Land Governance in an Interconnected World
ANNUAL WORLD BANK CONFERENCE ON LAND AND POVERTY
WASHINGTON DC, MARCH 19-23, 2018

A PESSIMISTIC IS AN OPTIMIST THAT KNOWS GEOGRAPHY

Farid Matuk, Luis Triveño

29x55, World Bank

Presenting Farid.Matuk@29x55.com

Paper prepared for presentation at the

"2018 WORLD BANK CONFERENCE ON LAND AND POVERTY"

The World Bank - Washington DC, March 19-23, 2018

DATA

With information concerning 91,875 georeferenced blocks in the city of Lima obtained from the 2007 Population and Housing Census of Peru, 11 indicators of social vulnerability have been built by the authors, for each of the blocks of the city that have more than 15 inhabitants and that have more than 100 square meters, which in total corresponds to 71,139 blocks.

On the other hand, there is another database where the average income per household at the block level has been imputed using information from household surveys for year 2007, but 1,730 of them have zero income and therefore have been excluded. Therefore, the total available blocks with complete information are 69,409.

MODEL

The model to be estimated in this paper seeks to identify which of the social vulnerability indicators has the greatest impact on household income, but at the same time to take into account the spatial contiguity of the data and to correct the spatial heteroscedasticity in the term of disturbance. The theoretical model is as follows:

$$Y = X \boldsymbol{\beta} + \boldsymbol{\lambda} W y + u$$
$$u = \boldsymbol{\rho} M u + \varepsilon$$
$$E(X'u) = 0$$
$$\varepsilon \text{ i.i.d. } N(0, \sigma)$$

Where Y is the income, X is the vulnerability indexes, W is the spatial contiguity matrix, and M is the spatial auto-correlation matrix. Elements of matrices W and M are made with the distance in meters between observations. Greek letters are parameters to be estimated.

ANALYSIS

The computation program used for the previously described specification is Stata, which has as intrinsic limitation that the square matrices attain a maximum dimension of 11,000. Therefore, it is necessary to reduce the information of 69,409 blocks to a smaller number of analytical units that allow the W and M matrices to be constructed.

Taking as an area of analysis the accumulated of the 69,409 blocks of the city of Lima, with the program ARCgis we proceeded to create two grids, one of 700 meters by 700 meters and another one of 400 meters by 400 meters, where each of them was awarded the average of the values of the income and the vulnerability indexes of the blocks they covered.

In the first case 2,362 cells were obtained and in the second case 6,307 cells. Images of 5 income quintiles of these cells are found in Annex 1 and Annex 2 respectively, being pertinent to note that larger cells tend to dilute differences that smaller cells reveal.

The conventional econometric model (equivalent to λ and ρ equal to zero) does not take into account the spatial location of the observations to be analyzed, because in the vast majority of cases it was not available. On the other hand, in cases where uncorrected spatial heteroscedasticity existed, a bias was generated that gave relevance to variables that in fact did not have it due to the underestimation of the variance of the perturbations.

As indicated above, the variable to be explained is the average income of households in the city of Lima in 700x700 cells and 400x400 cells. In the following chart we have the descriptive characteristics for both cases.

Household Income (700x700 cells)

	Percentiles	Smallest		
1%	115.8367	28.13467		
5%	264.1895	34.83987		
10%	364.2862	36.25323	Obs	2,362
25%	518.9028	43.69167	Sum of Wgt.	2,362
50%	742.6931		Mean	879.7915
		Largest	Std. Dev.	575.6456
75%	1029.224	3853.326		
90%	1647.806	4039.176	Variance	331367.9
95%	2155.435	4066.877	Skewness	1.957366
99%	2995.737	4856.371	Kurtosis	8.143215

Household Income (400x400 cells)

