Responsible Land Governance:
Towards an Evidence Based Approach
ANNUAL WORLD BANK CONFERENCE ON LAND AND POVERTY
WASHINGTON DC, MARCH 20-24, 2017

CALIBRATING HOUSING POLICIES WITH GEOSPATIAL CENSUS DATA

FARID MATUK, **LUIS TRIVEÑO**
29x55
farid.matuk@yahoo.com
World Bank
ltriveno@worldbank.org

Paper prepared for presentation at the
“2017 WORLD BANK CONFERENCE ON LAND AND POVERTY”
The World Bank - Washington DC, March 20-24, 2017

Abstract

A conventional problem for public policies is to focus in the social groups with the most acute needs; in most circumstances this imply a sort of field operation, usually under the form of a survey. This paper shows how to recycle an old product (housing and population census) with the cartography elaborated for it; in order to find population size and surface occupied with acute needs classified by type. In particular, which plots have housing without drainage systems, or housing with soil as floor material, or household crowdedness.

Key Words:

census cluster housing spatial statistics

While the need for housing is widespread, people have different individual requirements — depending on whether they are single, married, senior citizens, families with children, or if the household includes members with disabilities. Despite the best of intentions of policy makers, having "a roof overhead" remains an elusive goal for a large majority of the world's people. Most households cannot afford even the cheapest house that matches their needs and qualifies as "suitable," and no government can close this gap with subsidies alone. Nor is the formal housing sector on track to build the 300 million new units needed to close the world's housing gap by year 2030. What can be done to face this issue?

We believe it is possible to use geospatial based tools to inform social housing policies to improve the (i) population targeting; (ii) policy-mix; and (iii) potential for leveraging private resources.

STRATEGY

In order to show the advantages of our proposed methodology we calibrated housing programs in Lima (Peru). The methodology we developed and then applied is capable of (i) characterizing the housing needs of each family depending on their employment condition and situation in the life cycle; and (ii) clustering families according to their capacity of paying for housing. This methodology could be replicated in any country where a recent national population and housing census has been conducted, and could significantly reduce the costs of housing policy design.

Based on Peru's 2007 National Population and Housing Census and on its Province and District Poverty Maps, we built two databases: one on vulnerability and the other on payment capacity.

- The database on vulnerability for Lima has 90 physical attributes for 71,257 blocks (with more than 5 dwellings) and 128 population characteristics for 91,875 blocks. Even if some attributes are duplicated, there is nevertheless around 200 workable attributes. The city block is the smallest area surrounded by streets, which can have or not buildings for housing or commerce.
- The database on payment capability has synthetic information for household income by block, as aggregation of individual household income for Lima. This income was imputed by the Statistical Office in the document, "2007 Poverty Map by Province and District" (*Mapa de Pobreza Provincial y Distrital 2007*) published in February 2009 by Peru's National Institute of Statistics. This data has limited use for assessing household payment capability because it is unable to reproduce the characteristic asymmetry of income distribution.

CHARACTERIZING HOUSING NEEDS

In poverty surveys, the module on food and beverage acquisition provides the caloric intake by household, which is the main input for estimating the monetary value of the minimum caloric intake and thus determining the extreme poverty line. Subsequently, other non-caloric items are included to provide a monetary threshold for the total poverty line. Such questions are never asked in a census, but a proxy can be assessed with a combination of variables usually related with poverty; in order to build a social vulnerability index.

From all attributes found in the two databases previously mentioned, a selection of characteristics of households was made based on a methodology developed in year 2005 by the National Institute of Statistics of Peru. This methodology was created for a nascent conditional cash transfer program, which needed as an initial data input a sample of population in poverty. Traditional poverty surveys focus in food and beverage consumption as well as other expenditures, with a cost of $150 per questionnaire. However, given a budget restriction of $10 per questionnaire, a minimum subset of question was selected in order to optimize poverty measurement.

The method aggregated household surveys from 2001 – 2005 (around 100,000 questionnaires) and truncated them for rural areas (around 30,000 questionnaires) and established the poverty condition as a binary variable. Several variables were then tried in order to obtain the maximum explanatory power. These explanatory variables were also made binary. For example, education variables show several levels of schooling achievement but a new binary variable created was, "adult with primary education (yes/no)". This strategy was successful to identify the poor and later evaluations have shown "negative filtration"; meaning exclusion of poor instead of the typical filtration of identification of non-poor. This methodology has been successfully applied to identify poor households for *Juntos*, Peru's national cash-transfer program.

