2021, Volume 44, Issue 87, 115-145 / ISSN 2304-4306

ECONOMÍA

revistas.pucp.edu.pe/economia

FONDO EDITORIAL

www.fondoeditorial.pucp.edu.pe

Human Capital Constraints, Spatial Dependence, and Regionalization in Bolivia: A Spatial Clustering Approach

Carlos Mendez[a,*], Erick Gonzales[b]

[a]Graduate School of International Development, Nagoya University, Japan
✉ carlos@gsid.nagoya-u.ac.jp [*]Corresponding author

[b]United Nations Agency for Disaster Risk Reduction, Kobe, Japan
✉ erick.gonzalesrocha@un.org

Abstract

Using a novel dataset, this article studies the spatial distribution of human capital constraints across 339 municipalities in Bolivia. In particular, five human capital constraints are evaluated: chronic malnutrition in children, non-Spanish speaking population, secondary dropout rate of males, secondary dropout rates of females, and inequality in years of education. Through the lens of principal components, spatial dependence, and regionalization methods, the municipalities are endogenously classified according to their similarity in human capital constraints and geographical location. Results from the spatial dependence analysis indicate the specific location of significant hot spots (high-value clusters) and cold spots (low-value clusters). A regionalization analysis of the constraints indicates that Bolivia can be regionalized into seven or eight geographical regions. The article concludes discussing the potential complementary of these two analyses and their usefulness in identifying the location of policy priorities.

Article History: Received: July 20 2020 / Revised: September 08 2020 / Accepted: March 16 2021
Keywords: Human capital; Spatial dependence; Regionalization; Cluster analysis; Bolivia
JEL Classification: C38, J24, R10

Acknowledgements

We would like to thank the anonymous referees for their thoughtful comments and suggestions, which have significantly improved the manuscript. We also acknowledge the comments and suggestions from the members of the QuaRCS-lab (https://quarcs-lab.org).
The findings, interpretations, and conclusions expressed in this article are entirely those of the authors. They do not necessarily represent the view of their institutions. Their institutions do not guarantee the accuracy of the data included in this work.

https://doi.org/10.18800/economia.202101.007

1. Introduction

Human capital is central for understanding individual earnings, inequality, and economic growth (Becker et al., 1990; Barro, 2001; Gemmell, 1996; Collin and Weil, 2020; Mincer, 1984; Psacharopoulos and Patrinos, 2018). While several low-and-middle income countries have made progress in terms of school enrollment and attainment, the focus needs to turn towards closing the gap in terms of quality type issues such as cognitive skills, institutions, and exposure to better environments, among others (Hanushek, 2013; Hanushek and Woessmann, 2008; Chetty et al., 2016; Pritchett, 2001; Psacharopoulos and Patrinos, 2018). Bolivia is not apart from these dynamics. A considerable number of studies have shed light on topics such as returns to education, migration, gender, language, and ethnicity (Kelley, 1988; Godoy et al., 2005; Psacharopoulos, 1993; Patrinos and Psacharopoulos, 1993; Chiswick et al., 2000; Martínez, 1990). However, there is less evidence on specific sub-national constraints hindering the accumulation of national human capital.

In this article, we apply principal components, spatial dependence, and regionalization methods to identify clusters of municipalities facing similar human capital constraints. Specifically, using the novel dataset of SDSN-Bolivia (2020), we evaluate the spatial distribution of chronic malnutrition in children, non-Spanish speaking population, secondary dropout rate of males, secondary dropout rates of females, and inequality in the years of education. We first use a principal component analysis (PCA) to organize the variables into two components. The first component (PC1) mostly summarizes the regional variation in malnutrition in children, inequality of education, and non-Spanish speaking; while the second component (PC2) summarizes variation in the dropout rates.

Next, we use two spatial frameworks to identify geographically contiguous clusters in each component. The classical spatial dependence framework of Anselin (1995) help us identify regional hot spots (high-value clusters), cold spots (low-value clusters), and spatial outliers. The more recent regionalization framework of Duque et al. (2012) help us design a new map of Bolivia in which regional boundaries are endogenously derived from differences in human capital constraints.

The main results are as follows. Human capital constraints (PC1 and PC2) show a positive and statistically significant degree of spatial dependence across municipalities. Based on PC1 data, there is a large cluster near the center of Bolivia which is characterized by high malnutrition of children, education inequality, and non-Spanish speaking population. Based on PC2 data, there is another large cluster in the north of the country which suffers from high dropout rates of secondary education. The regionalization analysis, based on PC1 (PC2) data, indicates that Bolivia can be divided into eight (seven) geographical regions facing similar constraints. The borders of these newly identified regions largely differ from those indicated by the administrative map of the country. Thus, the design and monitoring of human development policies need to be coordinated across multiple local administrations.

The contribution of this article is three-fold. First, it considers the role of locational similarity as a binding restriction in the identification of regional clusters. Second, the number of clusters does not have to be set a priori. It is an endogenous result of the analysis. These two features differentiate our article from other studies, which typically use non-spatial clustering methods and need to define the number of clusters in advance.[1] Lastly, the paper explores the potential complementary of spatial dependence (local Moran clusters) and regionalization (Max-p clusters) as a unified framework for identifying robust, endogenous, and spatially contiguous clusters. The identification of such clusters is important as it could help both national and regional governments prioritize places where human development policies are more needed.

The rest of this article is organized as follows. Section 2 presents a survey of related literature. Section 3 describes the dataset and Section 4 introduces the methods of principal components analysis, spatial dependence, and regionalization. Section 5 presents the results for each of these methods. Section 6 discusses the potential complementary of the spatial analyses. Finally, Section 7 offers some concluding remarks.

2. Related Literature

2.1 Human Capital Constraints in Bolivia

For Bolivia, there are several studies shedding light on education and earnings. For example, Kelley (1988) used data collected in 1966 (a decade after the 1952 revolution) and concluded that 95 to 100 percent of differences on income are due to class components (family background, individual education, and occupation) and not because of ethnic differences. Psacharopoulos (1993) published a study after the mid-80s Stabilization and Structural Adjustment Program in Bolivia. This study uses the 1989 household survey and found that indigenous workers received lower returns to schooling and work experience (the effect was less pronounced in younger cohorts who are more educated and earn more). Patrinos and Psacharopoulos (1993) used data from the 1989 *Encuesta de Hogares* (it covers urban centers and focuses on males) and suggested the existence of higher returns to schooling (8.6 percent) and labor market experience (4.5 percent) for non-indigenous than for indigenous population (5.7 percent and 2.7 percent, respectively).

In terms of income differences, Patrinos and Psacharopoulos (1993) suggested that 71.7 percent could be explained by productive characteristics of individuals. The unexplained remaining (28 percent) may include differences in ability, quality of education, culture, or discrimination. Lower earnings for indigenous citizens seem to be mainly due to lower human capital endowment. Even among foragers and horticulturalists in communities distanced from the nearest towns and cities of Bolivia, so-called primitive economies, Godoy et al. (2005) found positive correlations between human capital and economic outcomes such as income, consumption, or wages.[2]

Also on earnings, Chiswick et al. (2000) used a 1993 household survey in Bolivia (*Encuesta*

[1] Examples of these methods are K-means, hierarchical clustering, DBSCAN, among others. See Niembro and Sarmiento (2020) and the references there in for a recent application of some of these clustering methods in the context of regional development.

[2] This study tries to account for skills. People with better arithmetic skills had higher farm output (71.4 percent) and overall income (12.8 percent), in particular among those closer to market towns. Moreover, after controlling for both arithmetic and reading skills, an additional year of education was correlated with higher income (4.5 percent) and wages (5.9 percent).

Integrada de Hogares), conducted by the National Institute of Statistics, covering the capital cities from each nine departments and found that earnings raise with years of schooling for both men and women (around 6.5 percent). Other factors related to higher earnings are labor-market experience, being born in an urban area, and longer residence in the city (for migrants).

Nevertheless, these studies do not strictly account for the quality of education. Patrinos and Psacharopoulos (1993) underlined that, for example, the type of schools attended could make a significant difference for earnings determination. It should also be noted that some forms of discrimination, malnourishment in early childhood, Spanish-speaking ability, or inherent types of inequality in years of education, income, etc. (which could generally be identified as constraints) negatively affect access to schooling, good quality schooling, performance in the labor market, etc. This subsequently leads to lower levels of schooling, earnings, and poverty if the cycle is not broken.

Among these constraints, language, in particular, plays an important role in Bolivia.[3] For bilingual speakers in Bolivia, poorer proficiency in Spanish is penalized with lower earnings. Chiswick et al. (2000) found that monolingual Spanish speakers earn around 25 percent more than those who speak both Spanish and an indigenous language. At the other end, women who speak only an indigenous language earn around 25 percent less than bilingual speakers. Patrinos (1997) found a similar result in Guatemala (another country in Latin America with large percentages of indigenous citizens in their population) where earnings of Spanish speakers are higher than any of the indigenous groups. Patrinos and Psacharopoulos (1993) noted that citizens whose mother tongue is not Spanish have higher dropout rates in the primary grades, repeated more grades and were less likely to attend school. They also can experience limitations for speaking Spanish without an accent.

Another serious constraint to human development is posed by malnutrition. Miranda et al. (2020) used the 2008 Bolivian Demographic and Health Survey (DHS) to estimate the prevalence of malnutrition by wealth, ethnicity, and educational level. Their results suggested that lower levels of stunting or short stature among children less than five years old are significantly correlated with mother's years of education (particularly those with 7 to 12 years or more than 12 years of education).[4] Malnourished children will be less equipped to engage in more meaningful learning processes.

Cetrángolo et al. (2017) stated that for Latin America, education is a key component to reduce inequality, foster economic growth, and strengthen democracy. For example, equality of opportunity in the access to good quality education (less barriers to human capital development) is one basic step to reduce inequality in the mid to long term. This is particularly relevant for Bolivia because it has a high level of income inequality (confirmed by a GINI index of more than 40 according to data from the World Bank). In turn, societies with less barriers to the education

[3] Martínez (1990) indicated that despite being a multilingual country, in practice, Bolivia is largely dominated by a single language and culture. Hornberger (1992) pointed out that despite Spanish being the dominant language, there are more than 30 other languages, seven of which, at that time, were spoken by at least 10,000 people.

[4] The same statistically significant relationship was found for mother's level of education and the levels of stunting and short stature among women 11 to 19 years old.

of their citizens are better positioned to reap the benefits of technical progress, innovation, and productivity. This is also a key issue for Bolivia because most of the evidence suggests that diminishing differences in education could improve labor-market outcomes.[5] Finally, a strong democracy requires the political participation of citizens that are better informed, capable to question information with critical capacity, and displaying civic culture.[6] The recent turbulent times, if anything, highlighted the importance of these factors in allowing the country to engage in sustainable development instead of being born again and tumbling towards progress every couple of decades.

2.2 Regional Disparities in Bolivia

Spatial data analysis methods for the case of Bolivia are mainly focused on issues of convergence in economic growth, unsatisfied basic needs, or poverty. The evidence is not conclusive, but it suggest that departments in Bolivia converge during recessions and diverge during expansions (Sandoval, 2003; Cuervo, 2003). For example, dispersion and lack of convergence for the period 1976–1992 (Evia et al., 1990), some convergence in 1988–1992 (Morales et al., 2000), and divergence in 1993–1997 (Sandoval, 2003). While more recent studies covering longer periods of time such as Soruco (2012); Montero and del Río (2013); Mendieta Ossio (2019) observed limited spatial dependence among departments for economic growth, Vargas (2004) found that location does influence poverty levels in municipalities considering the type of resources that can be exploited as well as the flows of commerce that are enabled.

