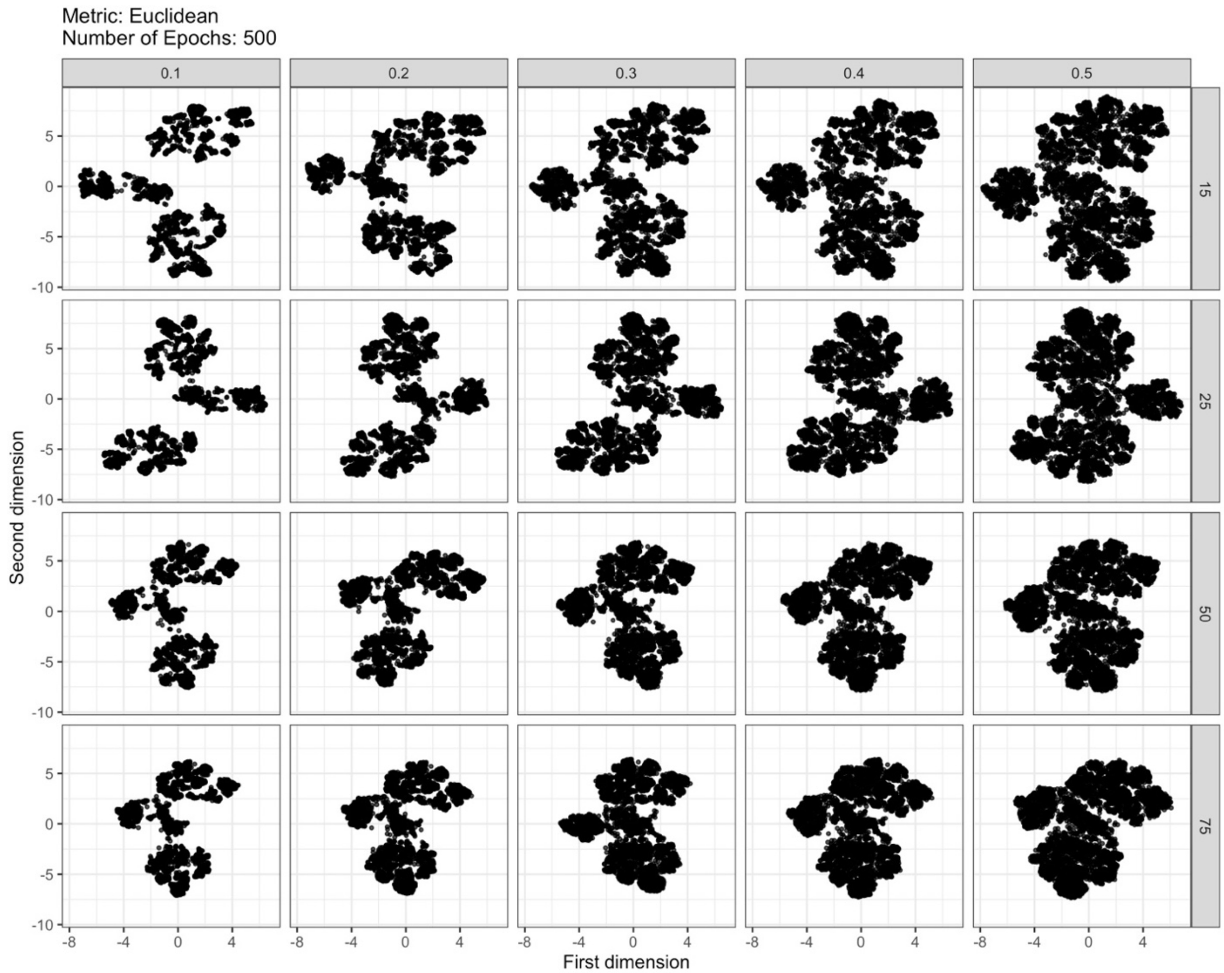


Figure 10. 2017 UMAP embeddings with different metrics (rows and varying minimum distance (columns)

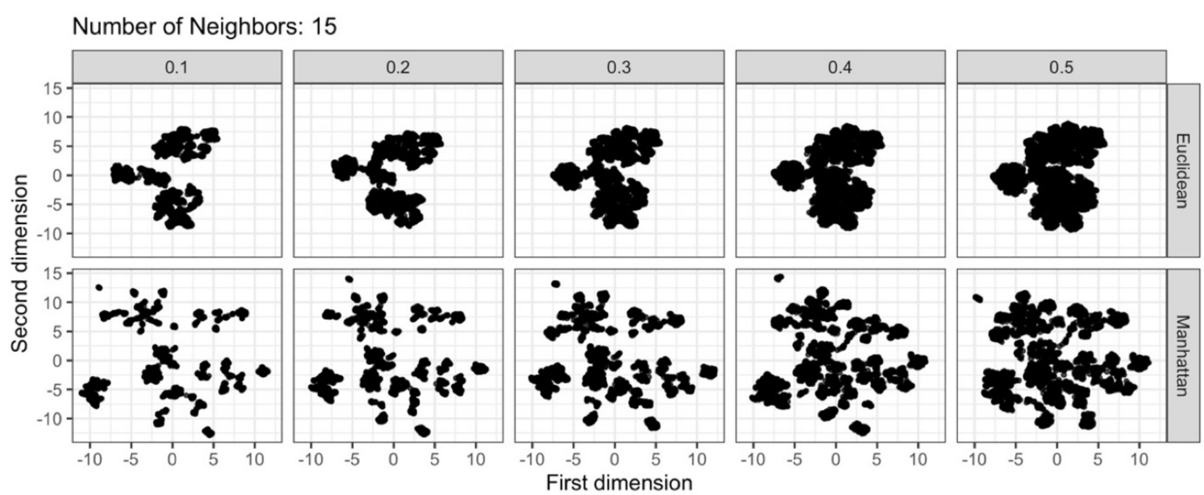
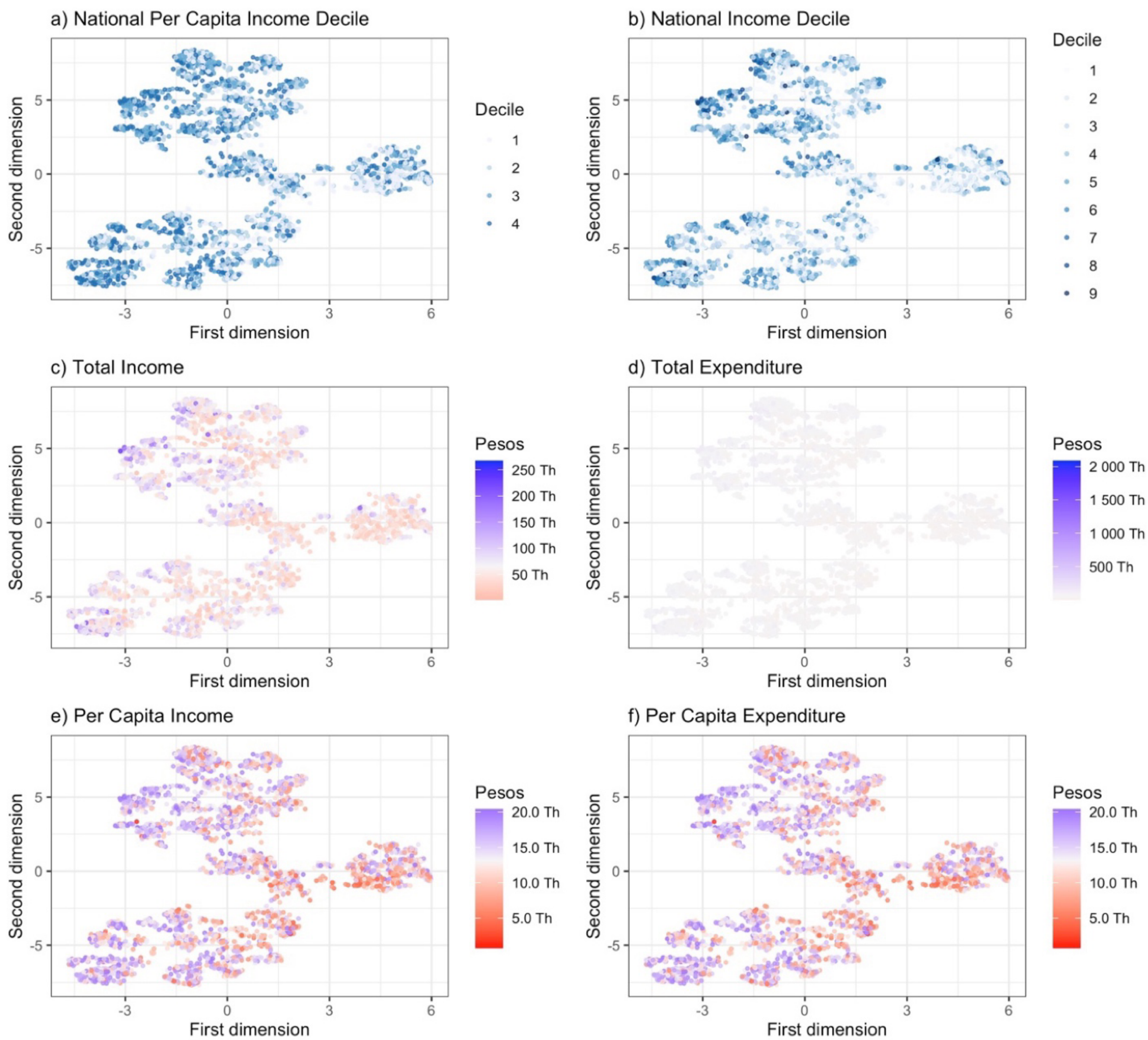


