Clustering Poverty Data from Community Based Monitoring System of Pasay City Using Hidden Markov Models

Ma. Cristina Panaligan, Jeramy Aaron Yao* and Rechel Arcilla
Mathematics and Statistics Department, De La Salle University
*Corresponding Author: jeramy_yao@dlsu.edu.ph

Abstract: The poverty incidence in Pasay City has movements over time. In 2006, Pasay City has 5.3% poverty incidence which eventually decreased to 1.7% in 2009 and later increased to 1.9% in 2012. Therefore, this study aims to apply a time series clustering algorithm which can help in making better poverty reduction strategies. Clustering and profiling the households can be useful because the policy makers will know where to focus their projects to alleviate the poverty in the city. In poverty analysis, the multidimensional aspect of poverty should be considered. Thus, this study used the Multidimensional Poverty Index as the indicator for the poverty state of households. In order to define a suitable distance between household trajectories, Hidden Markov Models (HMMs) were estimated using four hidden states. Partitioning Around Medoids algorithm was used to cluster the HMMs using the distance matrix based on the Kullback-Leiber Divergence. This approach was applied to the household poverty data of Pasay City during 2005, 2008, and 2011 obtained from the Community Based Monitoring System. Results showed that there were three clusters formed. Cluster 1 contains the households whose MPI poverty state moves back and forth through the three time points, Cluster 2 captured the households that are improving in their MPI poverty state throughout the time points, and Cluster 3 is composed of households that started as MPI non poor but became poor during the last time point.

Key Words: Pasay City; Community Based Monitoring System; Multidimensional Poverty Index; Hidden Markov Models; Partitioning Around Medoids

1. INTRODUCTION

Pasay City is the third smallest city in the National Capital Region which covers 18.5 square kilometers including the airport and the reclamation areas that consumes 9.5 and 4 square kilometers, respectively. Due to the large coverage of these areas, the available space for urban development is reduced which causes high density living for the residents of the city (Galing Pook Foundation, 2006). For the purpose of alleviating poverty in Pasay City, the city government in coordination with several benefactors and other government organizations such as the Department of Social Welfare and Development, Department of Education, Petron Tulong Aral-Petron Foundation, and UNICEF planned and enacted programs such as Pantawid Pamilyang Pilipino Program (4P's), Food for School, Paskong Ligtas sa Batang Kalye, and National Household Targeting System (Pasay City Social Welfare & Development, 2010). Despite these efforts, the poverty incidence rate in the city is still fluctuating.

In 2006, Pasay City has 5.3% poverty incidence which eventually decreased to 1.7% in 2009 and later increased to 1.9% in 2012 (Philippine Statistics Authority (PSA), 2012). Since there were movements in the poverty state over time, clustering and profiling the households will help provide findings that will assist policy makers in making poverty reduction strategies.

Ghassempour (2014) developed an approach for longitudinal health data clustering. She proposed an algorithm which takes into account both continuous and categorical variables using Hidden Markov Models (HMMs). Similar studies claims that the same approach can also be used in poverty analysis because similar to health conditions; there are efficient measurements of poverty that are categorical in nature particularly, the multidimensional poverty (Sen, 1992; Dagum, 1989; Bourguignon and Chakravarty, 2003; Deutsch and Silber, 2005; Lemmi and Betti, 2006; Costa and De Angelis, 2008).

The primary aim of this study is to identify clusters of households in Pasay City and determine

the characteristics of each cluster in terms of poverty level and its covariates such as employment status and income class through time. In particular, the movement of the households in each cluster will be monitored for years 2005, 2008 and 2011. This may help to determine the efficacy of poverty-alleviating programs implemented by government and non-government organizations.

2. METHODOLOGY

The approach used involves using Hidden Markov Models in order to cluster longitudinal categorical variables because it is challenging to find a suitable distance between them. On the other hand, Partitioning Around Medoids was employed rather than traditional clustering techniques such as K-Means Clustering because it is more robust to noise and outliers (Ghassempour, 2014).