	Percentiles	Smallest		
1%	128.859	12.61108		
5%	282.2675	19.14		
10%	381.3365	19.84849	Obs	6,307
25%	543.3834	29.77273	Sum of Wgt.	6,307
50%	783.2143		Mean	928.1216
		Largest	Std. Dev.	614.6939
75%	1079.239	4856.371		
90%	1707.945	4889.887	Variance	377848.6
95%	2209.305	5052.638	Skewness	2.119458
99%	3259.475	6334.344	Kurtosis	9.527089

As can be seen, for the 400x400 cells the average value and the standard deviation is higher than in the case of 700x700 cells, because the average of blocks per cell is lower. While in the cells of 400x400 cells the average is 21 blocks per cell, in the other case it is 49 blocks per cell. These differences are also observed in the images in Annexes 1 and 2.

The independent variables, obtained from the information by blocks of the 2007 Census, have been standardized between 0 and 1, in order to eliminate the effects of scale between the different natures of the variables. These variables are:

1. Soil floor (% of households)
2. No access to public drainage network (% of households)
3. Land tenure due to invasion / squatters (% of households)
4. Overcrowding – more than 3 persons per room (% of households)
5. Total number of children born alive
6. Household members older than 65 years of age (% of households)
7. Household members without elementary schooling (% of households)
8. Household members holding jobs in businesses with less than 10 persons (% of households)
9. Adult household members lacking national document of identity (% of households)
10. Household with members having a native tongue as first language (% of households)
11. Total number of household members

After analyzing the explanatory power of the 11 variables previously described with respect to income (in fact, to the logarithm of income), it was found that the variables with the best adjustment are overcrowding (SLUM_I), lack of schooling (BASIC_E), and the number of person per household (M_H). The latter has a positive sign because it is evident that the greater the number of people, the greater the collective income of the household; on the contrary, overcrowding and lack of schooling have both an accordingly negative sign.

A detail to underscore is that the R2 is intrinsically higher in the 700x700 cells cells because the number of elements of the average is greater than that of the 400x400 cells cells (49:21). This has been already highlighted by Cochran in the 1960s, who indicated that the averaged data reduce the variance of the original data. In our specific case, the ratio of the standard deviations is 576: 615.

When comparing the estimated coefficients of the independent variables, it can also be observed that there is no great difference between cells of 700x700 cells and cells of 400x400 cells.

OLS Income Logarithm (700x700 cells)

```
. reg LN_ING SLUM_I BASIC_E M_H

      Source |       SS           df       MS      Number of obs   =     2,362
-------------+----------------------------------   F(3, 2358)      =   1007.86
       Model |  515.812547         3  171.937516   Prob > F        =    0.0000
    Residual |   402.26544     2,358  .170596031   R-squared       =    0.5618
-------------+----------------------------------   Adj R-squared   =    0.5613
       Total |  918.077987     2,361  .388851329   Root MSE        =    .41303

------------------------------------------------------------------------------
      LN_ING |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      SLUM_I |  -2.099383   .1354085   -15.50   0.000    -2.364915   -1.833851
     BASIC_E |  -2.831216   .1990568   -14.22   0.000     -3.22156   -2.440871
         M_H |   12.13972   .2323625    52.24   0.000     11.68406    12.59537
       _cons |   5.928479   .0304564   194.65   0.000     5.868755    5.988203
------------------------------------------------------------------------------
```

OLS Income Logarithm (400x400 cells)

```
. reg LN_ING SLUM_I BASIC_E M_H

      Source |       SS           df       MS      Number of obs   =     6,307
-------------+----------------------------------   F(3, 6303)      =   2284.59
       Model |  1264.87707         3  421.625688   Prob > F        =    0.0000
    Residual |  1163.23012     6,303  .184551819   R-squared       =    0.5209
-------------+----------------------------------   Adj R-squared   =    0.5207
       Total |  2428.10718     6,306  .385047127   Root MSE        =    .42959

------------------------------------------------------------------------------
      LN_ING |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      SLUM_I |  -2.555403   .0877861   -29.11   0.000    -2.727494   -2.383313
     BASIC_E |  -2.240231   .1055303   -21.23   0.000    -2.447107   -2.033356
         M_H |   11.18741   .1418824    78.85   0.000     10.90927    11.46554
       _cons |   6.007641   .0188351   318.96   0.000     5.970718    6.044564
------------------------------------------------------------------------------
```