Figure 11. Final 2017 UMAP embeddings using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.

V. Discussion

The results from the three UMAP models show that households in the four poorest income deciles are alike across a range of non-income poverty indicators, such as household characteristics, household composition, human capital resources, and physical assets, regardless of their income and consumption. Based on these set of indicators, households can be naturally grouped into about four clusters; however, these groupings are unrelated to measures of income and expenditure. That is, there is no clear separation between the richest and poorest households in the bottom 40% of the population with reference to household- and individual-specific non-income poverty indicators used in the PMT models for the *Listahanan*. This provides some indication that the set of indicators used in the targeting system to estimate the income of households may not be able to accurately differentiate between poor and non-poor households.

Still, it must be noted that the analysis conducted does not provide the full picture since it does not include all the variables used in the second PMT models. Community-specific variables are not considered in the model as this information is not available. It is possible that the inclusion of these variables is crucial in classifying households according to their income. In fact, the variables were incorporated into the second PMT models precisely because they are found to explain households' per capita income in communities and are expected to increase the PMT models' goodness-of-fit and lower within-sample errors (Velarde 2018). Additionally, based on the UMAP results, there also appears to be evidence that adding more correlates does in fact help to differentiate households and create clearer separation between clusters. Although the UMAP models are not directly comparable, it is evident that the 2009 model with 83 variables does a better job in dividing the households into groups than the 2016 and 2017 models, which have only about half the number of variables. It may be the case that increasing the variables in the 2016 and 2017 models would result in patterns that are more similar to the 2009 results.

In terms of how the indicators may have changed over time, the results generally show consistent patterns in the clusters produced from the UMAP models across years. The 2009 UMAP embeddings appear very different from the 2016 and 2017 results as this uses a much larger number of variables. Nevertheless, there are noticeable similarities in the groups formed from the different datasets. Specifically, there are three to four large clusters in the results of all three models and households with higher income and higher expenditure tend to be located at the edges of the clusters. The 2016 and 2017 models, which only have a difference of one variable, have very similar patterns, including with regard to the sizes and shapes of the clusters and how households with different levels of income and expenditure are distributed. This result is as expected given that the data are collected only a year apart and variables used in the PMT models are those that do not easily change over a short period of time. A better comparison would have been over a longer period of time; however, there is no comparable data as the survey questionnaire for the APIS was modified in the following years. For instance, questions on households' assets were changed from asking the number of a specific appliance they owned into whether or not they owned the specific appliance.

The results discussed are only an initial examination of the non-income poverty indicators included in the second PMT models. Foremost, based on the results of this analysis, it may be worth exploring why households cluster into three to four groups based on the non-income poverty indicators in the *Listahanan 2*. There may be other significant variables that are related to the clustering patterns observed, which are not associated with households' income and expenditure. Furthermore, a more in-depth analysis on more detailed and complete data is necessary to fully understand whether these indicators reflect differences between households according to their income. For instance, the initial plan for this study was to train the UMAP model on the reference data and use the learned manifold to predict the embeddings for more recent years. With the 2009 FIES-LFS being the reference data, the UMAP model fit using this dataset should be able to predict embeddings from the 2015 FIES-LFS datasets. The results from this could then have been used to better assess whether patterns observed changed over time. Although the datasets are available upon request, the microdata that had been provided by the Philippine Statistics Authority to the author was incomplete. In particular, variables on housing characteristics were not included in the dataset despite being in the questionnaire. Moreover, as mentioned previously, a more thorough analysis should include the community-specific indicators from the Census of Population and Housing.

Even so, a clear separation of households in the dataset according to their income may not be needed for the PMT model to perform well. It should be pointed out that the UMAP and other unsupervised machine learning approaches do not make any prior assumptions about the dataset; hence, variables are typically normalized to have equal weights in the model. This contrasts with the objective of PMT models of assigning different weights to different variables to produce an estimate. Nonetheless, the aim of applying an unsupervised learning approach is to examine whether households of different income levels are different in terms of the non-income poverty indicators used to estimate their level of welfare and classify them into poor and non-poor. If results show that they are not very different from one another, then it is worth reevaluating which indicators would better represent their differences.

VI. Conclusion

This paper contributes to the literature on poverty targeting in the Philippines by exploring the non-income poverty indicators used in the Proxy Means Test for the *Listahanan*. Primarily, the results of the initial assessment using an unsupervised learning approach shows that there are natural groupings among lower income households according to individual- and household-specific characteristics deemed to be relevant for identifying poverty, but these differences are reflective of neither household income nor expenditure. This preliminary examination raises the question of whether poor and non-poor households can truly be differentiated based on the set of non-income poverty indicators utilized precisely for this purpose. Likewise, this calls attention to prospects that the indicators may be capturing differences among households across other factors unrelated to income and expenditure that may be significant as well. However, this study presents only partial results and inferences considering the limited data and information available.

Notwithstanding, the analysis conducted offers an approach to examine sets of indicators that may best reflect differences between poor and non-poor households without the need for information on model specifications. As specific details of the PMT models used for the *Listahanan* are kept confidential, it is difficult for independent researchers to directly assess the targeting performance of the model. To replicate the model, researchers would have to rely on published studies that only provide general information, such as which indicators are included in the model. Other essential information, including how each indicator, especially categorical variables, are transformed and incorporated into the model are not officially published. Although there are some studies with more detailed model specifications from academics who have proposed enhancements as part of the PMT model formulation process, such as Mapa and Albis (2013), it is unclear to what extent the authors' recommendations have been adopted into the final model.