Following the aforementioned methodology and given the closed set of total variables available, a first set of 11 variables were chosen as the ones usually included to establish the condition of poverty/non-poverty:

1) Composition of the dwelling's floor.
2) Characteristics of the dwelling's drainage system.
3) Legal status of the dwelling's plot tenure.
4) Household crowding (persons per bedroom).
5) Household women's total number of births.
6) Household members' age structure.

7) Household members' schooling achievement.
8) Household members' employment status.
9) Number of household adults who have a government issued ID.
10) Mother tongue of each household member (over the age of 5 years).
11) Household total number of members.

For each of the eleven variables, the information was provided as number of units with the characteristic of the variable, where units could be: 1) dwelling, 2) household, or 3) person.

As an example for the first variable, composition of the dwelling's floor, the data shows the number of dwellings with floor made up of: 1) soil, 2) concrete, 3) slate tiles, 4) polished wood or parquet, 5) rough wood, 6) asphalt tiles, and 7) other. In order to dichotomize this specific variable, factor analysis was applied for the 7 variables that describe the floor characteristics. The customary output is made by the number of factors that explain the input variables given a threshold, and the correlation of each input variable with each output factor. The result show that "soil" was the only variable with opposite sign to the other six variables, therefore a percentage variable was built for each block measuring the percentage of dwellings with soil floor.

A similar process was follow for the nine ordinal variables and the results were:

1) Percentage of dwellings with soil floor.
2) Percentage of dwellings with non-piped drainage.
3) Percentage of dwellings with no legal land tenure (squatters).
4) Percentage of households with more than 3 children.
5) Percentage of household with members under 65 years old.
6) Percentage of household with members who achieved only primary schooling.
7) Percentage of household members who work in businesses with less than 10 employees.
8) Percentage of household adults who lack government issued ID.
9) Percentage of household members whose mother tongue is an indigenous language.

For the variables of crowdedness and family size, an adjustment was made to re-scale the variable between 0 and 1. For crowdedness, the maximum found was 9 persons per bedroom; therefore the variable was divided by 9. For family size the maximum number was 17 persons; therefore the variable was divided by 17.

The eleven variables mentioned were the input for a cluster analysis with medians in order to avoid the conventional problem of large outliers. The arbitrary number of clusters was established at 10, against the result of the Calinski and Harabasz pseudo-F index and the Duda–

Hart Je(2)/Je(1) index of a maximum of 2 clusters. The main reason for this rejection is that the results from a hierarchical cluster analysis where dendograms that showed successive first branches made of 1 single block.

The practical reason to choose 10 clusters is to facilitate further analysis, but these do not imply 10 groups of equal size. The raw results are in Table 1

CLUSTERING HOUSEHOLDS ACCORDING TO PAYMENT CAPACITY

As mentioned before, Peru's Statistical Agency has imputed income for each household, but information available is only by city block. The method is based on a universe of 156 explanatory variables, and from these a subset with the best ones is selected for each of the 25 geopolitical regions in Peru. One of these is Metropolitan Lima and surrounding areas and our analysis will be based in this subset.

The first step based on income is to re-order previous tables based on the median income by block. The result is in Table 2.

Comparing this Table with the previous one, the top income cluster is made of 5,825 blocks (8.42 % of total) and the bottom income cluster is made of 7,613 blocks (11.00 % of total blocks). The gap between the top and the bottom clusters is 4:1.

A limitation of income values in Table 2 is that it includes monetary income and imputed income. For example, a squatter household does not pay for housing, but international practice recommends creating a hedonic value for housing. This hedonic value is included in the household's total income.

In order to evaluate affordability on the mortgage market, it is important to discriminate between households with a hedonic value for housing (and thus the dwelling has a market value), from those households who inhabit a house that lacks market value due to its squatter origin.

IDENTIFYING POTENTIAL SCHEMES TO INCREASE AFFORDABILITY

With the previous result a new table is made with the variables related to housing from the 11 variables identified for the Social Vulnerability Index (see Table 3)

Observing the table, targeting becomes more evident. For cluster 10, 83% of households declare themselves as squatters and lacking legal documents from land registry. A government plan targeting juridical security must take account of the blocks included in this specific cluster.

For clusters 9 and 10, 100% of households declare not having piped drainage. These 2 clusters represent 31% of all the blocks. A government plan targeting the attainment of piped drainage for every household can know the location of this need as well as the total surface and total households that need it; which allows a more accurate budgeting than usual.

Another target, highly related to health conditions, is when floor material is soil. Focusing on clusters 7, 9, and 10, the aggregate percentage of these three clusters is 47%. A government plan focused on this need may know in advance the total surface of new floor to be built and at the same time the income structure of the beneficiaries; taking in account that the income gap between the top 10% and the bottom 10% of these clusters is 3:1.