When the same routine is carried out taking into account the spatial location of the informants, there are substantive differences between estimates by Ordinary Least Squares (OLS) and estimates by Maximum Likelihood (ML), at absolute value of the estimated coefficients, as well as differences of estimated coefficients between 700x700 cells and 400x400 cells. When comparing the logarithm of likelihood between OLS and ML, it is found that in ML it is five times less than in OLS. The results are:

Spatial ML Income Logarithm (700x700 cells)

```
Spatial autoregressive model                    Number of obs     =       2362
(Maximum likelihood estimates)                  Wald chi2(3)      =    2681.07
                                                Prob > chi2       =     0.0000

------------------------------------------------------------------------------
      LN_ING |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LN_ING       |
      SLUM_I |  -1.033099   .1070407    -9.65   0.000    -1.242895   -.8233034
     BASIC_E |  -1.239203   .1536833    -8.06   0.000    -1.540417   -.9379889
         M_H |    10.3795   .2344058    44.28   0.000     9.920072    10.83893
       _cons |   5.367852   .0363922   147.50   0.000     5.296525    5.439179
-------------+----------------------------------------------------------------
lambda       |
       _cons |   .0232695   .0088371     2.63   0.008     .0059491    .0405899
-------------+----------------------------------------------------------------
rho          |
       _cons |   .9458222   .0110187    85.84   0.000     .9242259    .9674186
-------------+----------------------------------------------------------------
sigma2       |
       _cons |   .0615977   .0018377    33.52   0.000     .0579959    .0651995
------------------------------------------------------------------------------
```

Spatial ML Income Logarithm (400x400 cells)

```
Spatial autoregressive model                      Number of obs   =      6307
(Maximum likelihood estimates)                    Wald chi2(3)    =   6410.62
                                                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      LN_ING |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LN_ING       |
      SLUM_I |  -1.041058   .0723413   -14.39   0.000    -1.182845   -.8992721
|    BASIC_E |  -.5661883   .0826067    -6.85   0.000    -.7280945   -.4042822
         M_H |   8.925908   .1383623    64.51   0.000     8.654723    9.197093
       _cons |   5.353559   .0233147   229.62   0.000     5.307863    5.399255
-------------+----------------------------------------------------------------
lambda       |
       _cons |   .0231368   .0058056     3.99   0.000     .0117581    .0345156
-------------+----------------------------------------------------------------
rho          |
       _cons |   .9563739   .0061457   155.62   0.000     .9443286    .9684192
-------------+----------------------------------------------------------------
sigma2       |
       _cons |   .0611091   .0011161    54.75   0.000     .0589215    .0632967
------------------------------------------------------------------------------
```

In the following table all the coefficients of the four previous regressions can be seen:

Cells	**OLS**		**Spatial ML**	
	700x700 cells	**400x400 cells**	**700x700 cells**	**400x400 cells**
Household overcrowding	-2.099383	-2.555403	-1.033099	-1.041058
Lack of schooling	-2.831216	-2.240231	-1.239203	-.5661883
Household size	12.13972	11.18741	10.37950	8.925908
Intercept	5.928479	6.007641	5.367852	5.353559
Likelihood logarithm	-1260.9835	-3618.3816	-253.40335	-685.03308

CONCLUSIONS

A first conclusion obtained is the absolute value of the coefficients estimated with ML is lower than coefficients estimated with OLS. This implies that the effectiveness of a policy to provide better income to households through housing plans to reduce overcrowding and with better primary education programs will have less impact when taking account on the spatial location of households.