Mendez (2018a,b) focused on the regional distribution dynamics of the human development index and found that the formation and merging of several clusters can signal a reduction of inequality in human capital among metropolitan regions in Bolivia. For example, while the period 1992–2001 showed the existence of three clusters, the period 2001–2013 indicates the merge of the central cluster into the higher human capital cluster, suggesting forward mobility for some municipalities. However, there is also not so encouraging evidence when looking at the extremes of the distribution. Municipalities with the lowest levels of human capital are less likely to converge to higher equilibria in the long run. Conversely, municipalities with the highest levels of human capital appear to have some backward mobility.

Applying an exploratory spatial data analysis and spatial regression analysis, Delboy (2019) studied school attendance and presented some intuitive results such as higher levels of urbanization, labor market participation,[7] migration,[8] and child labor are significantly related to school attendance. Canelas and Niño-Zarazúa (2019) stated that short school days and lax legal frame-

[5] On the contrary, Maclsaac and Patrinos (1995), for example, suggested that a large portion of Peru's differences between indigenous and non-indigenous citizens are not explained by education or other observable factors.

[6] The current spreading of miss-information and serious cleavages (income, culture, etc.) render the need for better human capital to be more acute.

[7] This could be interpreted from the side of demand. Where there are more opportunities to find a job, and people indeed obtain formal employment, there might be incentives to engage in education. In other words, there is the expectation that education will pay-off.

[8] In a Latin American context, McKenzie and Rapoport (2011) also found a negative effect of migration both on attendance rates and attainment.

works may contribute to the finding of the Bureau of International Labor Affairs (2018), that about 15 percent of children between 7 to 14 years old engage in labor activities in Bolivia. The latter also noted that desertion rates dropped from 5 percent in 2006 to 2 percent in 2018, but secondary education attendance rates remain low in rural areas.

Information about possible constraints to human capital development in Bolivia is rather robust. However, studies evaluating the spatial distribution of those constraints are scarce and non-existent when it comes to identifying spatially contiguous clusters.[9] The literature suggests that returns to schooling (incomes, consumption, and wages) are largely influenced by human capital accumulation. Constraints to this accumulation include malnourishment, language ability, and inequalities (income, years of education, etc.), among others. There is a vacuum for understanding the distribution of those human capital constraints among municipalities and how this information can help in the identification of contiguous regions that may lead to cooperation in addressing shared challenges.

2.3 Regionalization and the Max-p Algorithm

Identifying geographically contiguous regions that share common features (demographics, economics, or politics) is important for regional planning and monitoring. Conceptually, this regionalization problem has been a topic of wide interest in the fields of statistics, quantitative geography, and machine learning (Duque et al., 2007; Law and Neira, 2019; Wise et al., 1997). According to Fischer (1980), a homogeneous region consist of a set of spatially contiguous areas which show a high degree of similarity regarding a set of attributes. In the context of this article, those attributes are chronic malnutrition in children, non-Spanish speaking population, secondary dropout rate of males, secondary dropout rates of females, and the Gini coefficient of years of education.

Regionalization, defined as a process of aggregating geographical areas into homogeneous regions, has been referred to by a large number of names, including conditional clustering (Lefkovitch, 1980), contiguity constrained clustering (Murtagh, 1992), clustering under connectivity constraints (Hansen et al., 2003), regional clustering (Maravalle and Simeone, 1995), regionalization (Wise et al., 1997), among others. As noted by Duque et al. (2007, 2011), regional scientists use spatial clustering methods not only for summarizing information or finding the real number of clusters, but as a means for designing suitable regions for analysis and monitoring.

Although, in many cases, the actual number of spatial clusters is unknown, some initial conditions or spatial constraints can be used to identify (endogenize) the number of clusters. Recently, Duque et al. (2012) have introduced a new method to endogenously identify the spatially constrained clusters. This method, known as the Max-p-regions problem, aggregates n geographical areas into an unknown maximum number p of homogeneous regions. Moreover, the method

[9] While regionalization analysis (max-p method in particular) has been applied to different areas such as statistics, geographic delimitation, public transport, urbanization, crime, etc. (see Arribas-Bel and Schmidt, 2013; Canavire-Bacarreza et al., 2016; Duque et al., 2013 and section 4.3), an extended review suggests that there are no studies applying the methodology in the field of human capital accumulation. For interesting spatial distribution studies on education in the Latin American region, though they do not use the max-p method, see Fujita et al. (2020); Elias and Rey (2011).

ensures that each aggregated region satisfies a minimum threshold value of a spatially extensive attribute such as the population per region, area per region, number of households per region, among others. The method is flexible and data-driven in the sense that it does not impose further constraints on the compactness of the regions; instead, it lets the data define the shape of each region.

A growing number of studies have used the Max-p approach to identify spatially contiguous regions that face similar challenges and opportunities. For instance, in the context of the Colombian municipalities, Church et al. (2020) identified spatial clusters based on industry-related variables and interactions. The authors argue that these clusters are particularly useful for designing innovation ecosystems. In the context of the Nigerian states, Lawal (2020) identified spatial clusters based on demographic, economic, and poverty characteristics. The author calls for a re-examination of the current regional design of Nigeria to ensure the formulation of development plans guided by evidence. The work of Rey and Sastré-Gutiérrez (2010) is one of the first studies to apply the Max-p regionalization scheme to study income inequality across states in Mexico. Their findings highlight the usefulness of this approach for understanding the spatial heterogeneity of regional inequality.

3. Data

3.1 A New Database to Study Regional Development in Bolivia

Data on human capital constrains are from the newly released Municipal Atlas of the Sustainable Development Goals in Bolivia (SDSN-Bolivia, 2020). From this novel database, the following five indicators are used:

1. **Chronic malnutrition in children:** This indicator measures the percentage of kids under five years with chronic malnutrition in the year 2016. This indicator is weighted by department and poverty level. The original source of the data is the Survey of Demography and Health of 2016 (*Encuesta de Demografía y Salud 2016*).

2. **Non-Spanish speaking population:** This indicator measures the percentage of the population, three years old or older, that do not have Spanish as their mother tongue, first or second language. The original source is the Census of Population and Housing 2012 (*Censo de Población y Vivienda 2012*).

3. **Secondary dropout rate of females:** This indicator measures the number of female students dropping out from secondary school as a percentage of matriculation. The original source is the Ministry of Education's Educational Statistics and Indicators System (*Sistema de Estadísticas e Indicadores Educativos 2017*).

4. **Secondary dropout rate of males:** This indicator measures the number of male students dropping out from secondary school as a percentage of matriculation. The original source is the Ministry of Education's Educational Statistics and Indicators System (*Sistema de Estadísticas e Indicadores Educativos 2017*).

5. **Gini coefficient of years of education:** This indicator measures the Gini coefficient, measuring inequality, in the years of schooling for the population in the segment of 25 to 65 years old. The original source are estimations by SDSN-Bolivia (2020) based on data from the Census of Population and Housing 2012 (*Censo de Población y Vivienda 2012*).

The indicators defined above were selected for two main reasons. One of them is that they represent indicators in the Municipal Atlas of the Sustainable Development Goals in Bolivia that are mainly related to SDG4 to "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (secondary dropout rate for females and males), SDG10 to "Reduce inequality within and among countries" (Non-Spanish speaking population and Gini coefficient of years of education), and SDG2 to "End hunger, achieve food security and improved nutrition and promote sustainable agriculture" (chronic malnutrition in children) because some of their fundamentals are related to human capital development constraints. The second reason is to select variables that conservatively aim to capture the studied concepts without overlapping.[10]

3.2 Descriptive Statistics and Maps

Table 1 provides an overview of the previously described indicators. Overall values are within expected ranges, but the summary also generates noteworthy observations. One of them is that a significant number of children in Bolivia experience chronic malnutrition.[11] The situation is particularly dire in municipalities where around half of all kids might be malnourished (several municipalities near the lower center of the country registered levels of 41, 49, and 53 percent). On the other hand, municipalities in the east have the lowest levels of malnutrition among children (8.5 percent). Another observation is that in some municipalities towards the center of Bolivia, more than 50 percent of the population does not have Spanish as their mother tongue.

For dropout rates, the dataset of SDSN-Bolivia (2020) provides a useful disaggregation by gender. For example, dropout rates for males are higher than those of females. It is possible that a higher dropout rate for males is due to cultural issues and/or economic need to enter the job market. Godoy et al. (2005) stated that among municipalities with communities further away from main towns, women are not usually expected to enter the market for wage labor, allowing them to stay longer or finish school.

Table 1
Descriptive statistics: human capital constraints.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max	Obs.
Chronic malnutrition in children (percent, 2016)	23.67	12.05	7.64	14.05	23.09	30.17	52.58	339
Non-Spanish speaking population (percent, 2012)	15.10	13.87	0.66	4.91	9.57	19.76	59.94	339
Secondary dropout rate (male percent, 2017)	5.00	2.88	0.00	3.15	4.69	6.45	21.15	339
Secondary dropout rate (female percent, 2017)	4.09	2.85	0.00	2.42	3.42	5.16	22.22	339
GINI coefficient of years of education (2012)	0.39	0.08	0.20	0.33	0.37	0.44	0.64	339

[10] The case of dropout rates makes use of the available disaggregation by females and males as it could enable more detailed interpretations.

[11] The municipalities are being weighted by levels of poverty, reducing the level of malnutrition in places where poverty is less prevalent.

The last indicator measures inequality in the years of education. An average value close to 0.4, indicates relatively high levels of inequality. While municipalities corresponding to several capital cities display more egalitarian distributions in the years of education (values between 0.21 and 0.25), municipalities in the lower center of the country experience severe inequality (values between 0.60 and 0.64). In the latter, while a few people have many years of education, the majority does not.

