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

Mendez, Carlos and Gonzales, Erick

15 November 2020

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

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

Abstract Using a novel municipal-level dataset and spatial clustering methods, this article studies the distribution of human capital constraints across 339 municipalities in Bolivia. In particular, the spatial distribution of 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 the inequality of years of education. Through the lens of both spatial dependence and regionalization frameworks, the municipalities of Bolivia are endogenously classified according to both their level of human capital constraints and their locational similarity. Results from the spatial dependence analysis indicate the location of hotspots (high-value clusters), coldspots (low-value clusters), and spatial outliers for each of the previously listed constraints. Results from the regionalization analysis indicate that Bolivia can be regionalized into six to seven geographical locations that face similar constraints in the accumulation of human capital. The article concludes by highlighting the usefulness of spatial data analysis for designing and monitoring human development goals.

Keywords Human capital · Spatial dependence · Regionalization · Cluster analysis · Bolivia

JEL Classifications C31 · J24 · R10

C. Mendez[a]
Graduate School of International Development, Nagoya University, JAPAN.
E-mail: carlos@gsid.nagoya-u.ac.jp; Web: carlos-mendez.rbind.io

E. Gonzales[b]
United Nations Agency for Disaster Risk Reduction, Kobe, JAPAN.
E-mail: erick.gonzalesrocha@un.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.

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 no stranger to 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; Patrinos and Hurst 2007; Martínez 1990). However, there is less evidence on specific regional constraints hindering the accumulation of human capital. Although data limitations are common when studying low-and-middle-income economies, in this article, we exploit a novel dataset from the forthcoming Municipal Atlas of the Sustainable Development Goals in Bolivia (SDSN-Bolivia 2020). This is a comprehensive cross-sectional dataset that includes a large set of indicators to measure the Sustainable Development Goals (SDGs) at the municipal level in Bolivia.

In this article, we apply recent advances in geospatial methods to identify clusters of regions 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. Two methodological approaches are implemented to identify geographically contiguous clusters. We first use the classical spatial dependence framework of Anselin (1995) and Anselin et al (2007) to identify regional hot spots (high-value clusters), cold spots (low-value clusters), and spatial outliers. Next, we use the more recent integer programming approach of 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 from the spatial dependence analysis are four fold. Chronic malnutrition of children is mostly located in the center-west and center-south

of Bolivia. Non-Spanish speaking populations are located in the center-south. Large secondary dropout rates (both male and female) are located in the north of the country. High education inequality is mostly located in the center-south of Bolivia. The main results from the regionalization analysis suggest that Bolivia can be divided into six to seven geographical regions that face similar human capital constraints. Interestingly, the borders of these new regions are largely different from those indicated by the political map of Bolivia. As these identified regions encompass multiple administrative units (both departments and municipalities), the design and monitoring of human development policies need to be coordinated across multiple local governments and supported by the national government.

The rest of this article is organized as follows. Section 2 presents a survey of related literature. Section 3 introduces the methods of spatial dependence and regionalization, and describes the dataset. Section 4 presents the results, and Section 5 discusses robustness, sequential analysis, and complementary results. Section 6 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.[1]

Also on earnings, Patrinos and Hurst (2007) used a 1993 household survey in Bolivia (Encuesta Integrada de Hogares), conducted by INE 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).

But, 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.[2] For bilingual speakers in Bolivia, poorer proficiency in Spanish is penalized with lower earnings. Patrinos and Hurst (2007) 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

[1] 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).

[2] 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.

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 mothers years of education (particularly those with 7 to 12 years or more than 12 years of education).[3] 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 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.[4] 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.[5] 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.

[3] 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.

[4] 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.

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

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 Gonzalez 2003). For example, dispersion and lack of convergence for the period 1976-1992 (Urquiola 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 Carballo (2012); Kuscevic and del Río Rivera (2013); Mendieta Ossio (2019) observed limited spatial dependence among departments for economic growth, Barrenechea 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,[6] migration,[7] and child labor are significantly related to school attendance. Canelas and Niño-Zarazúa (2019) stated that short school days and lax legal frameworks may contribute to the finding of Bureau of International Labor Affairs (2018) that about 15 percent of children between 7 to 14 years old en-

[6] 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.

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

gage 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.[8] 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 method

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

[8] 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 3.2), 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 Vernier Fujita et al (2020); Elias and Rey (2011).