2.1 Data

Longitudinal data takes into account an essential dimension to poverty measurement that cross-sectional data cannot provide (Rodgers & Rodgers, 2009). It can be useful in monitoring of the poverty situation and can capture vast information for policy-making. Thus, this paper used longitudinal data from the Community-Based Monitoring System (CBMS) of Pasay City taken in 2005, 2008, and 2011.

The CBMS data comprises of information regarding the households in the following aspects: health, nutrition, housing, water and sanitation, education, income, peace and order, migration, overseas remittances, community/political participation, access to programs, MDG indicators, and vulnerability indicators of impacts of climate change and disaster risks, among others. Moreover, the households can be profiled by demographics such as ethnicity groups, income class, purok, barangay, etc. (CBMS, 2014).

For the purpose of capturing the multidimensional aspect of poverty, the Multidimensional Poverty Index (MPI) was used as the measure of poverty in this study. First introduced by Alkire and Foster (2017), it uses three dimensions namely education, health, and living standard which are equally weighted.

Since some indicators were not available in the CBMS data, several modifications were made. For school attendance indicator, a school-aged child is defined as a household member aged 6 to 16 years old. Instead of using the flooring of a house, its roof and wall is used as the indicator wherein a household is said to be deprived if the roof and wall of the household is made up of weak and light materials (PSA, 2010). Moreover, cooking fuel and assets ownership indicators were not available in all time points of the CBMS data for Pasay City. Thus the respective weights for each unavailable indicator were distributed equally to the other indicators under the dimension of living standard. The modified MPI is presented in Table 2.1.

Table 2.1. Modified MPI Indicators

Dimension of poverty	Indicator	Related to	Weight
Education	Years of Schooling	MDG2	1/6
	Child Attendance	MDG2	1/6
Health	Child Mortality	MDG4	1/6
	Nutrition	MDG1	1/6
Living Standard	Electricity		1/12
	Improved Sanitation	MDG7	1/12
	Improved Drinking Water	MDG7	1/12
	Roof and Wall		1/12

If a household is deprived in an indicator, it received a weight corresponding to it. Otherwise, it received a weight of zero. Then the MPI for the household was computed by taking the sum of all the weights it received. Table 2.2 shows the basis of the Poverty Status of the households based on their computed MPI.

Table 2.2. MPI Indicator

Variable	Categories	Description
Poverty Status	Moderately Poor	$33\% \leq MPI < 50\%$
	Vulnerable to Poverty	$20\% \leq MPI < 33\%$
	Moderately Non Poor	$16\% \leq MPI < 20\%$

When working with trajectories in a span of several points in time, it would be unrealistic to assume that the Markov Chains associated with the hidden states and the emission probabilities are stationary through time. One solution that can be done to address this is to add time varying covariates. The Employment Status and the Income Class of the households were used because they were more likely to be correlated with MPI (Asian Development Bank, 2009). Table 2.3 describes the categories of the covariates used.

Table 2.3. Categories for the Covariates

Variable	Categories	Description
Employment Status	No Employed	No member is employed
	Single Employed	One member is employed
	Multiple Employed	Two or more members are employed
Income Class	Low Income Class	Under Php 40,000
	Moderate Income Class	Php 40,000 – Php 99,999
	High Income Class	Php 100,000 and over

The final data frame is a list of household trajectories in the form of a matrix with columns representing the variables and rows representing the time points denoted by *Ti*.

2.2 Analysis

A set of 1,344 trajectories was selected containing households that were in a different state in each of the three time points since the method that will be implemented needs a set of trajectories that is sufficiently complex and exhibits a high degree of variability.

2.2.1 Mapping the Trajectories into HMM

The implementation of the proposed approach starts by mapping the trajectories into HMMs. Each household trajectory *Ti* was mapped into a probability density over the space of trajectories such that the probability densities belong to the class of HMMs. The probability density associated to *Ti* will be denoted by $P\lambda i = P(T|\lambda i)$, where λi is a set of parameters that have been chosen to maximize the probability of observing the trajectory *Ti*. Parameters are then estimated using the Expectation Maximization algorithm in R package depmixS4. Once the parameters of a model λi have been estimated, the probability that the model i generates an arbitrary time series using the forward-backward algorithm is computed. This is necessary to define a distance $D(Ti, Tj)$ between trajectory *Ti* and trajectory *Tj* as D(Ti,Tj) = D(Pi;Pj). These probabilities are obtained by normalizing the log likelihood matrix.