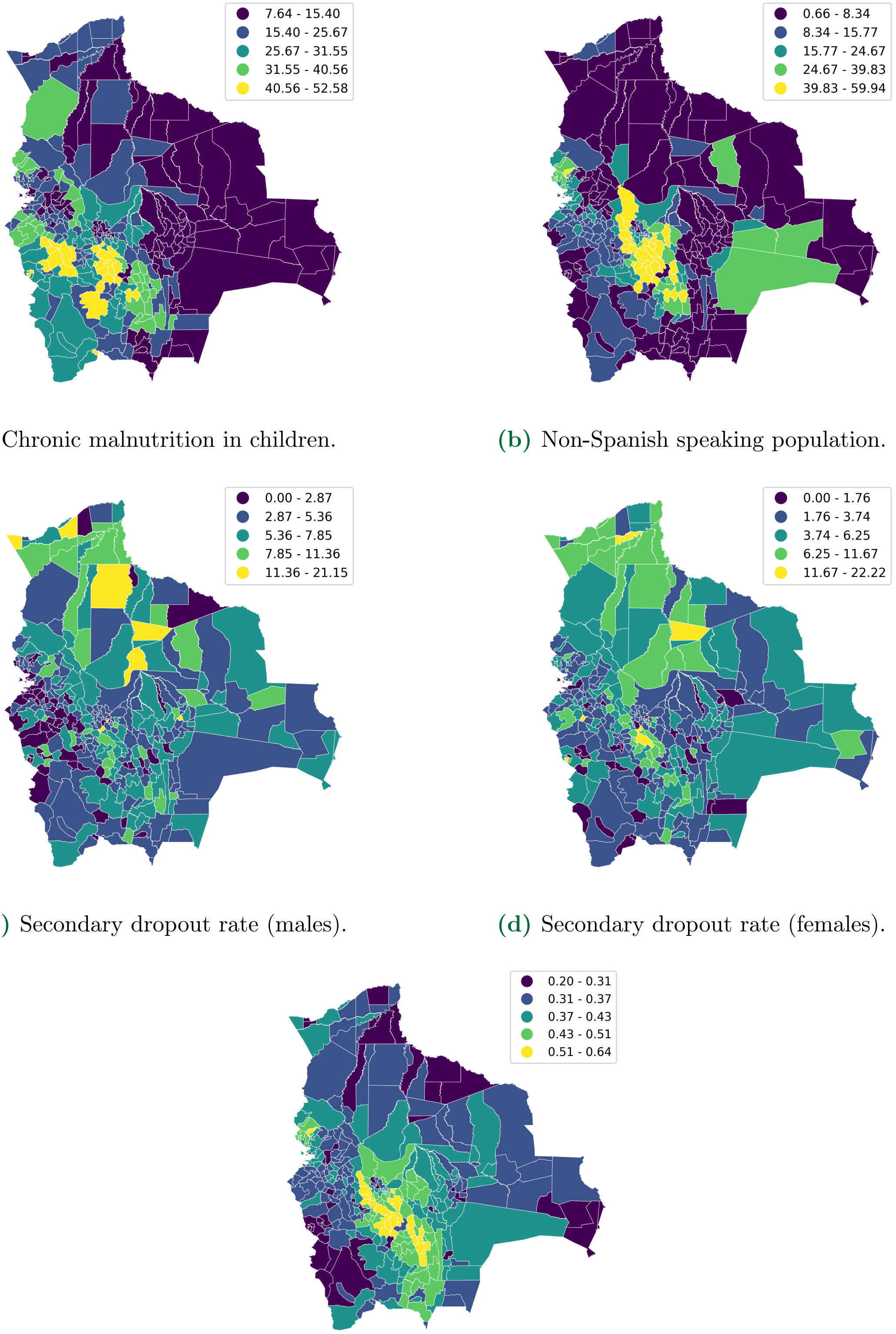
Figure 1 provides a first overview of the spatial distribution of each human capital constraint. Breaks for each choropleth map are optimally selected by using the natural breaks classification method of Fisher (1958) and Jenks (1977). This method applies a nonlinear algorithm to group regions in a way that maximizes within-group homogeneity. In essence, it is a one dimensional k-means clustering that finds groups with the largest similarity in the attribute being analyzed.

As a result, Figure 1 classifies municipalities into five groups ranging from lowest to highest values. Overall, for each human capital constraint, it appears that municipalities with high (low) values tend to be located near other municipalities with high (low) values. However, it is also clear that the identified clusters are not necessarily contiguous or spatially integrated. This is because the unidimensional clustering framework of Fisher (1958) and Jenks (1977) only maximizes attribute similarity without imposing any constraint on spatial contiguity. Another limitation is that the number of clusters is exogenous, that is, it has to be decided in advance. Motivated by these limitations, in the following section, we present the results of spatially integrated endogenous clusters.

Before moving to the next section, a brief interpretation of the five panels in Figure 1 can be two-fold. First, higher levels of malnourished children, non-Spanish speaking people in the population, and inequality in the years of education tend to overlap in municipalities towards the center of the country. Second, dropout rates tend to display a different distribution than the previous three variables showing higher levels of dropout rates towards the northeast of the country and lower levels towards the southwest.

In the first group, Panel (a) for chronic malnutrition in children also suggests a west (higher) and east (lower) divide. Likewise, Panel (b) depicts the prevalence of non-Spanish languages such as Quechua and Aymara in the center and west and significant pockets of native languages in the family of Tupí-Guaraní in the east (light green). Panel (d) is consistent in showing years of education are highly unequal in the center and south, but it also shows lower levels dispersed at the north, east and west of the country. In the second group, panels (c) and (d) also bring to attention that there are a few small pockets of high dropout rates towards the center of the country and, conversely, three clusters with low dropout rates for males in the north east.

Table 2 indicates a strong and positive correlation between inequality in the years of education and non-Spanish speaking populations in municipalities. Another relatively high correlation is found between non-Spanish speaking populations and rates of malnutrition in children. There are historic, institutional, and other factors for the prevalence of obstacles to human capital development in municipalities with larger percentages of non-Spanish speakers.



(a) Chronic malnutrition in children.

(b) Non-Spanish speaking population.

(c) Secondary dropout rate (males).

(d) Secondary dropout rate (females).

(e) GINI coefficient of years of education.

Figure 1. Spatial distribution of human capital constraints.

Table 2
Correlation matrix of human capital constraints.

	(A)	(B)	(C)	(D)	(E)
(A) Malnutrition in children	1				
(B) Secondary dropout rate (males)	0.0350	1			
(C) Secondary dropout rate (females)	0.271***	0.533***	1		
(D) GINI coefficient of years of education	0.356***	0.141**	0.189***	1	
(E) Non-Spanish speaking population	0.453***	0.0735	0.202***	0.730***	1

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4. Methods

From a spatial perspective, one could argue that all regions could be related. But, regions physically closer to each other are, intuitively, even more related. The physical proximity may imply that a region tends to interact more with those neighboring regions than with regions that are farther away. As a result, in public policy terms, municipalities within a determined region tend to have more similarities to the public policies from its neighbouring regions. This could happen by direct or indirect learning and/or influencing processes.

The article addresses the question of whether it is possible to find clusters that respond to two conditions: (1) similarities in attributes (variables representing constraints to human capital development) and (2) similarities in geographical location. To identify these contiguous regions, two spatial data analysis methods are implemented. First, a spatial dependence analysis based on the local indicators of spatial association framework of Anselin (1995) allows us to identify local spatial clusters (hotspots and coldspots) for each human capital constraint.[12] Second, a regionalization analysis based on the spatially constrained clustering framework of Duque et al. (2012) allows us to cluster all geographic areas (not only hotspots and coldspots) into a number of homogeneous regions. Additionally, this clustering framework is multidimensional, so it allows to evaluate all the human capital constraints simultaneously. In what follows, we provide a brief overview of these two spatial methods.

4.1 Principal Components Analysis

The Principal Components Analysis (PCA) method can be traced back to the work of Pearson K. (1901) and Hotelling (1933), but it was not until advances in electronic computing that its use became widespread. Jollife and Cadima (2016) defined the objective of the PCA as reducing the dimensionality of a dataset while trying to lose as little information as possible (preserving variability as statistical information).[13]

PCA preserves variability by looking for a few linear combinations that are linear functions of the original variables and could summarize the data. These new variables are uncorrelated with

[12] See Bivand and Wong (2018) for a recent survey and implementation options of the local indicators of spatial association (LISA).

[13] For more information, Manly and Navarro Alberto (2017) provided an introduction to PCA and Mardia et al. (1994) a more technical description.

each other and are supposed to maximize variance. To find these new variables, PCA solves an eigenvalue decomposition problem.

Specifically, consider a matrix X composed by k variables, each containing n observations. After standardizing each variable (to avoid scale differences among variables), we can define X^TX as a $k \times k$ cross-product correlation matrix. The goal of the PCA method is to find a smaller number of variables—called principal components—that explain a large fraction of the variance of the original variables. Intuitively, for each principal component z_j, the PCA method finds the coefficients a_j for $j = 1, 2, ..., k$ such that:

$$z_j = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_kx_k, \tag{1}$$

where z_j (the principal component j) is a linear combination of the original variables $x_1, x_2, \ldots x_k$.[14]

In this paper, the PCA is implemented based on five variables: chronic malnutrition of children, secondary dropout rates for men and women, inequality in years of education, and non-spanish speaking population. The goal is to identify a smaller set of variables (principal components) that summarize most of the variance of these five constrains to human capital accumulation.

4.2 Spatial Dependence Analysis

An analysis of spatial dependence integrates the notion of attribute similarity with locational similarity. In particular, an analysis of global spatial dependence evaluates the existence of an overall clustering pattern in the spatial distribution of an attribute. From a statistical inference point of view, the null hypothesis of a global spatial dependence test postulates the randomness of the spatial location. In other words, all regions are independent from each other, and their location on a map is irrelevant for informational purposes. The rejection of the null hypothesis suggests the existence of a spatial structure that provides additional information about the phenomenon under study. The most well-known test for evaluating global spatial dependence is Moran's I (Cliff and Ord, 1981). In the context of the variables of this study, this test is defined as:

$$I = \sum_i \sum_j w_{ij} \cdot (x_i - \mu) \cdot (x_j - \mu) \,/\, \sum_i (x_i - \mu)^2 \tag{2}$$

where w_{ij} is row-standardized and represents an element of the weighting matrix that summarizes the spatial structure of the data, x_i is the level of the human capital constraint of municipality i, x_j is the level of the human capital constraint of municipality j, and μ is the average level of the human capital constraint. Statistical inference is carried out based on a computational approach of random permutation and the simulation of reference distribution.[15]

For any spatial analysis, the notion of spatial weights w_{ij} deserves some additional clarification. The role space is introduced via a weights matrix W that summaries the spatial structure of the data. Non-zero values of w_{ij} represent a "neighbor" relationship in geographical space.

[14] See Chapter 2 of Lee and Verleysen (2007) for a detailed exposition of the solution to this problem.

[15] See Anselin (1995) for a detailed presentation of inferential procedures for the Moran's I test.

There are different perspectives on which values of the w_{ij} could take. Among the most common specifications, there is the simple Queen contiguity structure in which two regions are defined as neighbors when they share a common border or a vertex. Similarly, in Rook contiguity structure, regions are defined as neighbors when they share a common border. Other neighbor structures can also be specified based on distance thresholds, inverse distance, and k-nearest neighbors. Based on its simplicity and interpretability, we use a Queen contiguity structure in this article.

Anselin (1995) proposed the Moran scatter plot as a way to visualize the strength and type of the spatial dependence. This scatter plot shows the relationship between the spatially lagged variable (Wx) and the original variable (x). More intuitively, this scatter plot highlights the relationship between an attribute at a particular location (x) and the weighted average of its neighbors (Wx). The slope of the fitted line between these two variables is the Moran's I statistic. By its construction, the Moran scatter plot provides a useful categorization of spatial dependence. A positive slope indicates positive spatial autocorrelation and it represents the existence of an overall pattern clustering in the sense that values at a particular location are surrounded by similar values of their neighbors. A negative slope indicates negative spatial autocorrelation and it represents the dominance of spatial outliers in the sense that values at a particular location are surrounded by dissimilar values of their neighbors. An intuitive graphical representation of negative spatial autocorrelation is the pattern of a checkerboard.

Based on the layout of the Moran scatter plot, Anselin (1995) also proposed local indicators of spatial association (LISA). Specifically, the Local Moran statistic provides a means to evaluate local spatial patterns such as hotspots (relatively high values), coldspots (relatively low values), and spatial outliers (high values surrounded by low values and vice-versa).[16] The local Moran's I is computed for each spatial unit and it is defined as:

$$I_i = \frac{(x_i - \mu)}{\sum (x_i - \mu)^2} \sum_j w_{ij} \cdot (x_j - \mu) \tag{3}$$

where the notation and interpretation of the variables follows that of equation (2). Statistical inference is based on a conditional permutation approach (See Anselin, 1995 for details).

4.3 Regionalization Analysis

The Max-p method for identifying spatially constrained clusters is based on a mixed integer programming model. Specifically, as described in Duque et al. (2012), it is formulated as the solution to the following constrained optimization problem:

$$Min\ Z = \left(- \sum_{k=1}^{n} \sum_{i=1}^{n} x_i^{k0} \right) * 10^h + \sum_i \sum_{j|j>i} d_{ij} t_{ij}, \tag{4}$$

[16] A local analysis of spatial dependence complements the analysis of global in the sense that the latter only identifies the existence of a clustering pattern, while the former describes the specific location of the clusters and spatial outliers.

