Guatemala Census analysis

Juan Carlos Rosito Cuellar

Universita' degli studi di Milano-Bicocca

23-12-2015

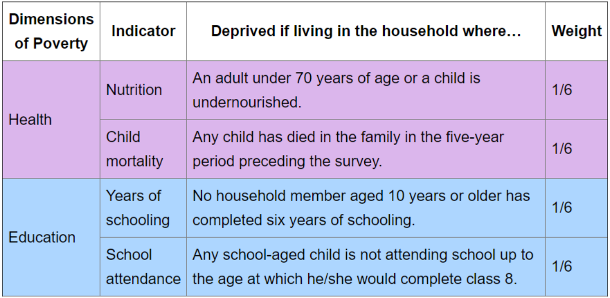
objectives

- Present the most significant variables that determine if a person is a poor person or not.
    - Define a metric that can measure the Multi-dimensional Poverty on Guatemala with the national census of 2002 data.
    - Make a Model that can predict if a person is, or not, a person categorized as a poor person by the Multi-dimensional Poverty Index(MPI).
    - Find the most relevant variables of the model

Introduction: Global Multidimensional Poverty Index (MPI) - 1

- “The MPI looks beyond income to understand how people experience poverty in multiple and simultaneous ways. It identifies how people are being left behind across three key dimensions: health, education and standard of living, comprising 10 indicators. People who experience deprivation in at least one third of these weighted indicators fall into the category of multidimensionally poor.” (UNITED NATIONS DEVELOPMENT PROGRAMME)
- The MPI is defined by the Oxfod Poverty And Human Development Initiative (OPHI) by the Oxford University.

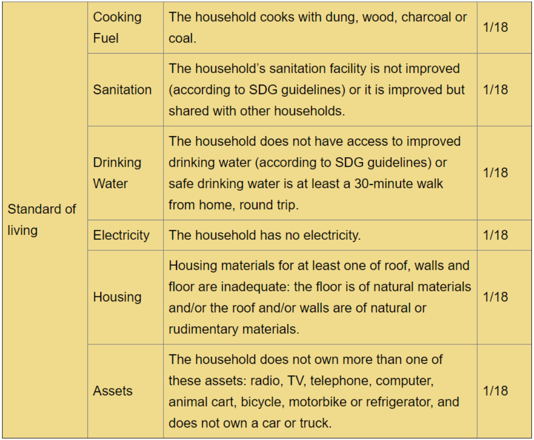
Introduction: Global Multidimensional Poverty Index (MPI) - 2



Dimensions of Poverty	Indicator	Deprived if living in the household where...	Weight
Health	Nutrition	An adult under 70 years of age or a child is undernourished.	1/6
	Child mortality	Any child has died in the family in the five-year period preceding the survey.	1/6
Education	Years of schooling	No household member aged 10 years or older has completed six years of schooling.	1/6
	School attendance	Any school-aged child is not attending school up to the age at which he/she would complete class 8.	1/6

Figure 1: Health and Education aspects and how they were taken into consideration by the MPI

Introduction: Global Multidimensional Poverty Index (MPI) - 3



Standard of living	Cooking Fuel	The household cooks with dung, wood, charcoal or coal.	1/18
	Sanitation	The household's sanitation facility is not improved (according to SDG guidelines) or it is improved but shared with other households.	1/18
	Drinking Water	The household does not have access to improved drinking water (according to SDG guidelines) or safe drinking water is at least a 30-minute walk from home, round trip.	1/18
	Electricity	The household has no electricity.	1/18
	Housing	Housing materials for at least one of roof, walls and floor are inadequate: the floor is of natural materials and/or the roof and/or walls are of natural or rudimentary materials.	1/18
	Assets	The household does not own more than one of these assets: radio, TV, telephone, computer, animal cart, bicycle, motorbike or refrigerator, and does not own a car or truck.	1/18

Figure 2: standard of living aspects and how it was taken into consideration by the MPI

Introduction: Guatemala Study of Generalized Poverty - 1

- The study that is taken as a reference is "Mapas de Pobreza en Guatemala al 2002" by the National Institute of statistics (INE) of Guatemala.
- In this study the poverty is taken as a generalized poverty, which take into account the living standard of a person.
- url: `http://fadep.org/wp-content/uploads/2016/10/D-5_MAPAS_DE_POBREZA_GUA_2002.pdf`