2.2.2 Forming the Distance Matrix

To determine the distance between the probability densities of trajectories, the symmetrized Kullback-Leibler (KL) divergence was used. The idea is that the likelihood can be seen as a probability density on the space of trajectories. Therefore, the distance between models λi and λj could be measured as the KL distance between the probability densities $P(T|\lambda i)$ and $P(T|\lambda j)$. The computation for the KL Divergence is also implemented using RStudio.

2.2.3 Clustering the Trajectories

After having computed the distance matrix, cluster analysis was performed using the PAM method. This can also be easily computed using the RStudio. The number of clusters chosen was based on the clustering criterions Dunn Index, DB Index, and Silhouette Index. The number of clusters chosen is where the Dunn and Silhouette Index are maximized while the DB Index is minimized.

3. RESULTS AND DISCUSSIONS

Table 3.1 shows the summary statistics of the provided data for the three time points. It can be seen that the mean MPI in 2008 is 27% which decreased in 2008 to 24.7% then eventually increased in 2011 to 26.5%. The average per capita income of households in Pasay City in 2005 is Php 42,460 which declined in 2008 to Php 40,320 then increased again to Php 62,776 in 2011. Meanwhile, the mean number of employed members of a household in the city is approximately 1 across the three time points which means that majority of the households only have 1 employed member.

Table 3.1 also shows the proportions of households that availed different types of programs. It can be seen that during 2005, the proportion of households that availed Health and

Educational programs are fairly low, while those that availed the housing programs are slightly higher but still below half. In 2008 and in 2011, it can be seen that majority of the households now availed each of the three programs.

Table 3.1. Summary Statistics

Variable	Summary Statistics	2005	2008	2011
MPI	*Mean*	0.270	0.247	0.265
	St. Dev	0.090	0.079	0.089
Per Capita Income	*Mean*	42460	40320	62776
	St. Dev	37482	36519	373564
Number of Employed Members	*Mean*	1.35	1.33	1.47
	St. Dev	0.88	0.76	0.91
Health Programs Availed	Proportion	0.272	0.999	1.000
Educational Programs Availed	Proportion	0.014	0.996	1.000
Housing Programs Availed	Proportion	0.452	0.995	No Avail. Data

The number of clusters was chosen to be three because it provided the best optimal agreement between the cluster criterions (Table 3.2). Results showed that the Dunn and Silhouette indices were relatively high while the DB index was the least when three clusters were considered.

Figure 3.1 shows the visualization of the clusters formed by using the PAM algorithm. This was performed using Multidimensional Scaling through the use of the clusterSim package in R. The plot of the clusters seems to agree that the optimal number of clusters is three.

Among the three clusters, Cluster 1 has the most number of households while Cluster 3 has the least (Table 3.3). Cluster 3 has the best clustering structure since it has the highest average silhouette width.

Table 3.2. The Dunn, Silhouette, and DB Index Values for Different Cluster Sizes

Cluster Size	Dunn Index	Silhouette Index	DB Index
2	**0.0388**	0.2626	1.8743
3	0.0314	0.2617	**1.4743**
4	0.0296	0.2537	1.6260
5	0.0297	**0.2930**	1.8132

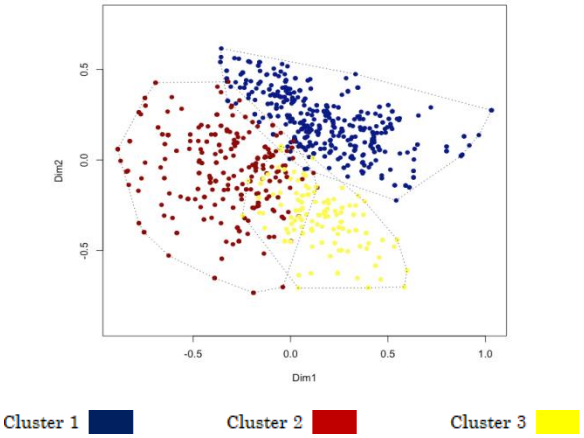

Fig 3.1. A two-dimensional MDS representation of the three clusters formed by the PAM algorithm