Subject to:

$$\sum_{i=1}^{n} x_i^{k0} \leq 1 \quad \forall k = 1, \dots, n \tag{5}$$

$$\sum_{k=1}^{n} \sum_{c=0}^{q} x_i^{kc} = 1 \quad \forall i = 1, \dots, n \tag{6}$$

$$x_i^{kc} \leq \sum_{j \in N_i} x_j^{k(c-1)} \quad \forall i = 1, \dots, n; \forall k = 1, \dots, n; \forall c = 1, \dots, q \tag{7}$$

$$\sum_{i=1}^{n} \sum_{c=0}^{q} x_i^{kc} l_i \geq \text{ threshold } * \sum_{i=1}^{n} x_i^{k0} \quad \forall k = 1, \dots, n \tag{8}$$

$$t_{ij} \geq \sum_{c=0}^{q} x_i^{kc} + \sum_{c=0}^{q} x_j^{kc} - 1 \quad \forall i, j = 1, \dots, n \mid i < j; \forall k = 1, \dots, n \tag{9}$$

$$x_i^{kc} \in \{0, 1\} \quad \forall i = 1, \dots, n; \forall k = 1, \dots, n; \forall c = 0, \dots, q \tag{10}$$

$$t_{ij} \in \{0, 1\} \quad \forall i, j = 1, \dots, n \mid i < j \tag{11}$$

The decision variables are:

$$t_{ij} = \begin{cases} 1, & \text{if areas } i \text{ and } j \text{ belong to the same region } k, \text{ with } i < j \\ 0, & \text{otherwise} \end{cases}$$

$$x_i^{kc} = \begin{cases} 1, & \text{if areas } i \text{ is assigned to region } k \text{ in order } c \\ 0, & \text{otherwise} \end{cases}$$

The parameters of the problem are:

$i, I =$ Index and set of areas, $I = \{1, \dots, n\}$

$k =$ index of potential regions, $k = \{1, \dots, n\}$

$c =$ index of contiguity order, $c = \{0, \dots, q\}$, with $q = (n - 1)$

$$w_{ij} = \begin{cases} 1, & \text{if areas } i \text{ and } j \text{ share a border, with } i, j \in I \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

$N_i = \{j \mid w_{ij} = 1\}$, the set of areas that are adjacent to area i

$d_{ij} =$ dissimilarity relationships between areas i and j, with $i, j \in I$ and $i < j$

$h = 1 + \left\lfloor \log\left(\sum_i \sum_{j|j>i} d_{ij}\right) \right\rfloor$, which is the number of digits of the floor function of $\sum_i \sum_{j|j>i} d_{ij}$, with $i, j \in I$

$l_i =$ spatially extensive attribute value of area i, with $i \in I$

threshold = minimum value for attribute l at regional scale.

Equation (4) is the objective function and it is composed by two terms. The first term controls the number of regions by adding the number of areas designated as root areas. The second term controls total heterogeneity by adding pairwise dissimilarities between the areas of a region. Equation (5) indicates that an aggregated region should not have more than one core area.

Equation (6) indicates that each area is allocated to only one region k and one contiguity order c. Equation (7) indicates that area i is allocated to region k at order c if an area j exists and is allocated to the same region k in order $c1$. Equation (8) indicates that when a region is created, there is a predefined *threshold* based on a spatially intensive attribute, which for the purpose of this article is 10 percent of the population. Equation (9) indicates that total heterogeneity is calculated from pairwise dissimilarities. Finally, equations (10) and (11) indicate that variable integrity should be preserved.

5. Results

5.1 Principal Components

Table 3 presents the results of the PCA analysis. The table is divided into three parts. Table 3a describes the fraction of the total variance that is explained by the principal components. The first (PC1) and second (PC2) components explain 45 and 27 percent of the variance, respectively. Cumulatively, these two components explain almost three fourths of the total variance. Table 3b shows the squared correlation between the components and the original variables. The magnitude of this correlation facilitates the interpretation of each component in terms of the original variables. For instance, the main three variables that characterize the first components (PC1) are chronic malnutrition of children, inequality in years of education, and non-spanish speaking population. The second component, on the other hand, is mostly characterized by the secondary dropout rates of both males and females. Table 3c provides a criterion to select the number components. Kaiser (1960) suggests using the number of eigenvalues exceeding one as

Table 3
Principal component analysis.

(a) Total variance and cumulative proportion.

	PC1	PC2	PC3	PC4	PC5
Proportion of variance	0.45	0.27	0.15	0.08	0.05
Cumulative proportion	0.45	0.72	0.87	0.95	1.00

(b) Squared correlations between components and the original variables.

	PC1	PC2	PC3	PC4	PC5
Chronic malnutrition in children	**0.44**	0.04	0.46	0.06	0.00
Male dropout rate (Secondary level)	0.16	**0.63**	0.05	0.16	0.00
Female dropout rate (Secondary level)	0.32	**0.44**	0.04	0.19	0.00
Inequality in years of education	**0.64**	0.10	0.14	0.00	0.11
Non-Spanish speaking population	**0.68**	0.14	0.04	0.01	0.14

Notes: The non-squared correlations between PC2 and the male and female dropout rates are negative.

(c) Criterion to select the number of components.

	PC1	PC2	PC3	PC4	PC5
Eigenvalues	2.25	1.35	0.74	0.41	0.26
Kaiser criterion	2				

criterion. In our case, this would indicate the first two components, which explain 72 percent of the total variance.

Now that we have a small number of variables (PC1 and PC2) that summarize most of the variance of the five human capital constraints, we proceed to study their spatial distribution. In particular, we explore the degree of spatial dependence and patterns of regionalization of the first two principal components. The ultimate goal of these two explorations is to identify spatially contiguous clusters that face similar human capital constraints.

5.2 Spatial Dependence

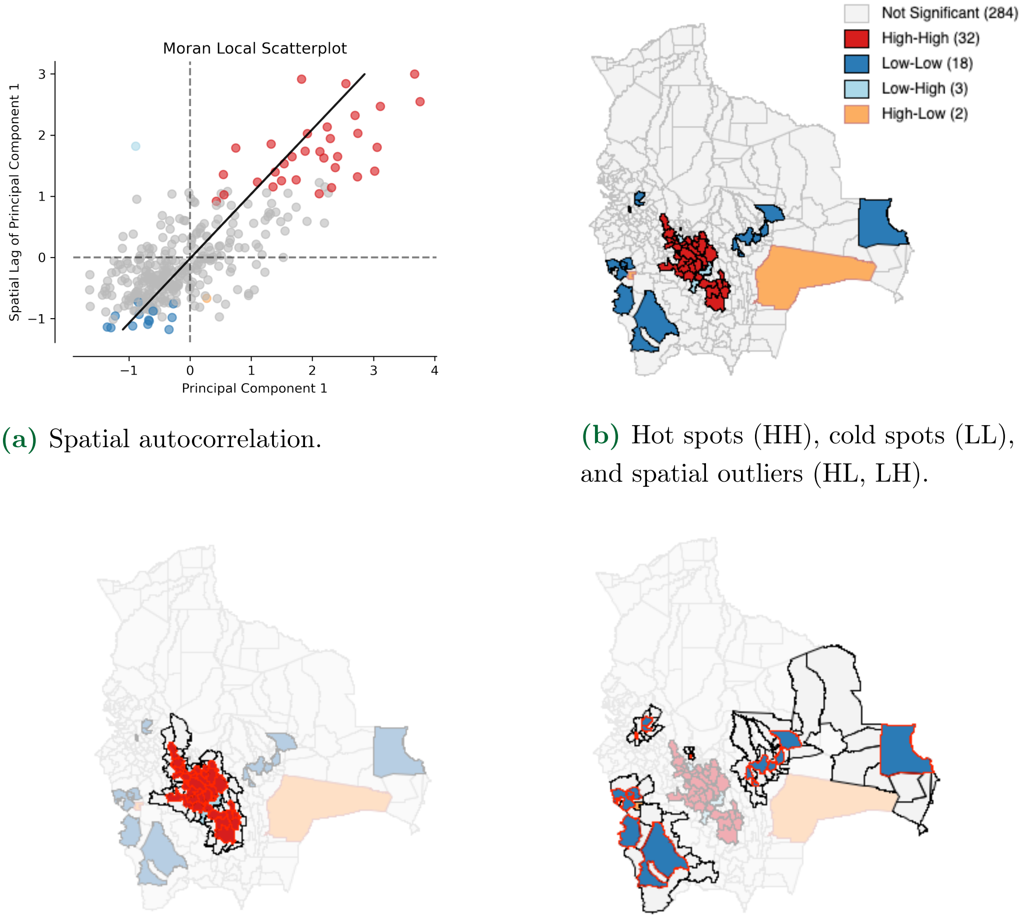
Panel (a) of Figures 2 and 3 shows the local Moran scatter plot of spatial autocorrelation. The horizontal axis in each panel indicates the standardized value of the principal component under study, PC1 in Figure 2 and PC2 in Figure 3. The vertical axis indicates weighted mean value of the surrounding neighbors (spatial lag). The regression line summarizes the overall degree of spatial dependence in the regional system. The Moran scatter plot is divided into four quadrants. The top-right (bottom-left) quadrant helps identify spatial clusters. These are cases in which a municipality and its neighbours share similar high (low) values of PC1.[17] The top-left (bottom-right) quadrant helps identify spatial outliers. These are cases in which a municipality has a high (low) value in PC1, but its geographical neighbours have low (high) values. The colored dots indicate statistically significant observations. To compute these statistically significant municipalities, a p-value of 0.01 is applied. Relative to the conventional 0.05 p-value, this lower significance level helps reduce concerns associated with the multiple comparison problem, which is common in the analysis of local indicators of spatial association (Anselin, 1995, 2020).

Panel (b) of Figures 2 and 3 shows the spatial distribution of statistically significant municipalities. Based on the quadrant location of the Moran scatter plot, municipalities are classified as high-high (hot spots) clusters, low-low (cold spots) clusters, and spatial outliers. Conceptually, a spatial cluster is composed by two parts: a core group of municipalities and their surrounding neighbours. In panel (b), the plotted hotpots and coldspots represent the cores of the spatial clusters. In panels (c) and (d), the spatial clusters include both core and neighbours.

The main findings related to Figure 2 are three-fold. First, there is a positive and statistically significant degree spatial dependence. That is, the human capital constraints (indicated by PC1) faced by the average municipality of Bolivia are highly similar to those faced by its geographical neighbors. The slope coefficient summarizing this relationship (the global Moran's I) is 0.57. Second, as shown in panel (c), clusters suffering from large human capital constraints are concentrated near the center of Bolivia. There are 32 municipalities classified as hotspots. That is, they are characterized by high malnutrition of children (37.5 percent), high education inequality (0.53 Gini index), and a large fraction of non-spanish speaking population (43.8 percent).[18] Third, as shown in panel (d), the clusters of municipalities facing fewer human capital constraints are

[17] In the analysis of spatial dependence, the qualification of high or low values is relative to the mean, which is normalized to zero.

[18] The numbers in parenthesis indicate the mean values for the 32 municipalities classified as hotspots. The values are presented in their original scale, similar those of Table 1 and Figure 1.



(a) Spatial autocorrelation.

(b) Hot spots (HH), cold spots (LL), and spatial outliers (HL, LH).

(c) High-PC1 clusters.