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 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 Methods and Data

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 deals with spatial autocorrelation to address 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. This is not straightforward given that some municipalities may have similar attributes but are not geographically close. Thus, some sort of balance needs to be found. In that process, trade-offs occur as some similarities in attributes may need to be sacrificed for geographical proximity and vice versa.

To identify contiguous regions facing similar human capital constraints, 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.[9] 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.

3.1 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

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

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:

$$I = \sum_i \sum_j w_{ij} \cdot (x_i - \mu) \cdot (x_j - \mu) \; / \; \sum_i (x_i - \mu)^2 \tag{1}$$

where w_{ij} represents a weights 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.[10]

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. 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 autocor-

[10] See Anselin (1995) and Anselin (2017) for a detailed presentation of inferential procedures for the Moran's I test.

relation 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).[11] The local Moran's I is computed for each spatial unit and it is defined as:

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

where the notation and interpretation of the variables follows that of Equation 1. Statistical inference is based on a conditional permutation approach (See Anselin (1995) and Anselin (2017) for details).

3.2 Regionalization analysis

The Max-p method for identifying spatially constrained clusters is based on a mixed integer programming model. Specifically, it is formulated as the solution to the following constrained optimization problem:

$$Min\ Z = \left(- \sum_{k=1}^{n} \sum_{i=1}^{n} x_i^{k0} \right) * 10^h + \sum_i \sum_{j|j>i} d_{ij} t_{ij}, \tag{3}$$

Subject to:

$$\sum_{i=1}^{n} x_i^{k0} \leq 1 \quad \forall k = 1, \ldots, n \tag{4}$$

$$\sum_{k=1}^{n} \sum_{c=0}^{q} x_i^{kc} = 1 \quad \forall i = 1, \ldots, n \tag{5}$$

$$x_i^{kc} \leq \sum_{j \in N_i} x_j^{k(c-1)} \quad \forall i = 1, \ldots, n; \forall k = 1, \ldots, n; \forall c = 1, \ldots, q \tag{6}$$

[11] 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.

$$\sum_{i=1}^{n}\sum_{c=0}^{q} x_i^{kc} l_i \geq \text{ threshold } * \sum_{i=1}^{n} x_i^{k0} \quad \forall k = 1, \ldots, n \tag{7}$$

$$t_{ij} \geq \sum_{c=0}^{q} x_i^{kc} + \sum_{c=0}^{q} x_j^{kc} - 1 \quad \forall i, j = 1, \ldots, n \mid i < j; \forall k = 1, \ldots, n \tag{8}$$

$$x_i^{kc} \in \{0, 1\} \quad \forall i = 1, \ldots, n; \forall k = 1, \ldots, n; \forall c = 0, \ldots, q \tag{9}$$

$$t_{ij} \in \{0, 1\} \quad \forall i, j = 1, \ldots, n \mid i < j \tag{10}$$

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, $I = \{1, \ldots, n\}$

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

$c =$ index of contiguity order, $c = \{0, \ldots, q\}$, 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 3 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 4 indicates that an aggregated region should not have more than one core area. Equation

5 indicates that each area is allocated to only one region k and one contiguity order c. Equation 6 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 7 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 8 indicates that total heterogeneity is calculated from pairwise dissimilarities. Finally, Equation 9 and 10 indicate that variable integrity should be preserved.

3.3 Principal components analysis

The Principal Components Analysis (PCA) technique 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).[12]

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 each other and are supposed to maximize variance. To find these new variables, PCA solves an eigenvalue problem. The PCA was implemented based on all the variables identified as constraints to human capital development: malnutrition, language, dropout rates for men and women, and inequality in years of education. The PCA results suggest the retention of one component which was named integrated human capital constraints.

Summarizing multiple variables into a single index (integrated human capital constraints, the first component of the PCA) will be useful for discussing robustness along with the sequential and complementary application of the local Moran and Max-p frameworks. For example, while the attributes (constraints) are analyzed individually using the local Moran method in the results section, the discussion section applies the local Moran method to the single index and compares it to the multivariate Max-p method (see Figure 14). Likewise, the discussion section applies a univariate Max-p method to

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

the single index and compares it to the multivariate Max-p method (see Figure 15). The PCA, thus, reduces dimensionality and enables the identification of clusters by accommodating multiple variables at the same time. Even though the PCA analysis attempts to explain most of the variation of the data and has a complementary use in the analyses, approaches considering the entire variation of the data might be preferred (multivariate Max-p).