Introduction: Guatemala Study of Generalized Poverty - 2

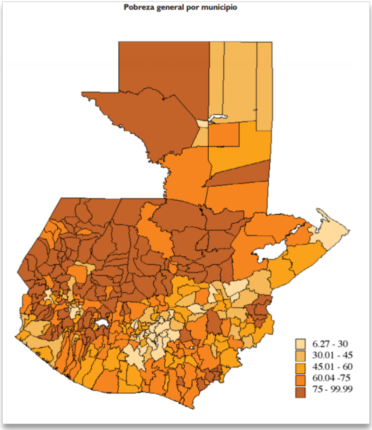
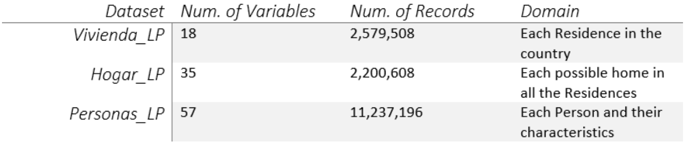


Figure 3: Guatemala State Map with a color scale of the generalized poverty

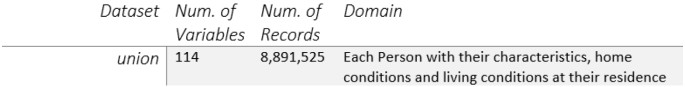
Census Dataset

- The Data is define in tree modules, the living conditions at the residence, the conditions of each house/apartment/room inside the residence, and finally the characteristics of each person that lives in each residence.



Dataset	Num. of Variables	Num. of Records	Domain
Vivienda_LP	18	2,579,508	Each Residence in the country
Hogar_LP	35	2,200,608	Each possible home in all the Residences
Personas_LP	57	11,237,196	Each Person and their characteristics

Figure 4: Datasets taking into consideration in order to make the analysis

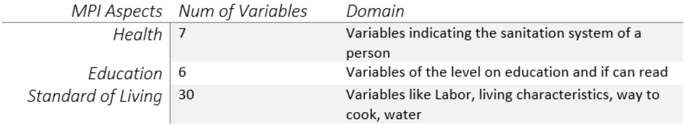


Dataset	Num. of Variables	Num. of Records	Domain
union	114	8,891,525	Each Person with their characteristics, home conditions and living conditions at their residence

Figure 5: Resulting dataset of the union of Vivienda_LP, Hogar_LP and Personas_LP; filtering records

Poverty Index Creation: MPI Aspects

- The variables of the Datasets has been measured and has been grouped by the aspects of the MPI of Oxford.



MPI Aspects	Num of Variables	Domain
Health	7	Variables indicating the sanitation system of a person
Education	6	Variables of the level on education and if can read
Standard of Living	30	Variables like Labor, living characteristics, way to cook, water

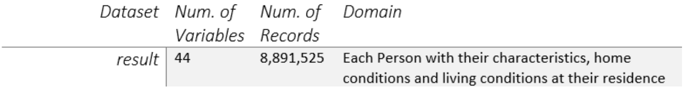
Figure 6: Variables took into consideration by the aspects of the MPI created, and the characteristics that the variables have

Poverty Index Creation: Filtering

- Each Variable taken into consideration has effects on the selected data and this lead to some records to be non applicable.
  - Health - the selected data requires that the residence is actually being inhabited by a person and is not a store
  - Education - the selected data requires that the person been interviewed has more than 6 years old
  - Living - all the persons that satisfy the requirements of Health and Education have the living data

Poverty Index Creation: Result

- The resulting dataset has been done by filtering only the variables that were involved on the Multidimensional Poverty Index (MPI) created, and filtering only the records that applied to MPI variables.