With scarce information on the model specifications, alternative approaches that do not rely on these information, such as the unsupervised learning method presented in this paper, are valuable for conducting extensive analyses.

Apart from the model itself, another challenge faced by independent researchers is having limited access to the datasets used for the *Listahanan*. While data from household surveys conducted by the Philippine Statistics Authority are available upon request, the geographic information included is restricted to the households' region of residence to prevent disclosing sensitive information. Due to this limitation, it is not possible to identify the corresponding community-specific information for each household in the dataset. With the present information and data constraints, only internally conducted studies can provide a full evaluation of the targeting system. Yet, as Dadap-Cantal, Fischer, and Ramos (2021) note, there are concerns on the partiality of reports produced or commissioned by the very authorities that have established the system as these tend to emphasize its merits rather than assess whether it is truly effective.

As the *Listahanan* is the primary system used for identifying poor Filipinos and is the basis for determining beneficiaries of the country's largest social protection programs and services, it is necessary to ensure that it is reaching who it intends to benefit. With this, it is crucial for the system to be thoroughly evaluated. Thus far, limited studies have been conducted on assessing the targeting performance of the system, including on its use of the Proxy Means Test. Most studies have been internal reports, likely due to scarce publicly available data. This study presents one approach to evaluate the PMT considering such constraints. However, for a comprehensive evaluation, access to more information is essential. The Philippine government must make more resources available to independent researchers to allow them to produce their own assessments and contribute to the literature. For instance, the Philippine Statistics Authority permits researchers to access some of their more sensitive data, such as the Annual Business Survey of Philippine Business and Industry, by providing access to their data enclave facility and only allowing results of statistical runs to be provided to the researchers. This enables independent researchers to access confidential data while still making certain that the data is protected. Offering similar solutions to address the issue of data and information constraints will allow more researchers to conduct their own evaluations of the targeting system. This can lead to more recommendations for improvements that can be made to increase the coverage of social protection programs and services to the poor in the country.

References

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102, no. 4 (June): 1206-1240. http://www.jstor.org/stable/23245454.

Australian Agency for International Development. 2011. *Targeting the Poorest: An assessment of the proxy means test methodology.* Canberra: Australian Agency for International development. https://www.dfat.gov.au/sites/default/files/targeting-poorest.pdf.

Banerjee, Abhijit, Rema Hanna, Benjamin A. Olken, and Sudarno Sumarto. 2020. "The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia." *Journal of Public Economics Plus* 1, no. 100001 (January). https://doi.org/10.1016/j.pubecp.2020.100001.

Besley, Timothy and Ravi Kanbur. 1990. "The Principles of Targeting." Policy Research Working Paper Series No. 385. Washington, DC: World Bank. https://documents1.worldbank.org/curated/en/212811468739258336/pdf/multi0page.pdf.

Besley, Timothy. 1990. "Means Testing versus Universal Provision in Poverty Alleviation Programmes." *Economica* 57, no. 225 (February): 119-129. https://doi.org/10.2307/2554085.

Coady, David P. and Susan W. Parker. 2009. "Targeting Performance under Self-selection and Administrative Targeting Methods." *Economic Development and Cultural Change* 57, no. 3 (April): 559-587. https://doi.org/10.1086/596615.

Coady, David, Margaret Grosh, John Hoddinott. 2004. *Targeting of Transfers in Developing Countries : Review of Lessons and Experience*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/14902.

Dadap-Cantal, Emma Lynn, Andrew M. Fischer, and Charmaine G. Ramos. 2021 "Targeting versus Social Protection in Cash Transfers in the Philippines: Reassessing a Celebrated Case of Social Protection." *Critical Social Policy* 41, no. 3 (August): 364-84. https://doi.org/10.1177/02610183211009891.

Department of Social Welfare and Development. n.d. *Listahanan Info Kit*. https://listahanan.dswd.gov.ph/wp-content/uploads/2019/11/listahanan_info_kit_7.pdf.

Department of Social Welfare and Development. n.d.a. "Listahanan Info Kit." Accessed April 24, 2022. https://listahanan.dswd.gov.ph/wp-content/uploads/2019/11/listahanan_info_kit_7.pdf.

Department of Social Welfare and Development. n.d.b. "Listahanan Info Kit." Accessed June 13, 2022. https://fo1.dswd.gov.ph/wp-content/uploads/2018/07/Listahanan-InfoKit.pdf.

Department of Social Welfare and Development. n.d.c. "National Results of Listahanan 2." Accessed April 24, 2022. https://fo1.dswd.gov.ph/wp-content/uploads/2021/01/Listahanan-2-National-Profile-of-the-Poor.pdf.

Grosh, Margaret E. and Judy L. Baker. 1995. *Proxy means test for targeting social programs*. Washington, DC: World Bank. https://doi.org/10.1596/0-8213-3313-5.

Hanna, Rema and Benjamin A. Olken. 2018. "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries." *Journal of Economic Perspectives* 32, no. 4 (Fall): 201-226. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.32.4.201.

Karlan, Dean and Bram Thuysbaert. 2019. "Targeting Ultra-Poor Households in Honduras and Peru." *The World Bank Economic Review* 33, issue 1 (February): 63-94. https://doi-org.proxy.uchicago.edu/10.1093/wber/lhw036.

Lavallée, Emmanuel, Anne Olivier, Laure Pasquier Doumer, and Anne-Sophie Robilliard. 2010. "Poverty Alleviation Policy Targeting: A Review of Experiences in Developing Countries." Document de Travail 2010-10, Dauphine Université Paris, Paris. https://dial.ird.fr/wp-content/uploads/2021/10/2010-10-Poverty-alleviation-policy-targeting-a-review-of-experiences-in-developing-countries.pdf.

Mapa, Dennis S. and Manuel Leonard F. Albis. 2013. "New Proxy Means Test (PMT) Models: Improving Targeting of the Poor for Social Protection." Paper presented at the 12th National Convention on Statistics, Mandaluyong City, Philippines, October 2013.

Petinglay, Annabel Consuelo. "DSWD completes enumeration of cash grant beneficiaries in Antique." *Philippine News Agency*, April 21, 2022. https://www.pna.gov.ph/articles/1172647.

Philippine Statistics Authority. n.d. "Annual Poverty Indicators Survey (APIS)." Accessed June 13, 2022. https://psa.gov.ph/content/annual-poverty-indicators-survey-apis.

Premand, Patrick and Pascale Schnitzer. 2021. "Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an Experiment in Niger." *The World Bank Economic Review* 35, issue 4 (November): 892-920. https://doi.org/10.1093/wber/lhaa019.

Rhys, Hefin I. 2020. *Machine Learning with R, the tidyverse, and mlr*. New York: Manning Publications Co.

Saavedra, John Rey. "'Listahanan' to boost R7 poor sector's healthcare program: DOH." *Philippine News Agency*, February 15, 2022. https://www.pna.gov.ph/articles/1167812.

Velarde, Rashiel. 2018. *The Philippines' Targeting System for the Poor: Successes, lessons, and ways forward.* World Bank Social Protection Policy Note No. 16. Manila: World Bank.

https://documents1.worldbank.org/curated/en/830621542293177821/pdf/132110-PN-P162701-SPL-Policy-Note-16-Listahanan.pdf.

Weiss, John, ed. 2005. *Poverty Targeting in Asia*. Tokyo: Asian Development Bank Institute. https://www.adb.org/sites/default/files/publication/159383/adbi-poverty-targeting-asia.pdf.

Weiss, John. 2004. *Poverty Targeting in Asia: Experiences from India, the Philippines, People's Republic of China and Thailand*. ADBI Research Policy Brief No. 9. Tokyo: Asian Development Bank Institute. https://www.adb.org/sites/default/files/publication/157277/adbi-rpb9.pdf.