Table 3.3. Characteristics of Clusters Formed by the PAM Algorithm

Characteristics	Cluster 1	Cluster 2	Cluster 3
Size	608	441	295
Medoids	599	7	81
Average Silhouette Width	0.2183	0.2089	0.3533

In terms of MPI, Cluster 1 contains a lot of households in Cluster 1 that were vulnerable to poverty in 2005 but all of them moved to either moderately poor or moderately non poor in 2008.

However, in 2011 the proportion of moderately poor and moderately non poor households decreased.

Cluster 2 shows an improving proportion of poverty status along the three time points. A huge proportion of the households in Cluster 2 were moderately poor in 2005. However, in 2008, the majority of the households were now vulnerable to poverty, and during 2011 the majority of the households were now moderately non poor.

Cluster 3, on the other hand, has households that are quite the opposite of those in Cluster 2. Households in Cluster 3 during 2005 were dominated by moderately non poor households. In 2008, all of these households became vulnerable to poverty while in 2011, majority of the households were moderately poor.

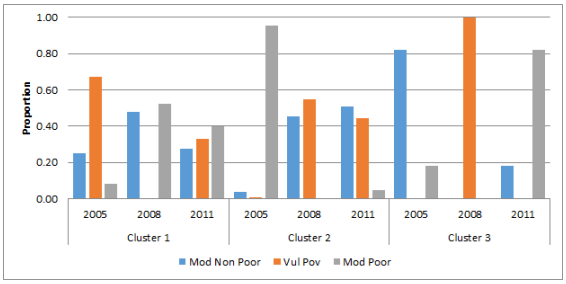

Figure 3.2. Proportion of households in each cluster based on their MPI poverty state during the three time points.

When it comes to the employment status, it can be seen in Figure 3.3 that Cluster 1 has a high proportion of households that has one or more employed members in 2005. In 2008 however, the proportion of the households with multiple employed members decreased while the proportion of households with single employed members increased which means that some of the households in Cluster 1 lost employed members. On the other hand, in 2011, the proportion is now similar to 2005 which means that some households gained more employed members.

Cluster 2 has a high proportion of households with single employed members in 2005 although in 2008 and in 2011, the proportion of households with multiple employed members increased which means that a lot of the households gained more employed members.

Cluster 3 on the other hand, has a high proportion of households that have one or more employed members in 2005, but in 2008, a lot of the households only had one employed member which means some of them became unemployed during this time. In 2011 however, the proportion of households that had multiple employed members increased a little which means some of the households gained more employed members.

Figure 3.3. Proportion of households in each cluster based on their employment status during the three time points.

Looking at Figure 3.4, it can be seen that in Cluster 1 a huge proportion of the households are either in a moderate or low income class in 2005. In 2008 and 2011 however, the proportion of households that are in the low income class became higher.

Cluster 2 on the other hand, has a high proportion of households that are in the low income class in 2005. Although in 2008 and 2011, the proportion of households that are in the moderate income class increased.

In Cluster 3, there is a huge proportion of households in either the moderate or low income class. However, the proportion of households which are in the moderate income class decreased in 2008 and 2011, while the proportion of households in the low income class increased at these time points.