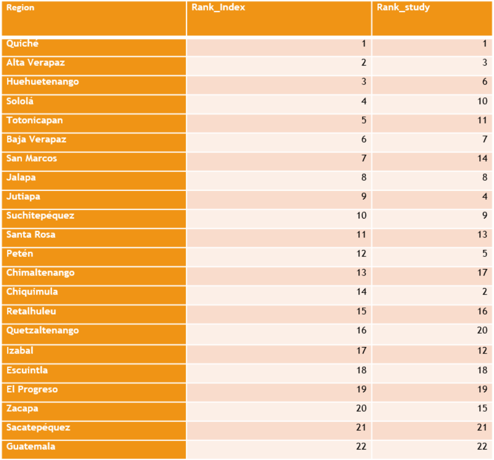
Dataset	*Num. of Variables*	*Num. of Records*	*Domain*
result	44	8,891,525	Each Person with their characteristics, home conditions and living conditions at their residence

Figure 7: Resulting dataset of the union of Vivienda_LP, Hogar_LP and Personas_LP; filtering variables and records

Poverty Index Comparison - 1

- The metric that it takes as comparison is the index general poverty of the Guatemalan study by INE, that also seek to measure the poverty not only by the income but by the aspects of living of a person.

Poverty Index Comparison - 2



Region	Rank_Index	Rank_study
Quiché	1	1
Alta Verapaz	2	3
Huehuetenango	3	6
Sololá	4	10
Totonicapan	5	11
Baja Verapaz	6	7
San Marcos	7	14
Jalapa	8	8
Jutiapa	9	4
Suchitepéquez	10	9
Santa Rosa	11	13
Petén	12	5
Chimaltenango	13	17
Chiquimula	14	2
Retalhuleu	15	16
Quetzaltenango	16	20
Izabal	17	12
Escuintla	18	18
El Progreso	19	19
Zacapa	20	15
Sacatepéquez	21	21
Guatemala	22	22

Figure 8: Index Table comparison, where Rank_index is the rank of the generated index with the MPI metrics. The Rank_study is the rank of the index of Generalized Poverty by the Guatemala Study

Poverty Index Comparison - 3

- The comparison was made with the paired difference test of "Wilcoxon Signed-Rank Test".
- The result is that the null hypothesis can't be refuted, where H0: is that the difference between the pairs follows a symmetric distribution around zero, whit a 0.05 of significance level. With a p-value = 0.8155. So we can't prove that they are different, making a well comparison of the index.

Classification Trees: Introducction - 1

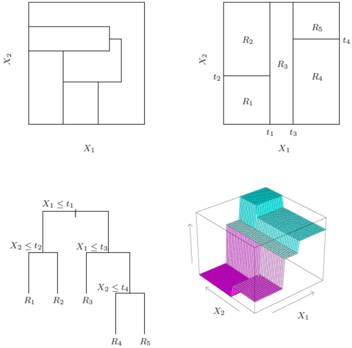


Figure 9: Rerpresents graphicly the regions of a classification tree

Classification Trees: Introducction - 2

Getting as a response variable Y *and* p *explicative variables as* X
the target categorical variable taken into consideration is poverty
where (x_i, y_i) *goes with* $i = 1, 2, \ldots, N$ *and* $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, *giving* $N = 2$ *and* $p = 43$
letting be k *as each class of the target variable and* m *as identifier of each node.*
In a node m, *representing* R_m *as the Region*
N_m *representing the quantity of observations in* m

It can be said that the proportions between each class k *of the target variable in each node* m *can be represented as* :

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(\, y_i = k)$$

Classification Trees: Introducction - 3

the splitting rule took was the Missclassification Error, where can be :

$$\begin{array}{cc}
\textit{Missclassification error} : & \frac{1}{N_m}\sum_{i\in R_m} I(\,y_i \neq k(\,m)\,)\; = 1 - \hat{p}_{mk(\,m)}\,; \\
\textit{Gini index} : & \sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(\,1 - \hat{p}_{mk}\,)\,; \\
\textit{Cross - entropy of deviance} : & -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}\,;
\end{array}$$

Classification Trees: Introducction - 2

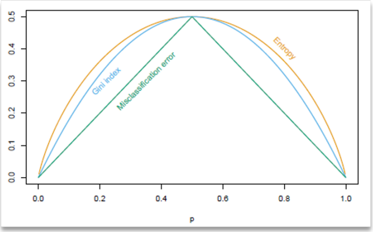


Figure 10: Shows the different spliting rules of the classification tree by the measure of impurity, in the x axis show the proportions of the binary target.