World Bank. 2018. *The State of Social Safety Nets*. Washington, DC: World Bank. http://hdl.handle.net/10986/29115.

World Bank. n.d. *Measuring income and poverty using Proxy Means Tests*. Dhaka: World Bank. https://olc.worldbank.org/sites/default/files/1.pdf.

Appendix

Table 3. Variables included in the 2009 UMAP model

No.	Variable name	Description
1	tenure_own	Tenure Status: Own or owner-like possession of house and lot
2	tenure_squatter	Tenure Status: Rent-free house and lot without consent of owner
3	b509_n_ref	Number of refrigerators
4	b5102_n_wash	Number of washing machines
5	b5151_w_phone	With telephone
6	b_5062_n_tv	Number of TVs
7	b5052_n_radio	Number of radios
8	b5072_n_vtr	Number of VCRs
9	b5082_n_stereo	Number of stereos
10	b5172_n_oven	Number of ovens
11	b5122_n_salaset	Number of sala sets
12	b5132_n_dining	Number of dining sets
13	b5142_n_car	Number of vehicles
14	b5182_n_motor	Number of motorcycles
15	b5112_n_aircon	Number of aircons
16	b5162_n_pc	Number of microcomputers
17	roof_strong	Type of roof: Strong material (galvanized,iron,al,tile,concrete,brick,stone,asbestos)
18	walls_strong	Type of wall: Strong material (galvanized,iron,al,tile,concrete,brick,stone,asbestos)
19	single_house	House building type: Single house
20	water_own	Main source of water: Own use, faucet, community water system; Own use, tubed/ piped well
21	water_shared	Main source of water: Shared, faucet, community water system; Shared, tubed/piped well
22	water_dug	Main source of water: Dug well
23	water_spring	Main source of water: Spring, river, stream, etc.
24	toilet_sealed	Toilet facility: Water-sealed
25	toilet_none	Toilet facility: None
26	electric	With available electricity
27	hhtype_single	Household type: Single Family
28	w_no_hh	Number of households in the housing unit
29	agind	Agricultural household
30	w_dom_helper	With domestic helper
31	w_urb2	Urban
32	head_ms_single	Household head marital status: Single
33	head_male	Household head sex: Male
34	fsize	Family size
35	z2021_h_age	Household head age
36	pr_bel_14	Proportion of household members aged 14 and below
37	pr_educ_ngc	Proportion of household members with no grade completed

38	pr_educ_elem_u	Proportion of household members who are elementary undergraduates
39	pr_educ_elem_g	Proportion of household members who are elementary graduates
40	pr_educ_hs_u	Proportion of household members who are highschool undergraduates
41	pr_educ_hs_g	Proportion of household members who are highschool graduates
42	pr_educ_col_u	Proportion of household members who are college undergraduates
43	pr_educ_col_g	Proportion of household members who are college graduates
44	pr_educ_pgrad	Proportion of household members with post-graduate education
45	pr_curr_sch	Proportion of household members who are currently attending school
46	pr_working	Proportion of household members who did work or had a job during the past quarter
47	occ_11	With family member whose primary occupation: officials of government and special-interest organizations
48	occ_12	With family member whose primary occupation: corporate executives and specialized managers
49	occ_13	With family member whose primary occupation: general managers or managing proprietors
50	occ_14	With family member whose primary occupation: supervisors
51	occ_21	With family member whose primary occupation: physical, mathematical and engineering science professionals
52	occ_22	With family member whose primary occupation: life science and health professionals
53	occ_23	With family member whose primary occupation: teaching professionals
54	occ_24	With family member whose primary occupation: other professionals
55	occ_31	With family member whose primary occupation: physical science and engineering associate professionals
56	occ_32	With family member whose primary occupation: life science and health professional associates
57	occ_33	With family member whose primary occupation: teaching associate professionals
58	occ_34	With family member whose primary occupation: related associate professionals
59	occ_41	With family member whose primary occupation: office clerks
60	occ_42	With family member whose primary occupation: customer service clerks
61	occ_51	With family member whose primary occupation: personal and protective services workers
62	occ_52	With family member whose primary occupation: models, salespersons and demonstrators
63	occ_61	With family member whose primary occupation: farmers and other plant growers
64	occ_62	With family member whose primary occupation: animal producers
65	occ_63	With family member whose primary occupation: forestry and related workers
66	occ_64	With family member whose primary occupation: fishermen
67	occ_65	With family member whose primary occupation: hunters and trappers

68	occ_71	With family member whose primary occupation: mining, construction and related trades workers
69	occ_72	With family member whose primary occupation: metal, machinery and related trades workers
70	occ_73	With family member whose primary occupation: precision, handicraft, printing and related trades workers
71	occ_74	With family member whose primary occupation: other craft and related trades workers
72	occ_81	With family member whose primary occupation: stationary-plant and related operators
73	occ_82	With family member whose primary occupation: machine operators and assemblers
74	occ_83	With family member whose primary occupation: drivers and mobile plant operators
75	occ_91	With family member whose primary occupation: sales and services elementary occupations
76	occ_92	With family member whose primary occupation: agricultural, forestry and fishery laborers
77	occ_93	With family member whose primary occupation: laborers in mining, construction, manufacturing and transport
78	occ_01	With family member whose primary occupation: armed forces
79	occ_09	With family member whose primary occupation: other occupations not classifiable
80	w_employer	With family member who is an employer
81	w_s_term	With family member whose nature of employment is short-term
82	w_bp_month	With family member whose basis of payment is monthly
83	w_ocw	With family member who is an overseas contract worker

Table 4. Variables used in the 2016 and 2017 UMAP models

No.	Variable name	Description
1	tenure_own	Tenure Status: Own or owner-like possession of house and lot
2	tenure_squatter	Tenure Status: Rent-free house and lot without consent of owner
3	pufeq6g	Number of refrigerators
4	pufeq6e	Number of washing machines
5	pufeq6ij	Number of cellular phone/ landline/ wireless telephone
6	pufeq6n	Number of TVs
7	pufeq6o	Number of radio/ cassette players
8	pufeq6m	Number of CD/ DVD/ DVD player
9	pufeq6k	Number of audio component/ stereo set
10	pufeq6f	Number of stove with oven/ gas range
11	pufeq6a	Number of car, jeep, van
12	pufeq6b	Number of motorcycle, tricycle
13	pufeq6d	Number of aircons
14	pufeq6h	Number of microcomputers

15	roof_strong	Type of construction materials of the roof: Strong material
16	walls_strong	Type of construction materials of the outer wall: Strong material
17	single_house	House building type: Single house
18	water_own	Main source of water: Dwelling; Yard/ Plot
19	water_shared	Main source of water: Public Tap
20	water_dug	Main source of water: Protected Well; Unprotected Well
21	water_spring	Main source of water: Developed Spring; Undeveloped Spring; Rivers/ Stream/ Pond/ Lake/ Dam
22	toilet_sealed	Toilet facility: Flush Toilet
23	toilet_none	Toilet facility: None
24	elec	With available electricity
25	head_ms_single	Household head marital status: Single
26	head_male	Household head sex: Male
27	fsize	Family size
28	pufh05_age	Household head age
29	pr_bel_14	Proportion of household members aged 14 and below
30	pr_educ_ngc	Proportion of household members with no grade completed
31	pr_educ_elem_u	Proportion of household members who are elementary undergraduates
32	pr_educ_elem_g	Proportion of household members who are elementary graduates
33	pr_educ_hs_u	Proportion of household members who are highschool undergraduates
34	pr_educ_hs_g	Proportion of household members who are highschool graduates
35	pr_educ_col_u	Proportion of household members who are college undergraduates
36	pr_educ_col_g	Proportion of household members who are college graduates
37	pr_educ_pgrad	Proportion of household members with post-graduate education
38	pr_curr_sch	Proportion of household members who are currently attending school
39	pr_working	Proportion of household members who did work or had a job during the past quarter
40	w_employer	With family member who is an employer
41	urb*	Urban household