Finally, on crowdedness it is difficult to envision a government plan due to the aggregate nature of blocks. The gap between cluster 1 and cluster 10 is barely 2:1, which hides the gaps inside each block.

CONCLUSION AND FURTHER ANALYSIS

As a conclusion, a government plan trying to calibrate a housing program should have access to census information at granular level in order to avoid noise cancellation, due to an aggregation process at block level.

A cluster analysis should be made at household level, instead of as it is done in this presentation, which is at block level, in order to obtain more robust clusters of households. Also a typology of each dwelling based in materials of floor, walls and roofing plus access to utilities should be made at a household level, in order to value the dwelling where the household inhabits.

This synthetic measurement of wealth could be more precise than a synthetic measurement of income, because we have statistical noise for income from the poverty survey, which is added to noise from econometric modelling. In fact, the measurement of wealth will have the classical error of any "Delphic" model (judgmental approach).

REFERENCES

- Betti, Gianni and Lemmi, Achille (2008). *Advances on Income Inequality and Concentration Measures*. New York, NY: Routledge
- Cameron, A. Colin and Trivedi, Pravin K. (2010) *Microeconometrics using Stata*. College Station, TX: Stata Press
- D'Ercole, Robert (2010) *Housing and Population data by block (shape file).* Retrieved from www.box.com

- Salvedra, Wiemer et. al. (2011) *The Oxford Handbook of Economic Inequality* (editors). New York, NY: Oxford University Press
- Stata Corp LP. *Stata User's Guide – Release 14.* (2015) Stata Press. PDF retrieved from www.stata.com

Table 1

Cluster	**Number of people**	**Number of households**	**Number of blocks**	**Block % by cluster**
1	482,980	124,593	4,444	6.42
2	782,210	208,588	5,825	8.42
3	896,217	227,827	5,463	7.90
4	1’542,537	362,586	8,483	12.26
5	1’330,010	300,157	9,108	13.16
6	862,718	193,396	6,471	9.35
7	396,842	89,356	4,897	7.08
8	246,405	59,135	2,791	4.03
9	757,087	191,403	14,098	20.37
10	349,747	96,329	7,613	11.00
Total	7’646,753	1’853,370	69,193	100.00

Table 2

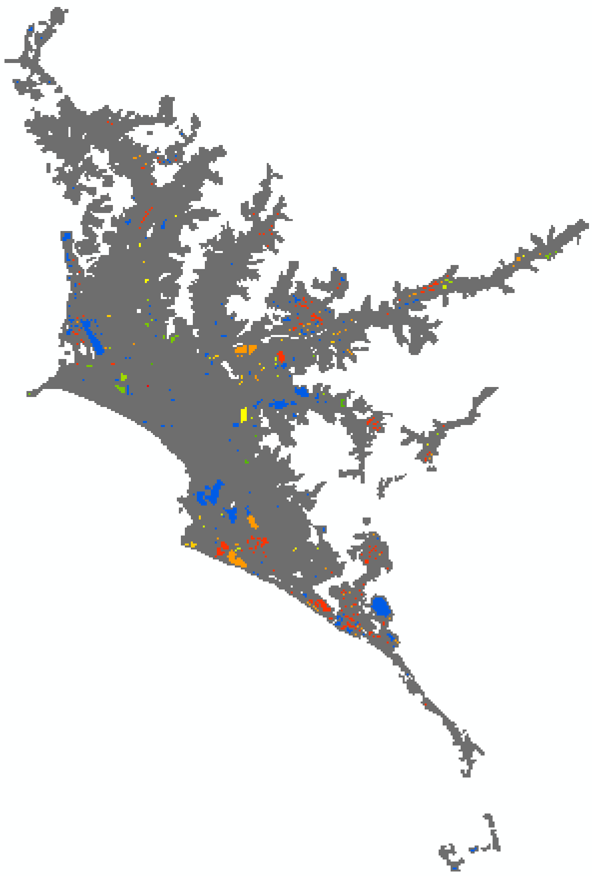
Cluster	**Median household income** (PEN)	**Total number of blocks**	**Cluster percentage**
1	2,079	4,444	6.42
2	2,213	5,825	8.42
3	1,298	5,463	7.90
4	1,171	8,483	12.26
5	1,017	9,108	13.16
6	971	6,471	9.35
7	760	4,897	7.08
8	814	2,791	4.03
9	662	14,098	20.37
10	564	7,613	11.00
Total	962	69,193	100.00