Classification Trees: Result - 1

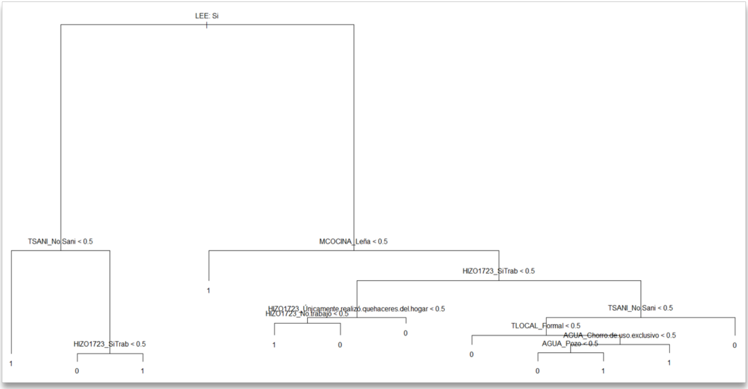


Figure 11: show the resultant tree of the dichotomous target variable and the most significant variables

Classification Trees: Result - 2

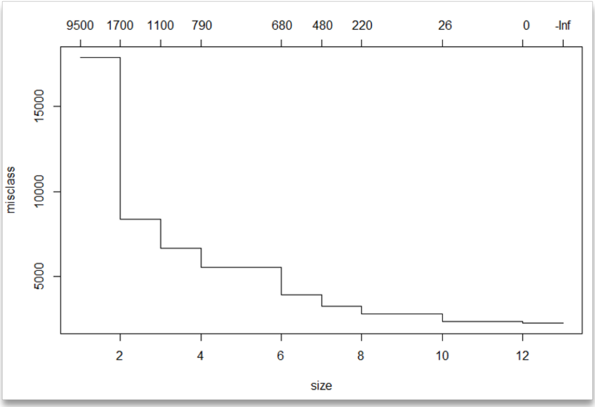


Figure 12: Missclassification rate of the prediction by the resultant model of the training of the Classification tree

LDA (Linear Discrimant Analysis): Introducction - 1

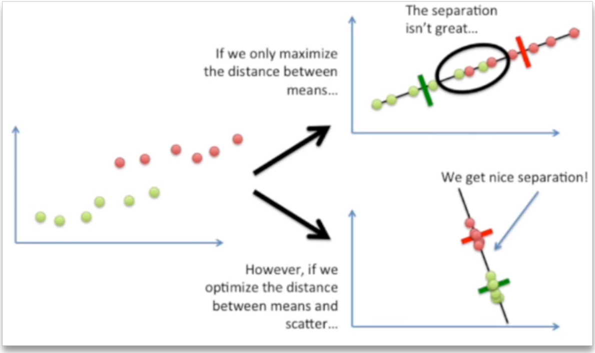


Figure 13: The Linear Discriminant Analysis tray to maximize the eucledean distance between groups separation and minimize the variation of each group

LDA: Introduction - 2

Given $f_k(x)$ *the class* $-$ *conditional density*
of X *in class* $G = k$ *and letting* π_k *be the prior probability of class* k,
Supposing that each class density is modeled
by a multivariate Gaussian, *the discriminant function is* :

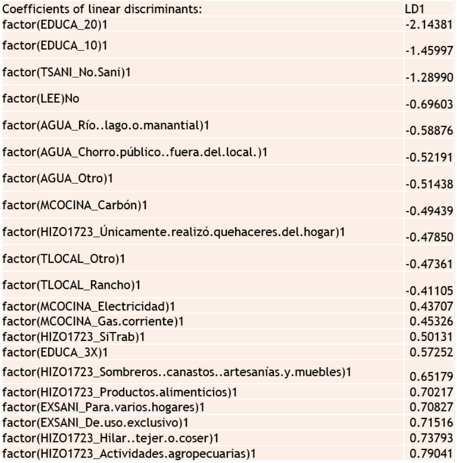
$$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + \log \pi_k;$$

$$\hat{\pi}_k = \frac{N_k}{N}, \text{ where } N_k \text{ is the number of class} - k \text{ observations};$$