** Only available in the 2017 APIS*

Figure 12. 2009 UMAP embeddings for areas outside the National Capita Region with varying number of neighbors (rows) and minimum distance (columns)

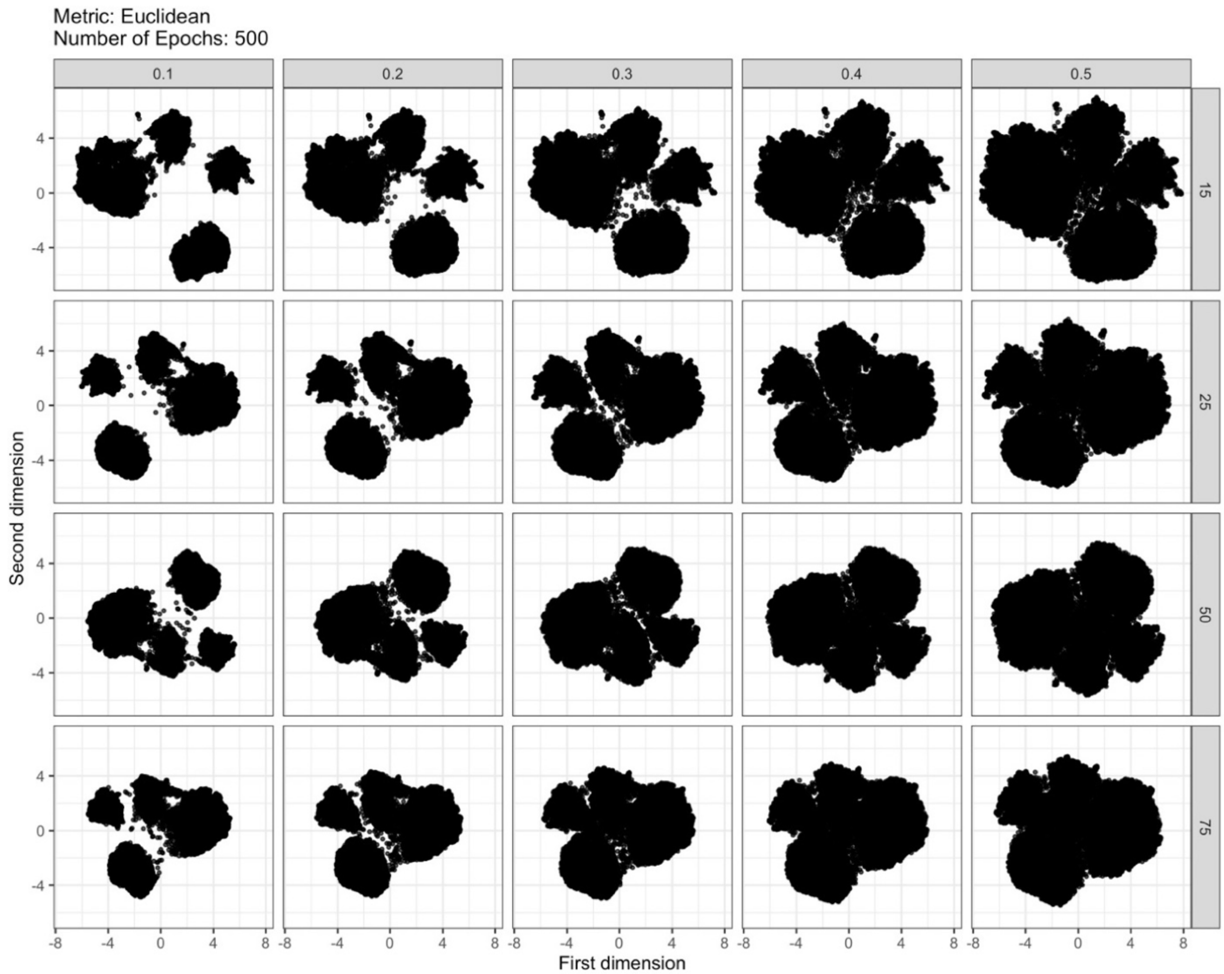


Figure 13. 2009 UMAP embeddings for areas outside the National Capital Region with different metrics (rows and varying minimum distance (columns)

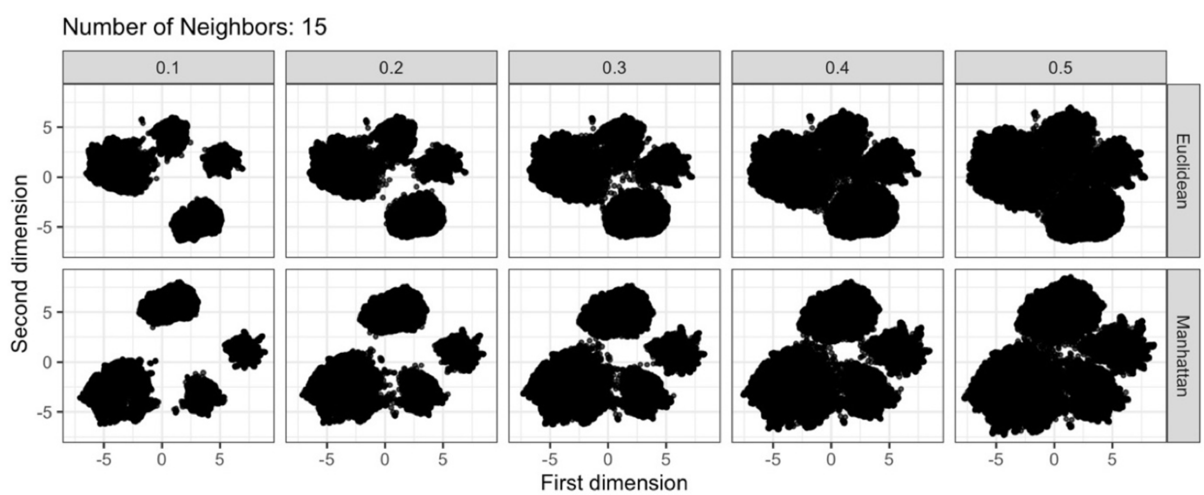
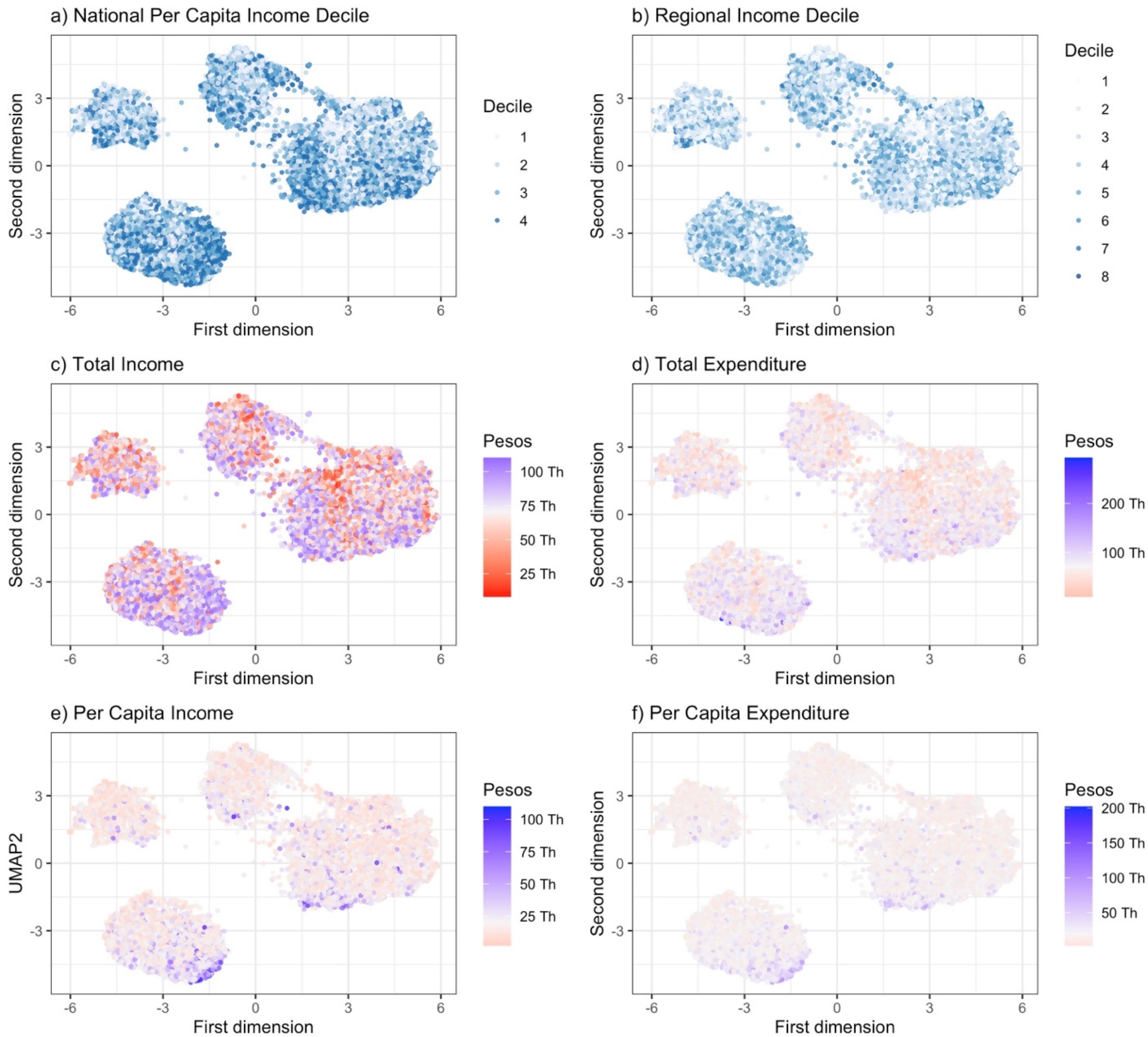