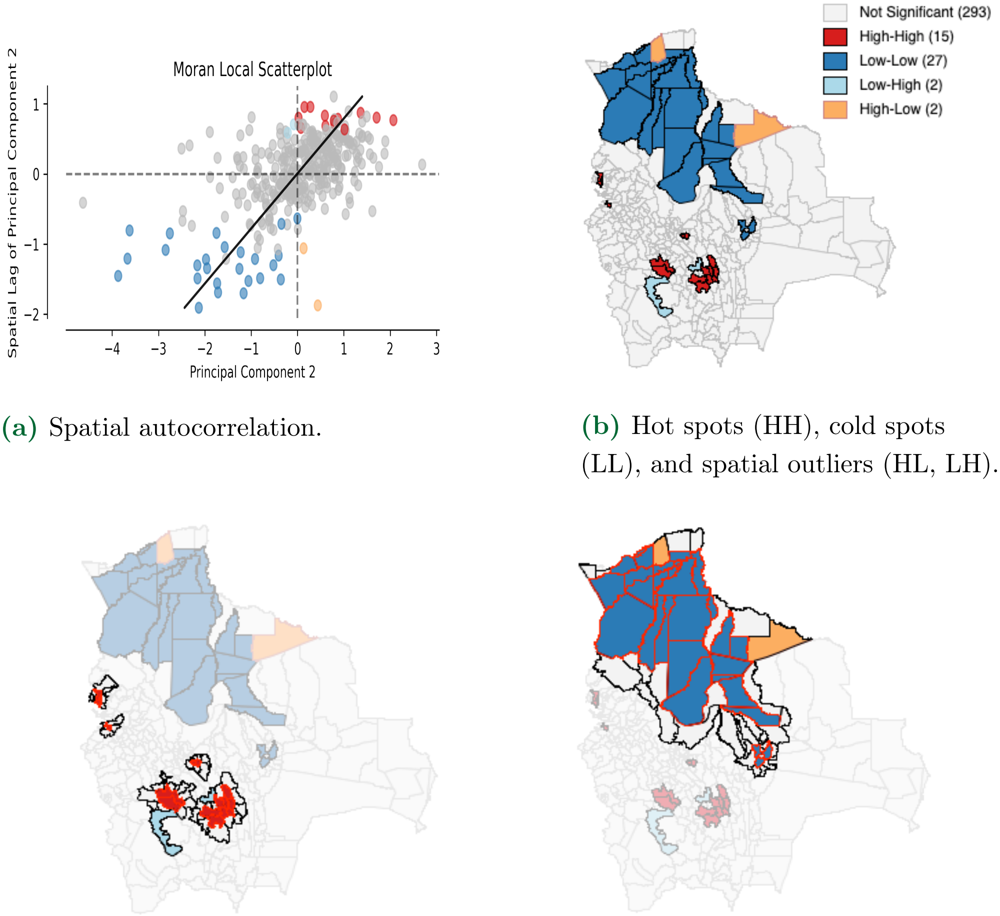
(d) Low-PC1 clusters.

Notes: Statistically significant municipalities are computed using a p-value of 0.01. The labels Low-High and High-Low indicate spatial outliers. The labels High-High and Low-Low indicate the *cores* of the spatial clusters. Given the results of Table 3, the interpretation of high-PC1 values is mostly associated with three human capital constraints: high chronic malnutrition of children, high inequality in years of education, and high non-Spanish speaking population.

Figure 2. Spatial distribution of the first principal component (PC1).

more scattered across the spatial distribution. There are 18 municipalities classified as coldspots. That is, they are characterized by lower malnutrition of children (19.2 percent), lower education inequality (0.33 Gini index), and a smaller fraction of non-speaking spanish population (5.2 percent).

Figure 3 shows the spatial dependence results for the second principal component (PC2). The main findings are also three-fold. First, there is a positive and statistically significant degree spatial dependence. The slope coefficient summarizing this relationship is 0.39. There is an important caveat in the interpretation of the clusters for PC2. As indicated in the notes of Table 3, there is a negative correlation between PC2 and the dropout rates in secondary education. Thus, the cluster classification in terms of the original variables is reversed. Hot spots of PC2 values should be interpreted as low-value observations while cold spots should be interpreted as high-value observations. Second, there is a large cluster suffering from high human capital



(a) Spatial autocorrelation.

(b) Hot spots (HH), cold spots (LL), and spatial outliers (HL, LH).

(c) High-PC2 clusters.

(d) Low-PC2 clusters.

Notes: Statistically significant municipalities are computed using a p-value of 0.01. The labels Low-High and High-Low indicate spatial outliers. The labels High-High and Low-Low indicate the *cores* of the spatial clusters. Given the results of Table 3a, the interpretation of low-PC2 values is mostly associated with two human capital constraints: high secondary dropout rate of males and high secondary dropout rate of females.

Figure 3. Spatial distribution of the second principal component (PC2).

constraints in the north of the country. There are 27 municipalities classified a cold spots. They are characterized by high rates of secondary education dropout of both males (8.1 percent) and females (7.6 percent). Third, there are only 15 municipalities classified as hot spots. In terms of the original variables, these municipalities are characterized by lower rates of secondary education dropout of both males (4.9 percent) and females (4.1 percent).

To sum up, compared to the choropleth maps of Figure 1, the Moran scatter plot has provided a mechanism to identify bi-dimensional clusters. The first dimension indicates the values of the human capital constraints (principal components) and the second dimension indicates the geographic proximity (spatial contiguity) of the municipalities. Nonetheless, one may still ask the following question: Is there a way to classify the non-statistically significant (grey) regions of Figures 2 and 3? In the next section, we aim to provide an answer to this question based on the Max-p clustering algorithm of Duque et al. (2012).

5.3 Regionalization

Panel (a) of Figure 4 shows the first-level subnational administrations of Bolivia. The classification and borders of these nine administrative regions (known as departments) have historic and political reasons. The second-and third-level subnational regions are provinces and municipalities. A large part of regionalization scheme shown in Figure 4a dates back to the colonial era, which indicates that the current regionalization scheme, based on subnational administrations, predates the establishment of the nation itself. At the same time, Montero and del Río (2013) underlined that often times the municipalities of each department are not only abstract administrative divisions, but the roads, local elites, and diversity gives them distinctive economic characteristics. In particular, underdeveloped transportation systems usually constraints the integration and development of various municipalities. Similar to weak transportation infrastructure, human capital underdevelopment and spatial disparities can limit the integration and development of the subnational administrations. In this context, panels (b) and (c) of Figure 4 provide a comparative regionalization scheme based on the identification of analytical regions.[19]

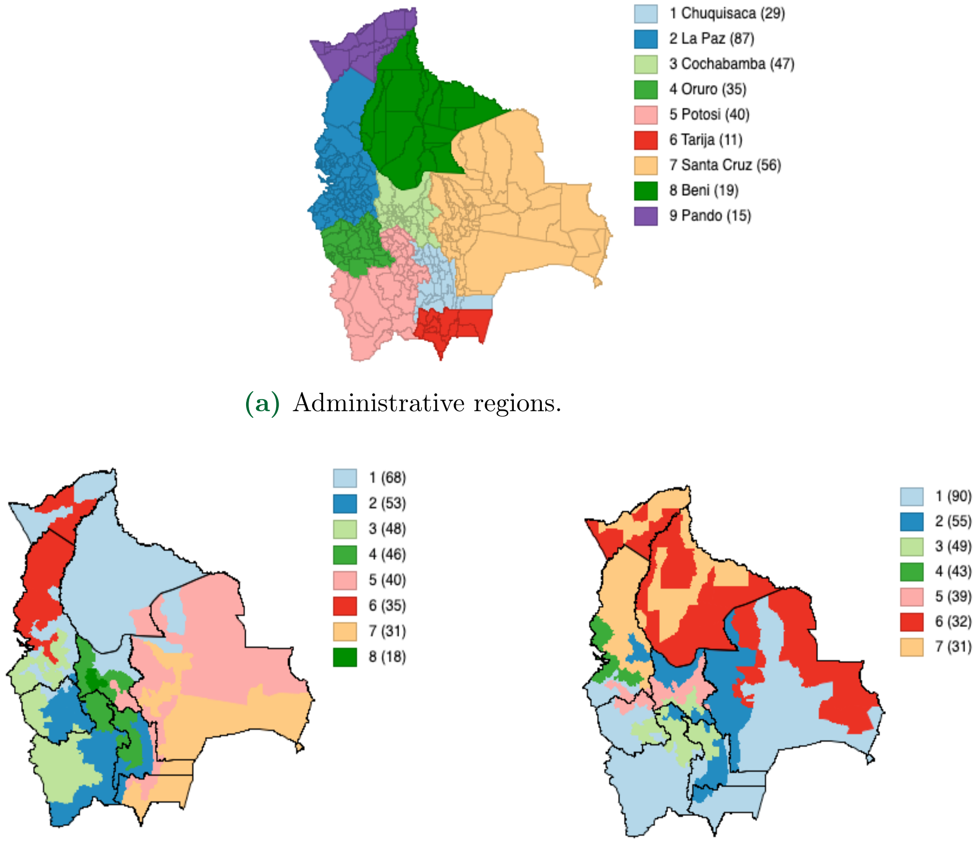
Through the lens of the Max-p clustering algorithm, analytical regions are identified using multiple human capital constraints (Table 1). In contrast to the commonly studied administrative regions (Figure 4a), the analytical regions of Figures 4a and 4b provide valuable information in two dimensions: attribute similarity and locational similarity. In contrast, the municipalities of each administrative region only share a similar location (Figure 4), but not necessarily common set of human capital constraints (recall Figure 1). In other words, the analytical regions of Figure 4 help us identify (1) spatially contiguous municipalities (clusters) facing similar human development constraints and (2) the degree of spatial heterogeneity within each administrative region. To illustrate the latter point, consider the administrative region of Cochabamba (that is, Region 3 in Figure 4). Based on the level of its human capital constrains (Figure 4b), there are four different spatial clusters within its administrative boundaries.

Overall, the main regionalization findings associated with Figure 4 are two fold. First, human capital constrains (represented by both PC1 and PC2) usually cross multiple administrate borders. Such crossing implies that the handling of human development issues requires the coordination of multiple regional governments. Second, some administrative regions are more heterogeneous and spatially more unequal than others. For instance, regions such as Cochabamba and La Paz encompass five spatial clusters while regions such as Pando and Beni only encompass two clusters.[20]

A better understanding of the spatial interactions between multidimensional indicators can inform both public and private investment decisions. Gaspar et al. (2019) estimated that to deliver the SDGs, low-and-middle-income countries may need 4 percent of GDP in additional spending every year. Countries such as Bolivia face multiple demands and resources are limited.

[19] In the context of the present paper, the word *analytical* refers to a group of regions that are identified as a result of solving an analytical problem of mathematical optimization (that is, equations (4) to (11)). The concept of *analytical* regions should not be confused with that of *functional* regions, see Duque et al. (2011) and Duque et al. (2007) for an detailed presentation of regionalization methods.

[20] See Appendix A for an evaluation of the Max-p clusters using alternative weights matrices.



(a) Administrative regions.

(b) Analytical regions based on human capital constraints (PC1).

(c) Analytical regions based on human capital constraints (PC1).

Notes: In panels (b) and (c), black outlines indicate administrative regions while fill colors indicate analytical regions. The numbers in parenthesis indicate the number of municipalities that belong to each region. The Max-p algorithm (Section 4.3) and data on the principal components (Table 3) of human capital constraints (Table 1) were used to estimate the analytical regions. PC1 mostly represents three human capital constraints: chronic malnutrition of children, inequality in the years of education, and non-spanish speaking population. PC2 mostly represents two human capital constraints: secondary dropout rate of males and secondary dropout rate of females.