Table 3

Cluster	**Median household income**	**Number of people per bedroom**	**Floor material: soil**	**Not piped drainage**	**Land tenure by squatting**
2	2,213	.84	.00	.00	.00
1	2,079	.90	.00	.00	.00
3	1,298	1.25	.00	.00	.00
4	1,171	1.36	.03	.00	.00
5	1,017	1.58	.07	.00	.00
6	971	1.59	.20	.00	.00
8	814	1.64	.25	.50	.00
7	760	1.82	.42	.06	.00
9	662	1.78	.43	1.00	.00
10	564	1.94	.58	1.00	.83
Total	962	1.5	.13	.03	0.00

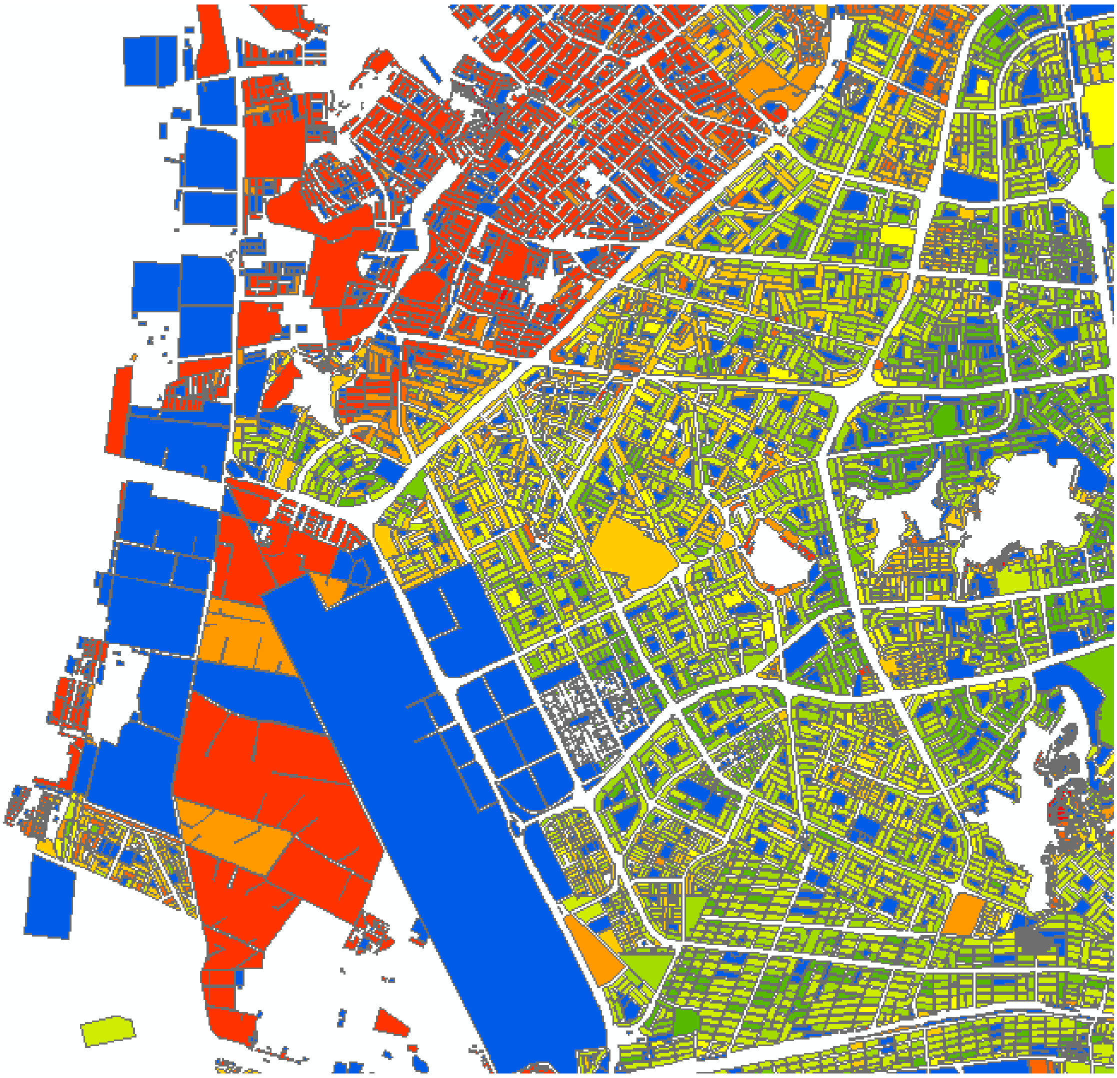
FIGURES

Metropolitan Lima (1:1'000,000)