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$$

supposing that :$\sum_k = \sum \forall k$;*where the target variable poverty is categorical having* 2 *classes*, 0 *or* 1

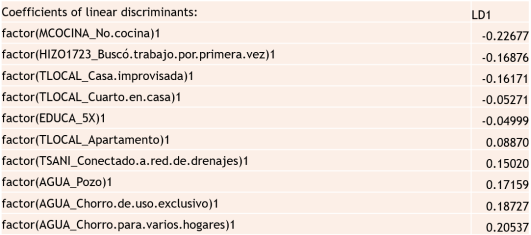
LDA: Result - 1



Coefficients of linear discriminants:	LD1
factor(EDUCA_20)1	-2.14381
factor(EDUCA_10)1	-1.45997
factor(TSANI_No.Sani)1	-1.28990
factor(LEE)No	-0.69603
factor(AGUA_Río..lago.o.manantial)1	-0.58876
factor(AGUA_Chorro.público..fuera.del.local.)1	-0.52191
factor(AGUA_Otro)1	-0.51438
factor(MCOCINA_Carbón)1	-0.49439
factor(HIZO1723_Únicamente.realizó.quehaceres.del.hogar)1	-0.47850
factor(TLOCAL_Otro)1	-0.47361
factor(TLOCAL_Rancho)1	-0.41105
factor(MCOCINA_Electricidad)1	0.43707
factor(MCOCINA_Gas.corriente)1	0.45326
factor(HIZO1723_SiTrab)1	0.50131
factor(EDUCA_3X)1	0.57252
factor(HIZO1723_Sombreros..canastos..artesanías.y.muebles)1	0.65179
factor(HIZO1723_Productos.alimenticios)1	0.70217
factor(EXSANI_Para.varios.hogares)1	0.70827
factor(EXSANI_De.uso.exclusivo)1	0.71516
factor(HIZO1723_Hilar..tejer.o.coser)1	0.73793
factor(HIZO1723_Actividades.agropecuarias)1	0.79041

Figure 14: The most significant variables that separate better the groups if a person is poor or not

LDA: Result - 2



Coefficients of linear discriminants:	LD1
factor(MCOCINA_No.cocina)1	-0.22677
factor(HIZO1723_Buscó.trabajo.por.primera.vez)1	-0.16876
factor(TLOCAL_Casa.improvisada)1	-0.16171
factor(TLOCAL_Cuarto.en.casa)1	-0.05271
factor(EDUCA_5X)1	-0.04999
factor(TLOCAL_Apartamento)1	0.08870
factor(TSANI_Conectado.a.red.de.drenajes)1	0.15020
factor(AGUA_Pozo)1	0.17159
factor(AGUA_Chorro.de.uso.exclusivo)1	0.18727
factor(AGUA_Chorro.para.varios.hogares)1	0.20537

Figure 15: The less significant variables that separate better the groups if a person is poor or not

LDA: Result - 3

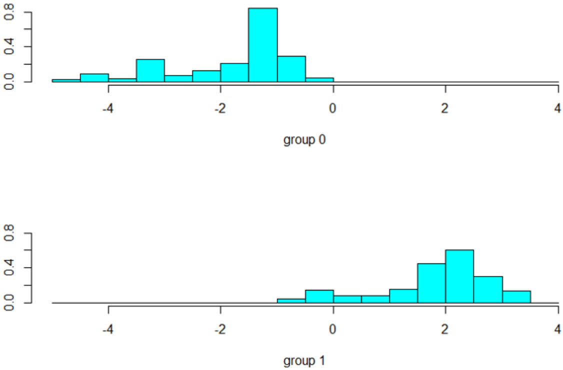


Figure 16: Histogram of the quantity of variables that separate each group, the group 0 determines if a person is poor and the group 1 if not

Conclusion

- The Alphabetization is the key variable that is more significant to determine if a person suffers of multidimensional poverty or not.
- To be out of poverty should consider to facilitate the living services of each house in a community.