Figure 4. Regionalization: Administrative regions vs. Analytical regions.

A better understanding of how human capital constraints are spatially distributed may inform the targeting of public policies, potentially increasing their effectiveness.[21]

6. Discussion: Moran Clusters or Max-p Clusters or Both?

Based on the human capital constraints indicators (PC1 and PC2), Figures 5 and 6 present a comparison of the spatial clusters that are derived from the local Moran and the Max-p frameworks. Next to the maps of each region, there is a parallel coordinate plot indicating the value

[21] Investments in education are associated with economic growth, perceived after a couple of decades (Aidt and Dutta, 2007; Bonfiglioli and Gancia, 2013; Atolia et al., 2019; Acosta-Ormaechea and Morozumi, 2017; Bose et al., 2007). Yet, political incumbents often prioritize other investments, such as roads, with benefits perceived sooner, while they hold office (Cetrángolo et al., 2017).

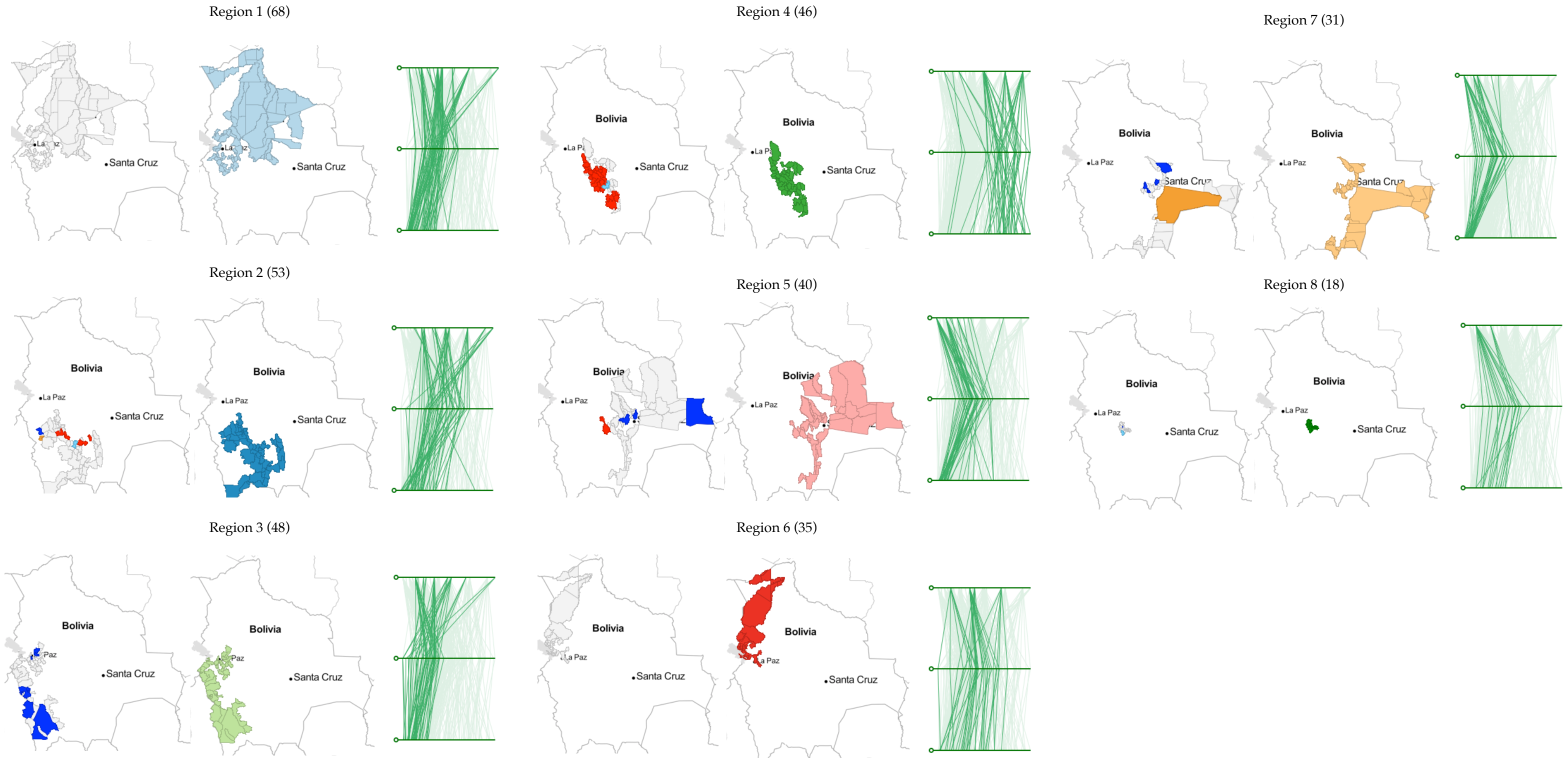
distribution of the original variables, represented as horizontal lines. The role of this plot is to facilitate the identification and interpretation of multidimensional clusters. In Figure 5, the first horizontal line of each parallel coordinate plot shows how the values of chronic malnutrition of children vary across municipalities. Similarly, the second and third horizontal lines indicate the value distribution of educational inequality and non-Spanish speaking population, respectively. As an example, consider the municipalities of Region 5. Based on their distribution along the first and third horizontal lines, these municipalities are characterized by low rates of malnutrition and low rates of non-Spanish speaking population. The distribution along the second horizontal line, however, indicates a level of education inequality in the lower-middle range.

As shown in Figure 5, the three spatial clusters identified by local Moran analysis are highly consistent with the clusters identified by the Max-p analysis. For instance, the high-PC1 cluster identified in Figure 2c is highly co-located with Region 4 of the Max-p analysis. The parallel coordinate plot also indicates that the geographically contiguous municipalities of this region are mostly characterized by high malnutrition of children, high education inequality, and a large fraction of non-Spanish speaking population. Similarly, the west low-PC1 cluster and east low-PC1 cluster identified in Figure 2d are partially consistent with Region 3 and Region 5, respectively.[22] In Region 3, municipalities tend to have middle-low malnutrition of children, low educational inequality, and a low fraction of non-Spanish speaking population. In Region 5, municipalities tend to have low malnutrition of children, middle-low educational inequality, and a low fraction of non-Spanish speaking population. In Figure 6, similar results between the local Moran and Max-p analysis are also evident, particularly for Region 6 and Region 7.

Figures 5 and 6 also highlight the differences between the two clustering methods. For the PC1 component, Region 2 and Region 5 include both hot spots and cold spots. A similar pattern is observed for the PC2 component regarding Region 1 and Region 2. These differences naturally raise concerns on the substitutability of the clustering methods, as they are not perfectly consistent. Differences in results, however, also indicate the complementary of the methods. If we are interested in finding spatial clusters for all the municipalities, only the Max-p is approach is suitable for that task. For instance, in Figure 5, the municipalities of Region 1 and Region 6 are not classified as clusters, according to the local Moran analysis. The Max-p analysis, nevertheless, finds that the municipalities of these regions are similar in both location and human capital constraints.

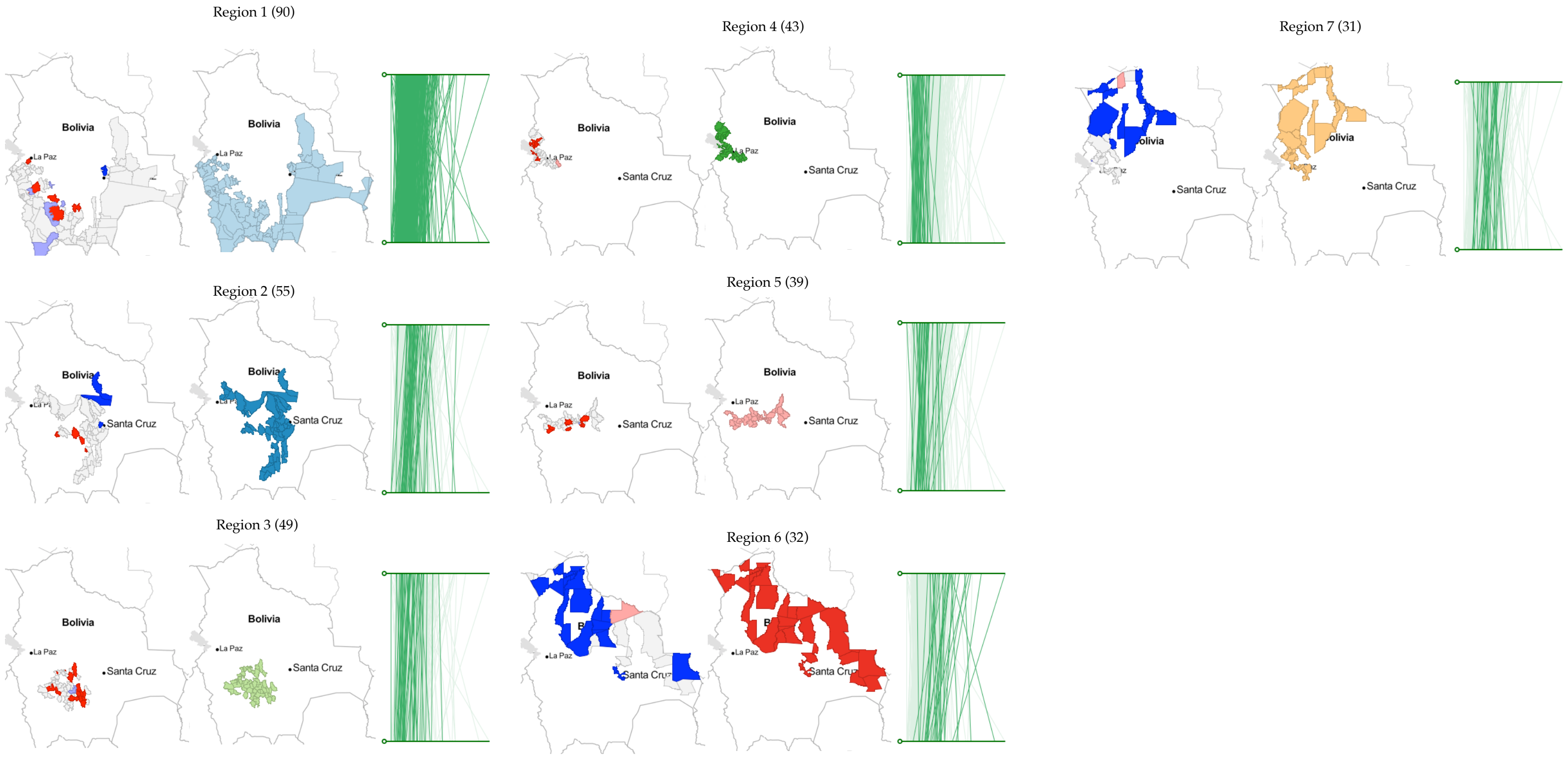
Overall, the similarities and differences of the spatial clusters suggest that the local Moran and the Max-p analyses could be better used as complementary tools. In a first stage, the local Moran analysis could be used to identify extreme clusters: hot spots and cold spots. Then, in a second stage, the Max-p analysis could be used to study the cluster classification of the remaining municipalities. Although perfect co-location of clusters is not guarantied, this sequential analysis could help identify robust spatial clusters that deserve further academic investigation. From a policy standpoint, these clusters could also help both national and regional governments prioritize

[22] Since Figure 5 only displays the cores of the clusters of the local Moran analysis, the comparison with the Max-p clusters of Region 3 and Region 5 is not as intuitive as in Region 4. Refer to Figure 2d for a more accurate comparison. In Figure 2d, the definition of cluster includes both core and neighbors, and thus it is more comparable with the results of Region 3 and Region 5.