Figure 14. Final 2009 UMAP embeddings for areas outside the National Capital Region using a Euclidean metric, 25 neighbors, 0.1 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.

Figure 15. 2016 UMAP embeddings for areas outside the National Capita Region with varying number of neighbors (rows) and minimum distance (columns)

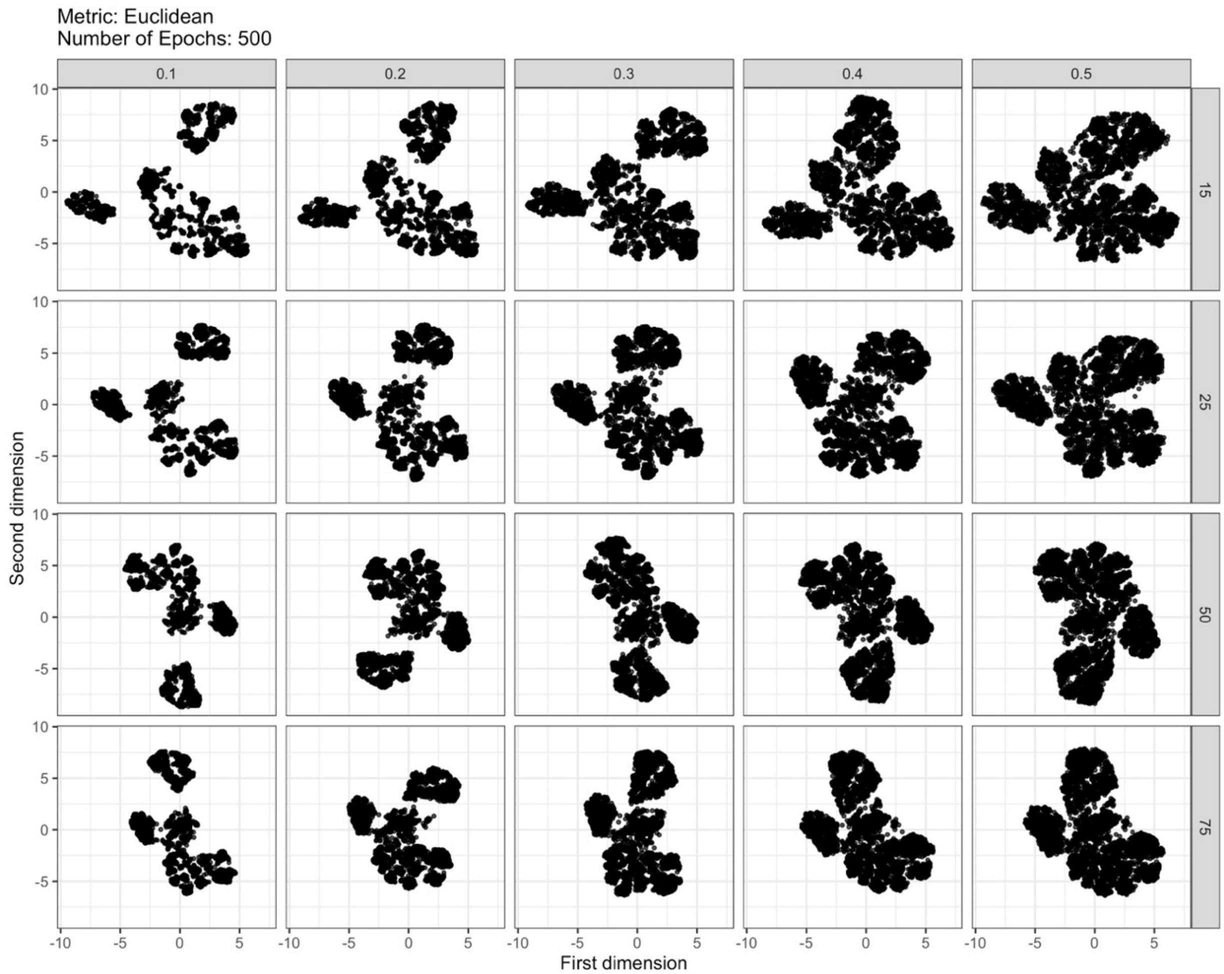


Figure 16. 2016 UMAP embeddings for areas outside the National Capital Region with different metrics (rows and varying minimum distance (columns)