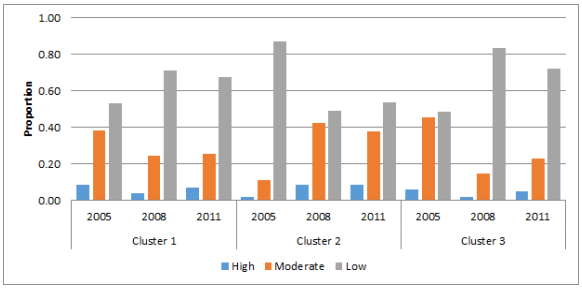

Figure 3.4. Proportion of households in each cluster based on their income class during the three time points

Figure 3.5 shows the per capita income of the households depending on their employment status. Cluster 1 is seen to have a low per capita income during 2008 given that it has households that have a single or multiple employed members. However, given the same employment status, the per capita income of the households in Cluster 1 increased in 2011, especially those that only have a

single employed member. It can also be seen that the per capita income of the households in Cluster 1 that have no employed members is gradually decreasing through the three time points.

It can be seen that the per capita income of Cluster 2 is almost always increasing through time regardless of their employment status although it can be seen that their highest per capita income is when they have multiple employed members.

On the other hand, Cluster 3 seems to have households that have a high per capita income in 2005 and a low per capita income during 2008 regardless of their employment status. It can be noticed though, that the per capita income of the households that have no employed members is higher than that of those with a single employed member.

Figure 3.5. Trend of per capita income of households that have a single employed member

4. CONCLUSIONS

4.1 Summary

Overall, the clusters formed are meaningful when it comes to profiling Cluster 2 and Cluster 3. Cluster 2 captured the households that are improving over time when it comes to their MPI poverty status while Cluster 3 captured the households that are transitioning into a worse MPI poverty status over time. Cluster 1, on the other hand, is similar to Cluster 3 which shows a high proportion of moderately poor households in 2011 although it is slightly lower than that of Cluster 3.

4.2 Recommendations

Henceforth, it is recommended that government and non-government organizations concerned in the reduction of poverty in Pasay City to focus on the households in Clusters 1 and 3 and provide more job opportunities for them. For future researches, it is suggested to explore more covariates and add more time points so that the HMMs could p roject all the possible scenarios.

5. ACKNOWLEDGEMENTS

The authors thank CBMS for providing the poverty data as well as Mrs. Merlita Lagmay, the City Planning Coordinator of Pasay City. Moreover, the authors express their gratitude to Dr. Shima Ghassempour for providing sample codes used in the study.

6. REFERENCES

Alkire, S., and Robles, G. (2017). Multidimensional poverty index- summer 2017; brief methodological notes and results. Oxford Poverty and Human Development Initiative, University of Oxford.

Bourguignon, F., and Chakravarty, S. (2003). The measurement of multidimensional poverty. Journal of Economic Inequality 1:25-49.

Costa, M., and De Angelis, L. (2008). The multidimensional measurement of poverty: a fuzzy set approach. Statistica 68:303-319.

Dagum, C. (1989). Poverty as perceived by the Leyden evaluation project. A survey of Hagenaars' contribution on the perception of poverty. Economic Notes 24:115-134.

Deutsch, J., and Silber, J. (2005). Measuring multidimensional poverty: An empirical comparison of various approaches. Review of Income and Wealth 51:145-174.

Galing Pook Foundation. (2006). Harnessing the families for the MDGS - Pasay City Philippines

Ghassempour, S. (2014). Clustering longitudinal health data using hidden markov models. International Journal of Environmental Research and Public Health,11(3), 2741-2763. doi:10.3390/ijerph110302741

Lemmi, A., and Betti, G. (2006). Fuzzy set approach to multidimensional poverty measurement. Berlin: Springer.

Pasay City Social Welfare & Development. (2010). Retrieved November 30, 2017, from http://www.pasay.gov.ph/Departments/PCCWC.html

Philippine Statistics Authority. (2010). Retrieved November 30, 2017, from https://psa.gov.ph/content/characteristics-poor-families-philippines-findings-2008-annua-poverty-indicators-survey

Rodgers, J. R., and Rodgers, J. L. (2009). Contributions of Longitudinal Data to Poverty Measurement in Australia. *Economic Record, 85*. doi:10.1111/j.1475-4932.2009.00587.x

Sen, A.K. (1992). Inequality reexamined. Cambridge, MA: Harvard University Press.