Notes: For each region (or cluster of municipalities), the first map on the left is based on the spatial dependence results of Section 4.2. In this map, the red and blue municipalities indicate cluster cores. The map at the center is based on the regionalization results of Section 5.3. The parallel coordinate plot next to the central map indicates the value distribution of the original variables of PC1: chronic malnutrition of children (first horizontal line), inequality of years of education (second horizontal line), and share of non-Spanish speaking population (third horizontal line).

Figure 5. Local Moran clusters PC1 vs Max-P clusters PC1.



Notes: For each region (or cluster of municipalities), the first map on the left is based on the spatial dependence results of Section 4.2. In this map, the red and blue municipalities indicate cluster cores. The map at the center is based on the regionalization results of Section 5.3. The parallel coordinate plot next to the central map indicates the value distribution of the original variables of PC2: secondary education dropout rates of males (first horizontal line) and secondary education dropout rates of females (second horizontal line).

Figure 6. Local Moran clusters PC2 vs Max-P clusters PC2.

places where human development policies are more needed.[23]

Finally, there are two main caveats when comparing the spatial clusters of these methods. First, the number of observations to be clustered is different. The local Moran analysis only uses the observations from the high-high and low-low quadrants, while the Max-p analysis uses all observations. Second, the size of the Max-p clusters tends to be larger. In the local Moran analysis, the definition of spatial contiguity is usually of first order.[24] In contrast, the Max-p analysis is not restricted to first-order neighbors. In sum, these caveats emphasize that the two methodologies should not be treated as substitutes. As the number of observations and the extent of spatial contiguity differs, both analyses provide different insights.

7. Concluding Remarks

In this article, through the lens of a principal component analysis and geospatial analytical methods, we identify clusters of regions facing similar human capital constraints. Specifically, using a novel dataset of 339 municipalities in Bolivia, we evaluate regional disparities in chronic malnutrition in children, non-Spanish speaking population, secondary dropout rate of males, secondary dropout rates of females, and inequality in the years of education. Using a principal components analysis, we first aggregate these five variables into two components. The first principal component (PC1) mostly represents the municipal variation in chronic malnutrition of children, inequality in the years of education, and non-Spanish speaking population. The second principal component (PC2) mostly represents the variation in secondary dropout rate of males and secondary dropout rate of females. Next, we identify groups of municipalities that are characterized by both similar human capital constraints and similar location (spatial contiguity).

Two methodological approaches are implemented to identify geographically contiguous clusters. On the one hand, we use the local Moran analysis of spatial dependence developed by Anselin (1995) to identify regional hot spots (high-value clusters), cold spots (low-value clusters), and spatial outliers. On the other, we use the Max-p algorithm developed by Duque et al. (2012) to design a new map of Bolivia in which regional boundaries are endogenously derived from differences in human capital constraints.

The main results of the local Moran analysis are three-fold. First, there is a positive and statistically significant degree of spatial dependence in the regional system. That is, the human capital constraints (indicated by PC1 and PC2) faced by the average municipality of Bolivia are highly similar to those faced by its geographical neighbors. Second, there is a large cluster suffering from high human capital constraints concentrated near the center of Bolivia. Municipalities belonging to this cluster are characterized by high malnutrition of children (37.5 percent), high education inequality (0.53 Gini index), and a large fraction of non-spanish speaking population (43.8 percent). Third, there is another large cluster suffering from high human capital constraints

[23] Policy interventions are particularly needed in municipalities that constitute spatial underdevelopment traps. Specifically, when the underdevelopment of some municipalities reinforces that of their immediate neighbors, and vice-versa.

[24] That is, only the first-order (most proximate) neighbors are considered to determine the degree of spatial dependence.

in the north of the country. Municipalities belonging to this cluster are characterized by high rates of secondary education dropout of both males (8.1 percent) and females (7.6 percent).

Results of the regionalization analysis are largely consistent with those of the spatial dependence analysis. Moreover, based on the PC1 (PC2) data, the Max-p analysis indicates that Bolivia can be divided into eight (seven) geographical regions. The borders of these regions are largely different to those indicated by the administrative map of the country. This difference suggests that constraints to human capital accumulation frequently cross current administrative boundaries. Thus, the design and monitoring of human development policies need to be coordinated across multiple local administrations. Finally, some administrative regions are more heterogeneous and spatially unequal than others in terms of their human capital constraints.

The results of this article also indicate that a combined analysis of spatial dependence and regionalization helps overcome the limitations of each of these analyses when implemented separately. For instance, a single analysis of local spatial dependence only focuses on high and low value clusters and leaves many middle-value regions without classification. A single analysis of regionalization classifies all the regions, but it is difficult to identify core clusters and spatial outliers. The sequential implementation of spatial dependence and regionalization analyses helps overcome these issues and provides a more comprehensive evaluation of the geographical system.

Lastly, since this is the first article to study human capital constraints in Bolivia using a spatial clustering approach, there are still several avenues for further research. At least two extensions seem particularly promising and manageable in the context of the available data. First, the sensitivity of the Max-p algorithm can also be re-evaluated using alternative initialization and size parameters. Next, the regionalization of Bolivian municipalities can be re-evaluated using alternative clustering frameworks. Among them, we consider the spatially constrained clustering approach of Assunção et al. (2006) as the closest alternative.

Appendix A – Max-p Clusters Based on Alternative Connectivity Structures

As explained in Section 4, the point of departure of most spatial analyses is the definition of a spatial connectivity structure (spatial weights matrix). In this paper, spatial connectivity is defined based on a queen contiguity criterion. That is, the neighbors of a region are those who share a border or a corner. A contiguity criterion is not only parsimonious and intuitive, but also a requirement for the identification of Max-p clusters. To illustrate its importance, this section uses data of the first principal component (PC1) to evaluate alternative connectivity structures.[25]

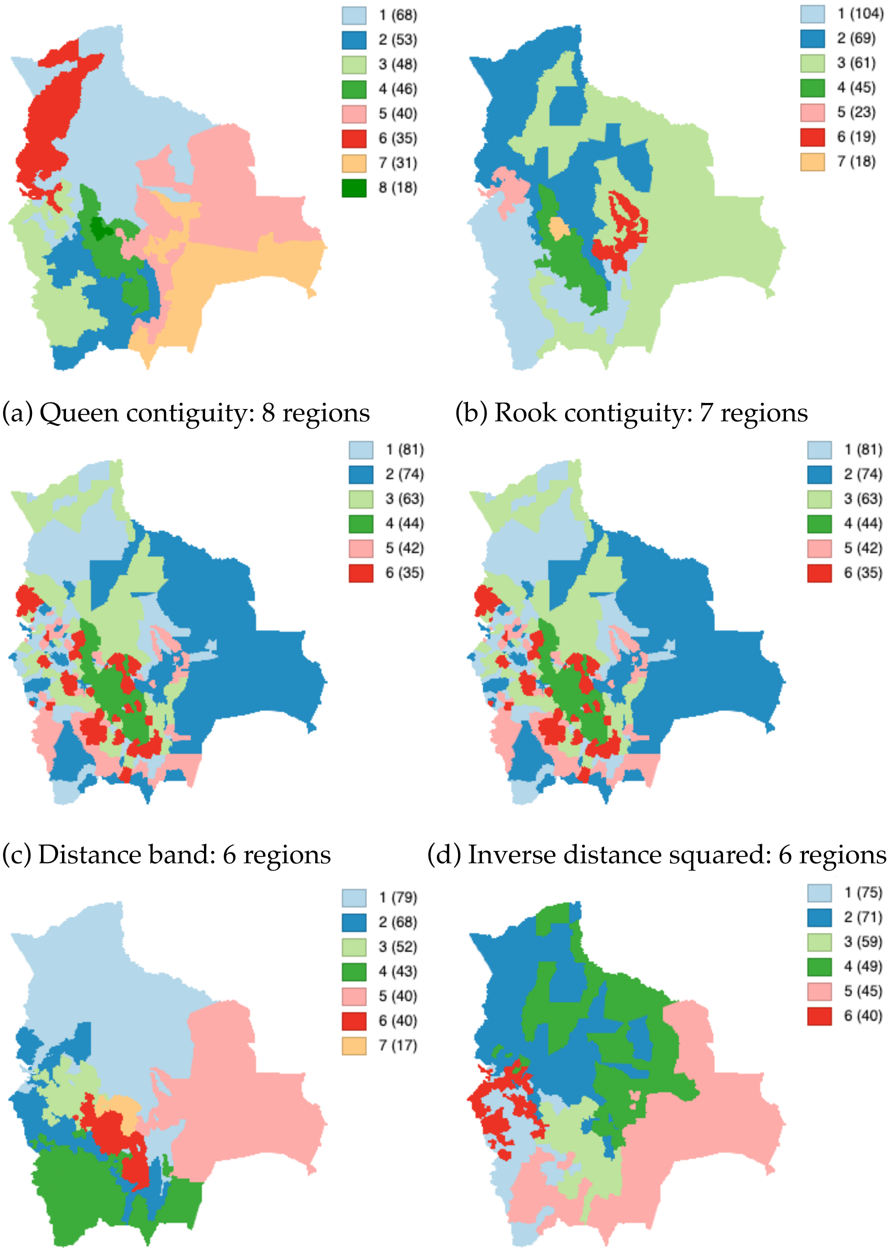
Figure A.1 shows the similarities and differences of the Max-p clusters across various connectivity structures. Panels (a) and (b) are based on two alternative definitions of contiguity. Compared to queen contiguity, the rook contiguity criterion identifies more compact regions. This result is expected as the rook criterion defines regional neighbors only based on common borders, not corners. More importantly, the main result of this comparison is that a higher degree of regional compactness implies changes in the number, size, and composition of the clusters.

Panels (c) and (d) are based on two alternative definitions of distance: minimum distance band and inverse distance squared. In both cases, the number, size, and composition of the clusters are the same. Regional contiguity, however, is a missing feature. The main result of this comparison is that regions facing similar human capital constraints are not necessarily contiguous, but closely located.

Panels (e) and (f) are based on the k-nearest neighbors approach. Compared to the contiguity and distance criteria, the k-nearest neighbors approach is particularly useful to identify clusters with a degree of high regional compactness. Nevertheless, spatial contiguity within each cluster is not assured. For instance, some regions in the south of Panel (e) and the east of Panel (f) are disconnected from their respective clusters.

Taken together, the results of Figure A.1 suggest that the identification of Max-p spatial clusters is sensitive to alternative connectivity structures and regional design objectives. On the one hand, if the objective is to achieve both spatial contiguity and high regional compactness, the rook contiguity structure appears to be the most suitable alternative. On the other, if the objective is to maximize regional compactness while accepting a small degree of discontinuity, the k-nearest neighbor structure would be most suitable. Finally, spatial connectivity structures based on distance are less suitable for identifying contiguous and compact clusters.

[25] Results for the PC2 component show similar patterns and are available from the authors upon request.



(a) Queen contiguity: 8 regions (b) Rook contiguity: 7 regions

(c) Distance band: 6 regions (d) Inverse distance squared: 6 regions

(e) Six nearest neighbors: 7 regions (f) Eight nearest neighbors: 6 regions

Figure A.1. Max-p clusters of PC1 based on alternative connectivity structures.

References

Acosta-Ormaechea, S., and Morozumi, A. (2017). Public Spending Reallocations and Economic Growth Across Different Income Levels. *Economic Inquiry 55(1)*, 98–114.