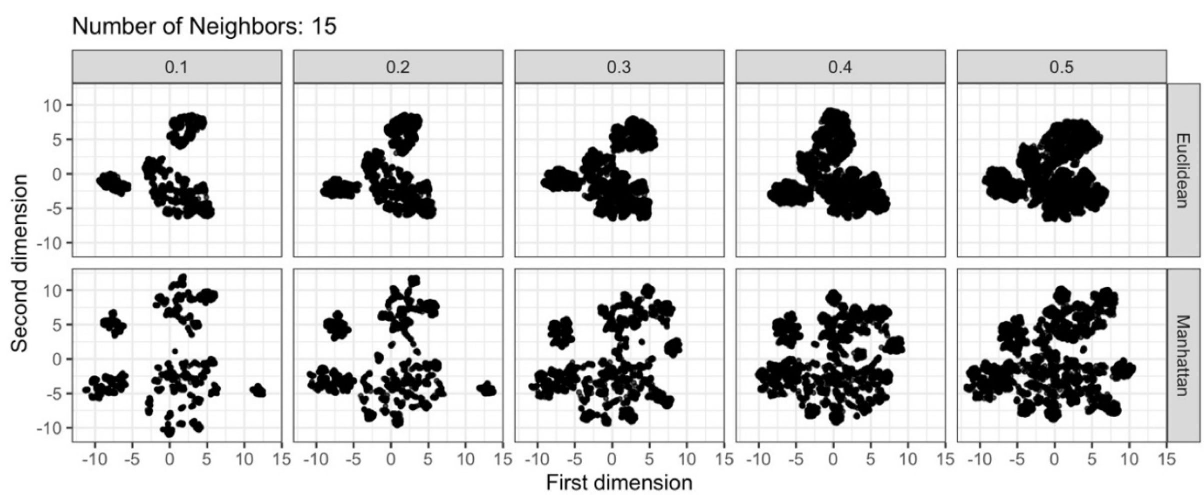
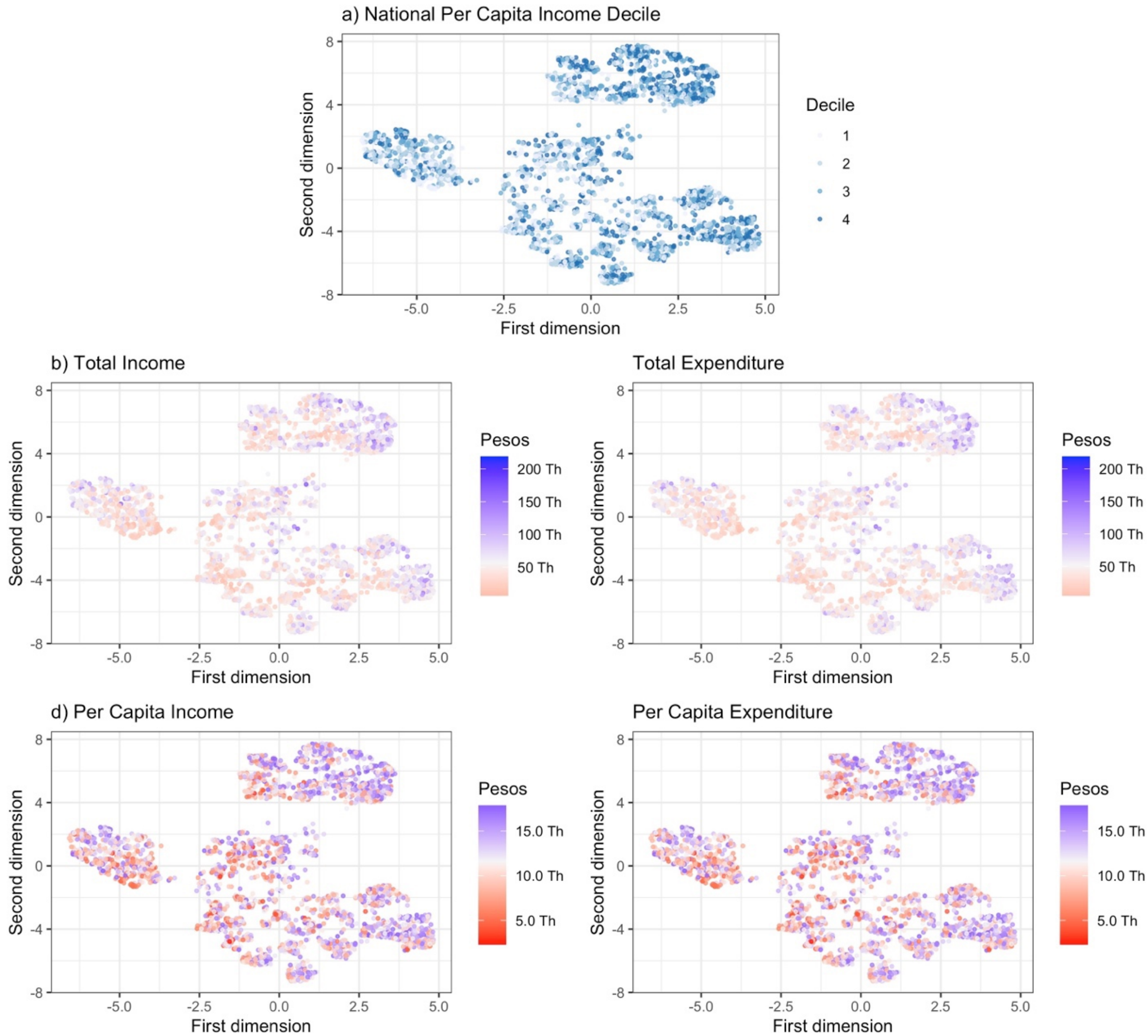


Figure 17. Final 2016 UMAP embeddings for areas outside the National Capital Region using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.

Figure 18. 2017 UMAP embeddings for areas outside the National Capita Region with varying number of neighbors (rows) and minimum distance (columns)