3.4 Data and measurements

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

- **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).
- **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).
- **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).
- **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).
- **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 estima-

tions 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.[13]

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.[14] The situation is particularly dire in municipalities where maximum values indicate that around half of all kids might be malnourished. Some examples include municipalities such as El Choro, Corque, Choque Cota, and 7 others in Oruro (malnourishment levels reach 53 percent), Tinguipaya, Urmiri, Chuquihuta, and 12 others in Potosí (49 percent), or Poroma, Azurduy, and 17 others in Chuquisaca (41 percent). On the other hand, municipalities in the department of Santa Cruz have on average the lowest levels of malnutrition among children (8.5 percent). These summary results can be visualized and will be explained in more detail later in Figure 1. Though the numbers come from the latest census in 2012, another observation is that in some municipalities of Bolivia such as San Pedro in Potosí, or Vila Vila in Cochabamba, along with a number of other municipalities, more than 50 percent of their population do not have Spanish as their mother tongue.

When it comes to indicators measuring basic features of the education systems such as dropout rates, the dataset of SDSN-Bolivia (2020) provides

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

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

a useful disaggregation by gender. For example, dropout rates for males are higher than those of females. It is possible that cultural issues or incentives to enter the job market are stronger for man and may induce them to dropout. Godoy et al (2005) stated that among municipalities with a high proportion of communities away from main towns, women do not usually enter the market for wage labor.

The last indicator measures inequality in years of education. With an average value higher than 0.3, there is a relatively high inequality in the distribution of this indicator in Bolivia. While municipalities corresponding to the capital cities in La Paz (0.21), Oruro (0.25), or Cobija (0.25) display more egalitarian distribution of citizens in terms of years of education (we can assume higher levels in years of education as well), municipalities such as Ocurí (0.64) in Potosí, Arque (0.64) in Cochabamba, or Tarvita (0.60) in Chuquisaca experience severe inequality. In other words, while a few people have many years of education, the majority does not. As it will be more apparent in the next figures, a somewhat atomized geographical location and other diverse characteristics of municipalities in the country are revealed when looking beyond averages and urban centers.

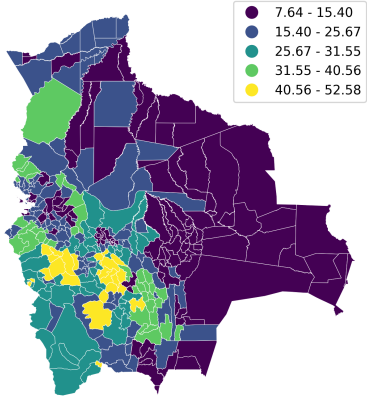
Table 1: Descriptive statistics: Human capital constraints

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Chronic malnutrition in children (percent, 2016)	24.00	12.00	7.60	14.00	23.00	30.00	53.00
Non-Spanish speaking population (percent, 2012)	15.00	14.00	0.66	4.90	9.60	20.00	60.00
Secondary dropout rate (male percent, 2017)	5.00	2.90	0.00	3.20	4.70	6.40	21.00
Secondary dropout rate (female percent, 2017)	4.10	2.90	0.00	2.40	3.40	5.20	22.00
GINI coefficient of years of education (2012)	0.39	0.08	0.20	0.33	0.37	0.43	0.64

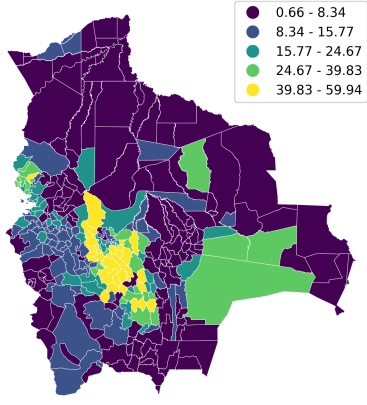
Figure 1 provides a first overview of the spatial distribution of each human capital constraint. The breaks of each choropleth map are optimally selected by using the natural breaks classification method of Fisher (1958); Jenks (1977). This method uses a nonlinear algorithm to group regions in a way that maximizes within-group homogeneity. In essence, this algorithm 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 from all maps that the identified clusters are not necessarily contiguous or spatially integrated. This is because the uni-dimensional clustering framework of Fisher (1958); 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.