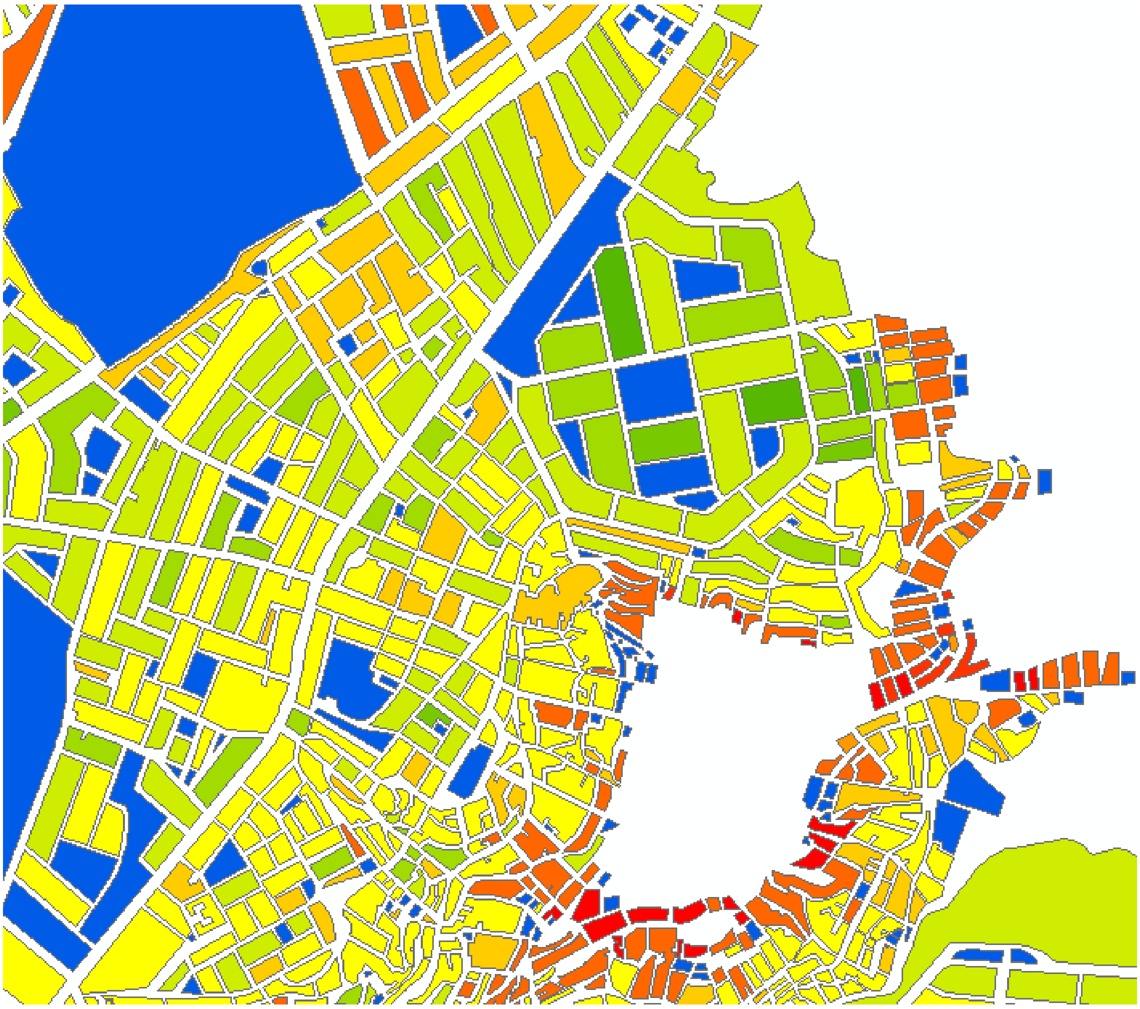
Metropolitan Lima (1:50,000)

Pacific Ocean (in white) is on the right side, next (in blue) storage buildings and the airport as a large rectangle. In upper right section, blocks (in red) made from squatting public land. In left section, former farm land developed as housing area.



Metropolitan Lima (1:10,000)

Old Cemetery (in blue) is at top left corner. Bottom center (in white) un-occupied hill due to steepness. Around hill (in red) recent squatting, circled by older squatters (in yellow). In both cases observe block shape.



Metropolitan Lima (1:25,000)

A hill (in white) split high income (in green) and low income (in red) households. In upper section is La Molina municipality and in lower section is San Juan de Miraflores municipality.