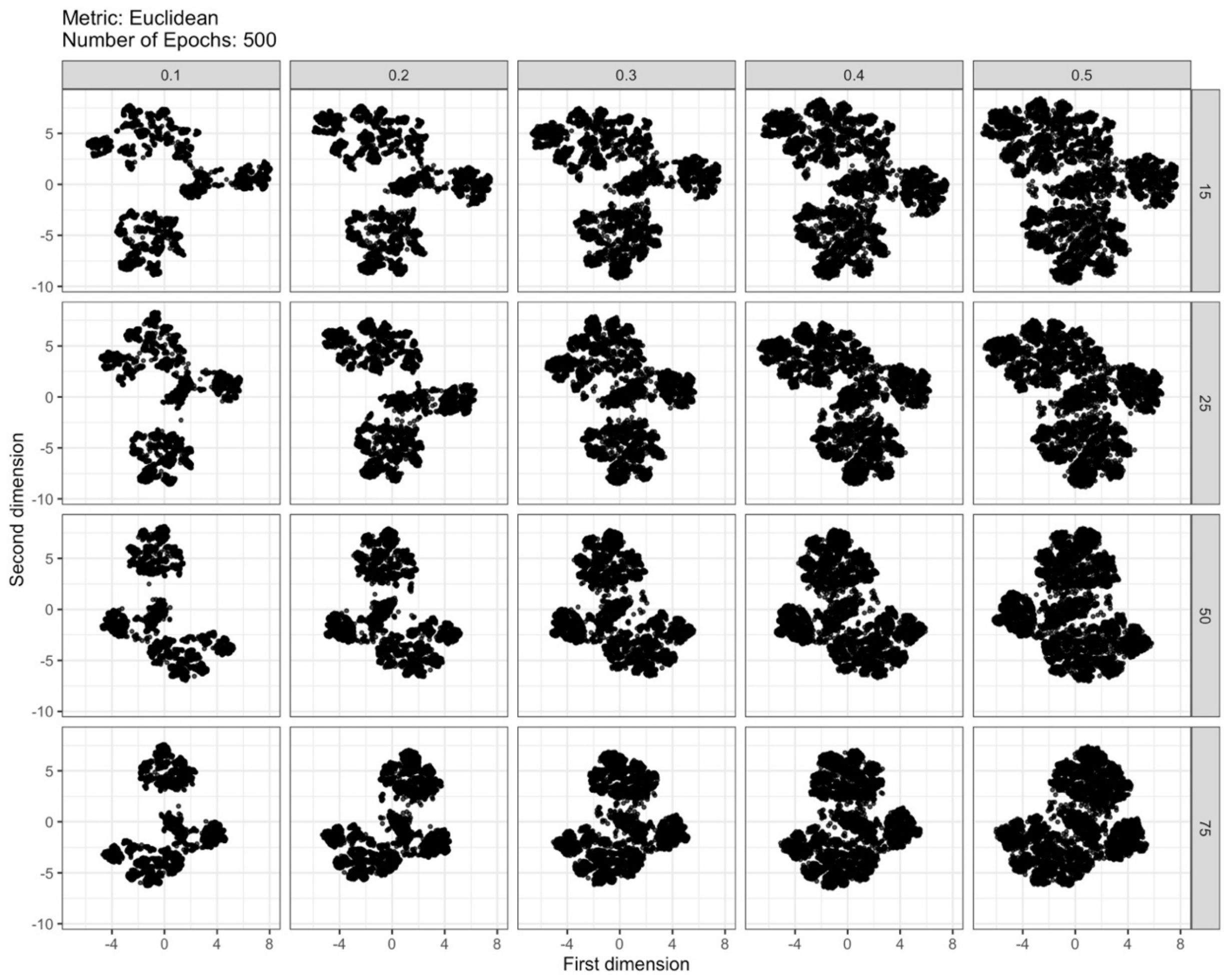


Figure 19. 2017 UMAP embeddings for areas outside the National Capital Region with different metrics (rows and varying minimum distance (columns)

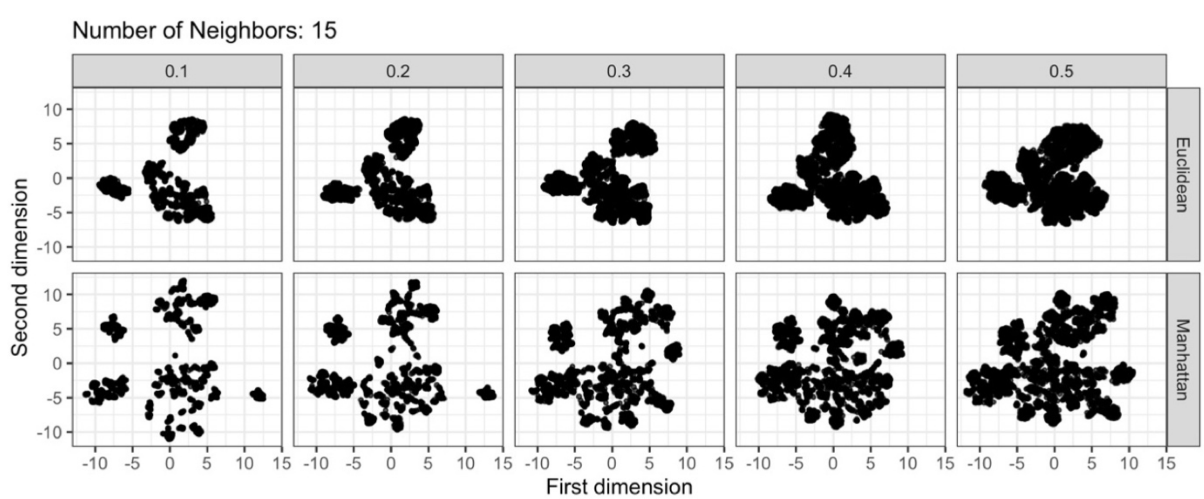
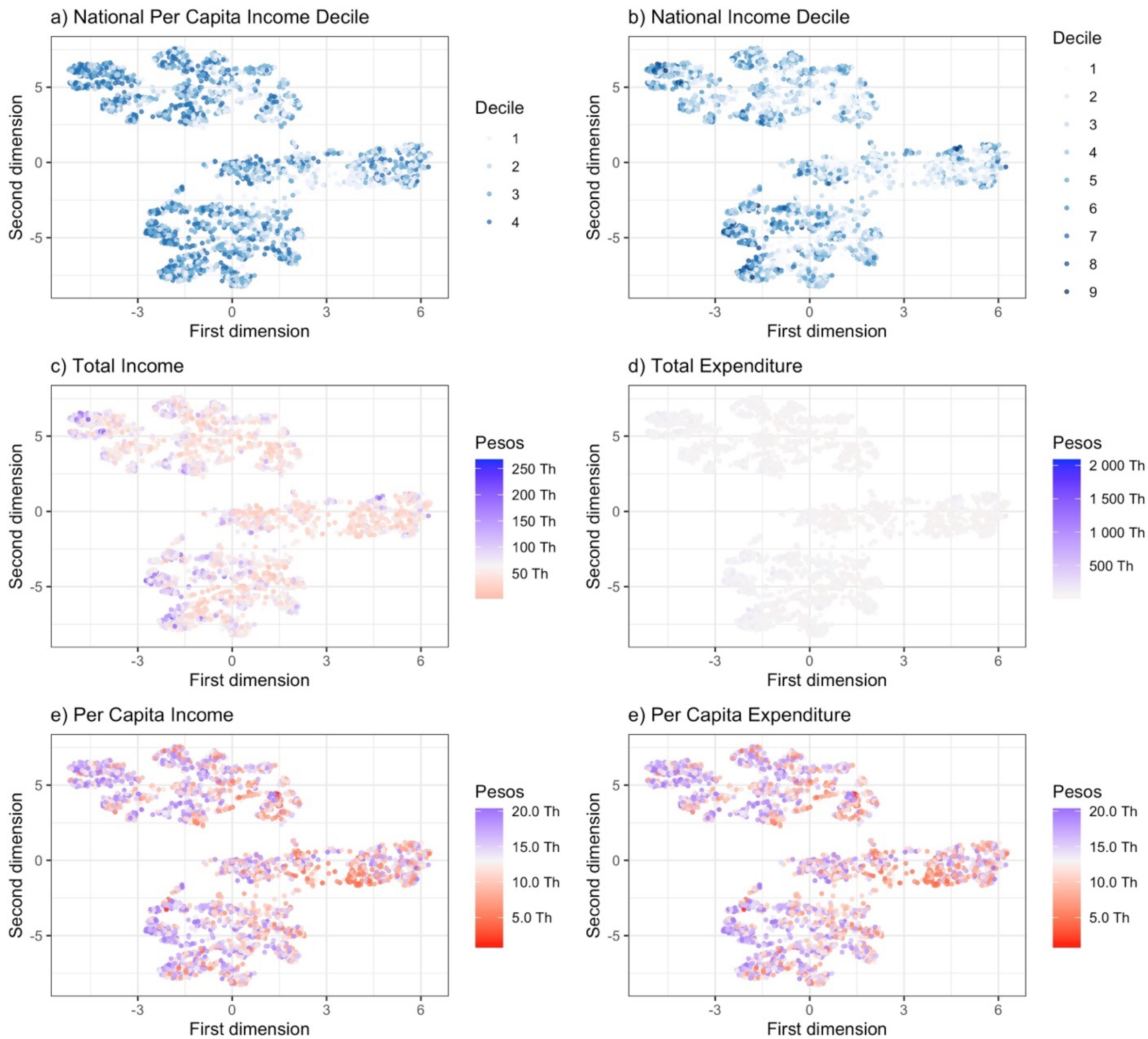


Figure 20. Final 2017 UMAP embeddings for areas outside the National Capital Region using a Euclidean metric, 25 neighbors, 0.2 minimum distance, and 500 epochs



Note: For continuous scales, dark blue represents the maximum value, gray represents the mean, and dark red represents the minimum value.