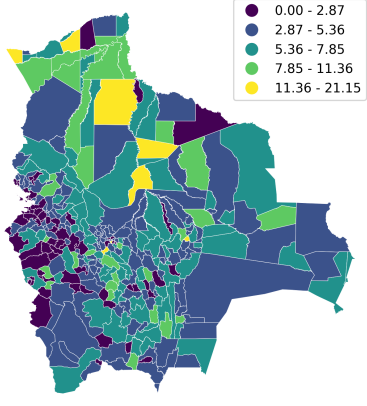
Before moving to the next section, a brief interpretation of the five panels in Figure 1 can be made summarized in three stages: percentages of malnourished children and non-Spanish speaking people in the population; dropout rates; and inequality in the years of education. The problem of malnourished children could divide the country in three broad regions. The east with overwhelmingly low levels of children malnutrition. A middle that includes municipalities from the north to the south of the country displaying not so severe levels of malnutrition. And, the west where malnutrition in municipalities could range from 25 to 50 percent. There is also, however, some clusters shown in yellow which are particularly worrisome. These clusters of municipalities display the highest levels of children malnutrition and are mainly located in Oruro and the north of Potosí.



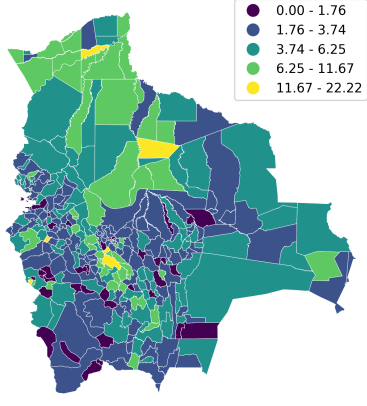
(a) Chronic malnutrition in children



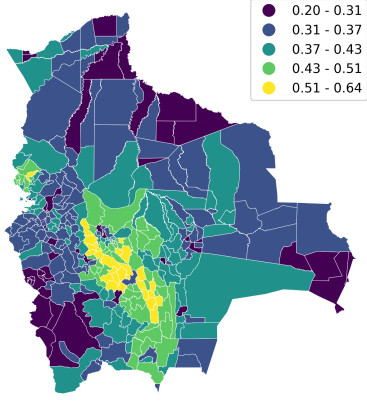
(b) Non-Spanish speaking population



(c) Secondary dropout rate (males)



(d) Secondary dropout rate (females)



(e) GINI coefficient of years of education

Fig. 1: Spatial distribution of human capital constraints

Panel b of Figure 1 denotes the prevalence of non-Spanish languages such as those in the family of Tupí-Guaraní, Quechua and Aymara. While the departments of Beni, Pando and Tarija overwhelmingly have municipalities with low levels of non-Spanish speaking populations, the rest of the departments show a more diverse picture. In this case, municipalities with high percentages of non-Speaking populations (see color yellow) are located across the borders of Chuquisaca, Potosí (north) and Cochabamba.

In the second stage, panels c and d show that for both females and males, there are higher dropout rates in the north of the country (the departments of Beni and Pando). But, there are also high dropout rates (see color yellow) in municipalities outside those departments: north of Potosí, Oruro, Cochabamba, and one isolated case corresponding to the municipality of Colpa Belgica in Santa Cruz for the case of males. In terms of lower dropout rates, there are good performer municipalities (in dark blue) particularly, but not exclusively, towards the south and west.

The last stage indicates that there are lower levels of inequality in years of education (dark blue) towards the western and eastern borders of the country. Conversely, higher inequality in years of education is clearly found in the middle-south of the country with other groups in between.

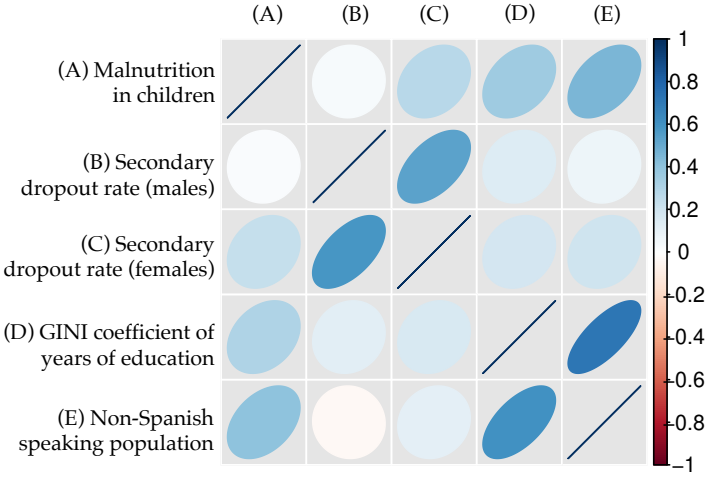


Fig. 2: Correlation matrix of human capital constraints

Notes: Pearson (Spearman) correlations above (below) the diagonal.