Aidt, T. S., and Dutta, J. (2007). Policy myopia and economic growth. *European Journal of Political Economy 23*(3), 734–753.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical analysis 27(2)*, 93–115.

Anselin, L. (2020). Local Spatial Autocorrelation (1). https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html.

Arribas-Bel, D., and Schmidt, C. R. (2013). Self-organizing maps and the US Urban Spatial Structure. *Environment and Planning B: Planning and Design 40(2)*, 362–371.

Assunção, R. M., Neves, M. C., Câmara, G., and Da Costa Frietas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science 20(7)*, 797–811.

Atolia, M., Li, B. G., Marto, R., and Melina, G. (2019). Investing in Public Infrastructure: Roads or Schools? *Macroeconomic Dynamics*, 1–30.

Barro, R. J. (2001). Human Capital and Growth. *American Economic Review 91(2)*, 12–17.

Becker, G. S., Murphy, K. M., and Tamura, R. (1990). Human Capital , Fertility , and Economic Growth. *Journal of Political Economy 98(5)*, 12–37.

Bivand, R. S., and Wong, D. W. (2018). Comparing implementations of global and local indicators of spatial association. *TEST 27(3)*, 716–748.

Bonfiglioli, A., and Gancia, G. (2013). Uncertainty, Electoral Incentives and Political Myopia. *Economic Journal 123(568)*, 373–400.

Bose, N., Haque, M. E., and Osborn, D. R. (2007). Public Expenditure and Economic Growth: A Disaggregated Analysis for Developing Countries. *The Manchester School 75(5)*, 533–556.

Bureau of International Labor Affairs (2018). 2014 Findings on the Worst Forms of Child Labor. Technical Report. Washington D.C: United States Department of Labor.

Canavire-Bacarreza, G., Duque, J. C., and Urrego, J. A. (2016). Moving Citizens and Deterring Criminals: Innovation in Public Transport Facilities. CAF Working Paper No. 2016/15.

Canelas, C., and Niño-Zarazúa, M. (2019). Schooling and Labor Market Impacts of Bolivia's *Bono Juancito Pinto* Program. *Population and Development Review 45(S1)*, 155–179.

Cetrángolo, O., Curcio, J., and Calligaro, F. (2017). *Evolución reciente del sector educativo en la región de América Latina y el Caribe: Los casos de Chile, Colombia y México*. Serie: Macroeconomía del Desarrollo No. 191. Santiago de Chile: CEPAL.

Chetty, R., Hendren, N., and Katz, L. F. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review 106(4)*, 855–902.

Chiswick, B. R., Patrinos, H. A., and Hurst, M. E. (2000). Indigenous Language Skills and the Labor Market in a Developing Economy: Bolivia. *Economic Development and Cultural Change 48(2)*, 349–367.

Church, R., Duque, J. C., and Restrepo, D. (2020). The p-innovation ecosystems model. *Physics* *and Society arXiv:2008.05885*.

Cliff, A. D., and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Collin, M., and Weil, D. N. (2020). The Effect of Increasing Human Capital Investment on Economic Growth and Poverty: A Simulation Exercise. *Journal of Human Capital 14(1)*, 43–83.

Cuervo, L. M. (2003). *Evolución reciente de las disparidades económicas territoriales en América Latina: estado del arte, recomendaciones de política y perspectivas de investigación*. Serie: Gestión Pública No. 41. Santiago de Chile: CEPAL.

Delboy, M. (2019). Determinants of School Attendance rate for Bolivia: A spatial econometric approach. *Unpublished Paper.*

Duque, J., Church, R., and Middleton, R. (2011). The p-Regions Problem. *Geographical Analysis 43(1)*, 104–126.

Duque, J. C., Anselin, L., and Rey, S. J. (2012). The Max-p-regions Problem. *Journal of Regional Science 52(3)*, 397–419.

Duque, J. C., Patino, J., Ruiz, L. A., and Pardo, J. E. (2013). Quantifying Slumness with Remote Sensing Data. Documentos de Trabajo: Economía y Finanzas No. 13-23.

Duque, J. C., Ramos, R., and Suriñach, J. (2007). Supervised Regionalization Methods: A Survey. *International Regional Science Review 30(3)*, 195–220.

Elias, M., and Rey, S. (2011). Educational Performance and Spatial Convergence in Peru. *Région et Développement 33*, 107–135.

Evia, J. L., Urquiola, M. S., Andersen, L., Antelo, E., and Nina, O. (1990). Geography and Development in Bolivia: Migration, Urban and Industrial Concentration, Welfare, and Convergence: 1950-1992. IDB Working Paper No. 115.

Fischer, M. (1980). Regional Taxonomy: A Comparison of Some Hierarchic and Non-Hierarchic Strategies. *Regional Science and Urban Economics 10(4)*, 503–537.

Fisher, W. D. (1958). On Grouping for Maximum Homogeneity. *Journal of the American Statistical Association 53(284)*, 789–798.

Fujita, L. D. V., Bagolin, I. P., and Fochezatto, A. (2020). Spatial distribution and dissemination of education in Brazilian municipalities. *The Annals of Regional Science 66*, 255–277.

Gaspar, V., Amaglobeli, D., Garcia-Escribano, M., and Soto, M. (2019). Fiscal Policy and Development: Human, Social, and Physical Investment for the SDGs. IMF Staff Discussion Note No. 19/03.

Gemmell, N. (1996). Evaluating the Impacts of Human Capital Stocks and Accumulation on Economic Growth: Some New Evidence. *Oxford Bulletin of Economics and Statistics 58(1)*, 9–28.

Godoy, R., Karlan, D. S., Rabindran, S., and Huanca, T. (2005). Do modern forms of human capital matter in primitive economies? Comparative evidence from Bolivia. *Economics of Education Review 24(1)*, 45–53.

Hansen, P., Jaumard, B., Meyer, C., Simeone, B., and Doring, V. (2003). Maximum Split Clustering Under Connectivity Constraints. *Journal of Classification 20(2)*, 143–180.

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review 37*, 204–212.

Hanushek, E. A., and Woessmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature 46(3)*, 607–668.

Hornberger, N. H. (1992). Literacy in South America. *Annual Review of Applied Linguistics 12*, 190–215.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24(6)*, 417–441.

Jenks, G. F. (1977). Optimal data classification for choropleth maps. Occasional Paper. Department of Geographiy, University of Kansas.

Jollife, I. T., and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065).*

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement 20(1)*, 141–151.

Kelley, J. (1988). Class conflict or ethnic oppression? The cost of being Indian in rural Bolivia. *Rural Sociology 53(4)*, 399–420.

Law, S., and Neira, M. (2019). An unsupervised approach to geographical knowledge discovery using street level and street network images. *Computer Vision and Pattern Recognition arXiv:1906.11907.*

Lawal, O. (2020). Spatially constrained clustering of Nigerian States: Perspective from Social, Economic and Demographic Attributes. *International Journal of Environment and Geoinformatics 7(1)*, 68–79.

Lee, J. A., and Verleysen, M. (2007). *Nonlinear dimensionality reduction.* New York: Springer-Verlag.

Lefkovitch, L. P. (1980). Conditional Clustering. *Biometrics 36*, 43–58.

Maclsaac, D. J., and Patrinos, H. A. (1995). Labour Market Discrimination Against Indigenous People in Peru. *The Journal of Development Studies 32(2)*, 218–233.

Manly, B. F. J., and Navarro Alberto, J. A. (2017). *Multivariate Statistical Methods: A Primer.* Boca Raton: CRC Press.

Maravalle, M., and Simeone, B. (1995). A spanning tree heuristic for regional clustering. *Communications in Statistics - Theory and Methods 24(3)*, 625–639.

Mardia, K. V., Kent, T. J., and Bibby J M (1994). *Multivariate Analysis.* London: Academic Press.

Martínez, P. P. (1990). Towards standardization of language for teaching in the Andean countries. *Prospects 20(3)*, 377–384.

McKenzie, D., and Rapoport, H. (2011). Can migration reduce educational attainment? Evidence from Mexico. *Journal of Population Economics 24(4)*, 1331–1358.

Mendez, C. (2018a). Beta, Sigma and Distributional Convergence in Human Development? Evidence from the Metropolitan Regions of Bolivia. *Latin American Journal of Economic Development 30*, 87–115.

Mendez, C. (2018b). On the distribution dynamics of human development: Evidence from the metropolitan regions of Bolivia. *Economics Bulletin 38(4)*, 2467–2475.

Mendieta Ossio, P. (2019). A Regional Landscape of Bolivian Economic Growth. *Latin American*

Journal of Economic Development 31, 77–98.

Mincer, J. (1984). Human Capital and Economic Growth. *Economics of Education Review 3(3)*, 195–205.

Miranda, M., Bento, A., and Aguilar, A. M. (2020). Malnutrition in all its forms and socioeconomic status in Bolivia. *Public Health Nutrition 23(S1)*, 1–8.

Montero, C., and del Río , M. (2013). Convergencia en Bolivia: Un enfoque espacial con datos de panel dinámicos. *Revista de Economía del Rosario 16(2)*, 233–256.

Morales, R., Galoppo, E., Jemio, L. C., Choque, M. C., and Morales, N. (2000). Bolivia: Geografía y Desarrollo Económico. Research Network Working Paper No. R-387. Inter-American Development Bank.

Murtagh, F. (1992). Contiguity-constrained clustering for image analysis. *Pattern Recognition Letters 13(9)*, 677–683.

Niembro, A., and Sarmiento, J. (2020). Regional development gaps in Argentina: A multidimensional approach to identify the location of policy priorities. *Regional Science Policy and Practice*, *Forthcoming (Early view)*.

Patrinos, H. A. (1997). Differences in Education and Earnings across Ethnic Groups in Guatemala. *Quarterly Review of Economics and Finance 37(4)*, 809–821.

Patrinos, H. A., and Psacharopoulos, G. (1993). The Cost of Being Indigenous in Bolivia: An Empirical Analysis of Educational Attainments and Outcomes. *Bulletin of Latin American Research 12(3)*, 293–309.

Pearson K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine 2*, 559–572.

Pritchett, L. (2001). Where Has All the Education Gone? *The World Bank Economic Review 15(3)*, 167–391.

Psacharopoulos, G. (1993). Ethnicity, Education, and Earnings in Bolivia and Guatemala. *Comparative Education Review 37(1)*, 9–20.

Psacharopoulos, G., and Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics 26(5)*, 445–458.

Rey, S. J., and Sastré-Gutiérrez, M. L. (2010). Interregional Inequality Dynamics in Mexico. *Spatial Economic Analysis 5(3)*, 277–298.

Sandoval, F. (2003). *Situación, tendencias y perspectivas de la convergencia regional en Bolivia 1980–1997*. La Paz: Banco Central de Bolivia.

SDSN-Bolivia (2020). *Atlas Municipal de los Objetivos de Desarrollo Sostenible en Bolivia.* La Paz: SDSN-Bolivia.

Soruco, C. F. (2012). Espacio, convergencia y crecimiento regional en Bolivia: 1990–2010. Documento de trabajo No. 01/2012. Banco Central de Bolivia.

Vargas, M. (2004). A spatial study about municipal poverty in Bolivia. MPRA Paper No. 6108.

Wise, S., Haining, R., and Ma, J. (1997). Regionalisation Tools for the Exploratory Spatial Analysis of Health Data. In M. Fischer and A. Getis (Eds.), *Recent Developments in Spatial Analysis* (pp. 83–100). Berlin, Heidelberg: Springer.