A second conclusion is coefficient of impact of the lack of education at the income level is reduced by half when we go from estimating 700x700 cells to 400x400 cells. This is because the variable BASIC_E is distributed normally for 700x700 cells whereas it is not distributed normally in 400x400 cells, as can be seen in Annex 3.

A third conclusion is that it is convenient to specify that the reduction in the number of observations to be analyzed, due to the limitations of the computer program, has severe consequences when evaluating the impacts of the public policies to be executed.

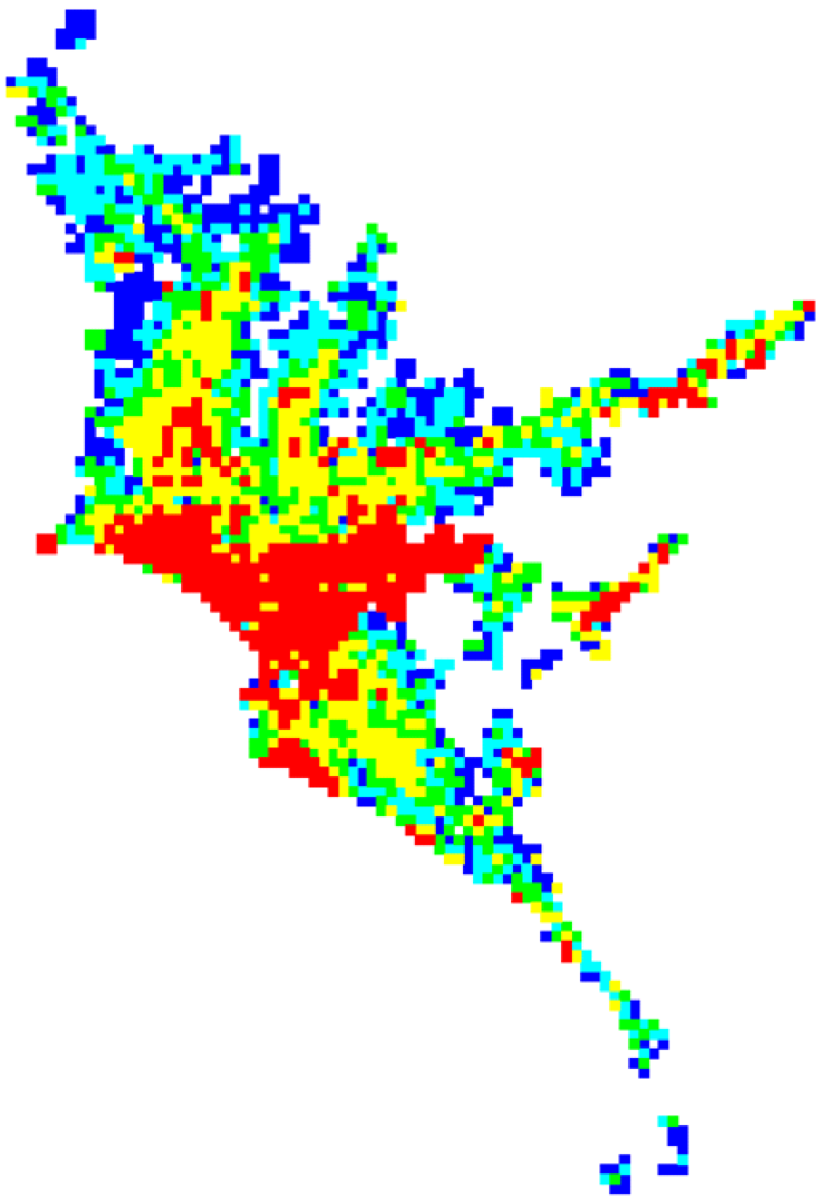
Finally, a fourth conclusion from 11 variables analyzed as social vulnerability component, only 2 variables are significant for the income level:

1. Household overcrowding
2. Lack of elementary schooling

Annex 1

Income by Quintile (Cells 700x700)

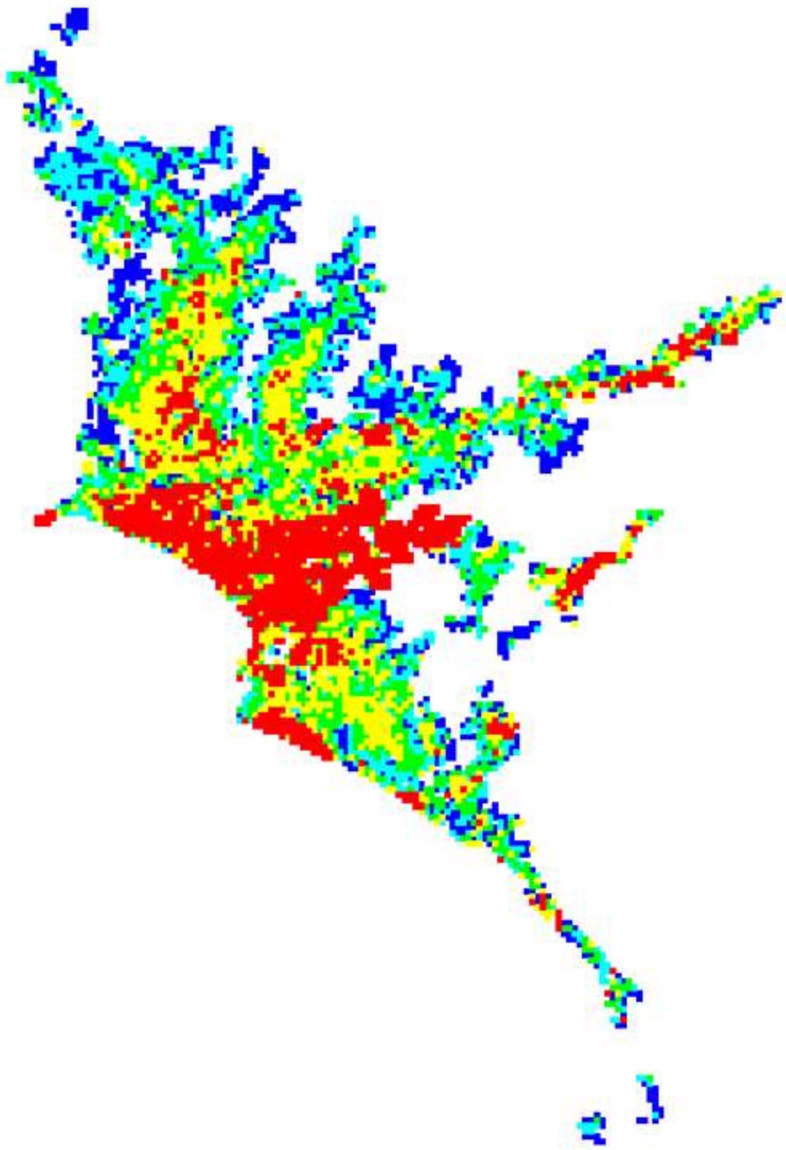
Red = High Yellow=High-Med Green=Medium Turquoise=Med-Low Blue=Low



Annex 2

Income by Quintile (Cells 400x400)

Red = High Yellow=High-Med Green=Medium Turquoise=Med-Low Blue=Low



Annex 3

Normal distribution test for Basic Education (BASIC_E) in 700x700 cells

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	 Prob>chi2
BASIC_E	**2,362**	**0.8554**	**0.2029**	**1.65**	**0.4373**

Normal distribution test for Basic Education (BASIC_E) in 400x400 cells

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	 Prob>chi2
BASIC_E	**6,307**	**0.0000**	**0.0000**	.	**0.0000**

Basic Education variable histogram

BASIC_E Histogram

700x700 cells	400x400 cells