Before proceeding to the results section, the correlation among the studied variables was explored. Figure 2 indicates a strong and positive correlation between higher levels of inequality in the years of education and higher percentages of non-Spanish speaking population in municipalities. Another relatively high correlation is found between non-Spanish speaking populations and rates of malnutrition in children. As it will be observed in the next section, there can be historic, institutional, and other factors for the prevalence of obstacles to human capital development in municipalities with larger percentages of non-Spanish speakers. Thus, it is interesting to see a low correlation between non-Spanish speaking populations and secondary dropout rates in particular for the case of males.

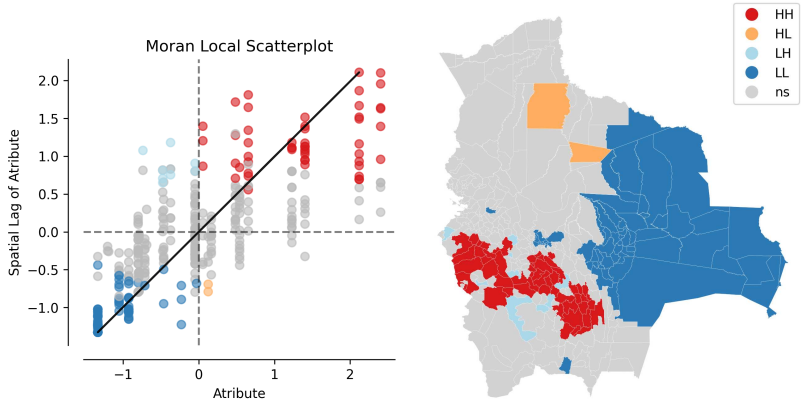
4 Results

4.1 Spatial dependence

In Figure 3, the Moran scatter plot of spatial autocorrelation is displayed on the left. The horizontal axis represents the attribute being analyzed, that is, the percent of children less than five years old with chronic malnutrition. The vertical axis represents the spatial lag of the variable that appears in the horizontal axis. Conceptually, the spatial lag of a municipality is defined by the average value of the neighbouring municipalities. The top-right (bottom-left) quadrant identifies cases where both a municipality as well as its neighbouring municipalities have high (low) values in the variable being analyzed. The colored dots, in particular, show cases where the relationship is statistically significant. This is one way to identify clusters that share both attribute and locational similarity. Translating those results into a map (right side of Figure 3), we can identify the hot spots in red (high values of chronic malnutrition) and cold spots in blue (low values of chronic malnutrition).

The cluster of municipalities where malnutrition among children is less prevalent (cold spots) are heavily located on the east part of the country. In particular, this cluster largely overlaps with the department of Santa Cruz. On the other hand, high levels of malnutrition (hot spots) are found in the lower center and west of the country. Hot spots mainly span across three departments: the west of Chuquisaca, the north of Potosí, and Oruro.

The Moran scatter plot also allows us to identify spatial outliers. Observations identified as statistically significant are colored in orange and sky blue.



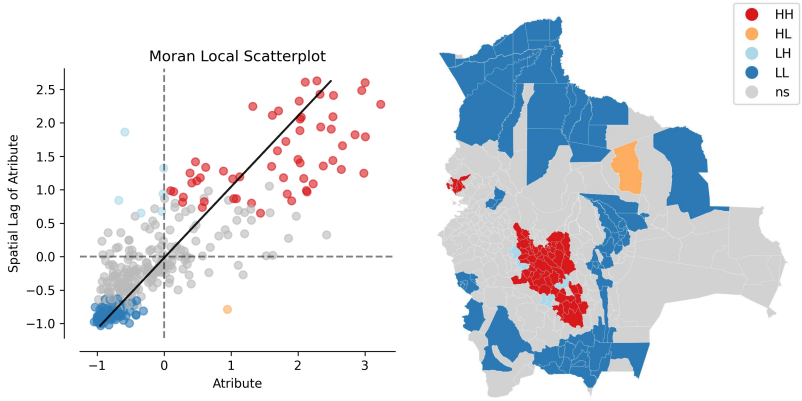
(a) Spatial autocorrelation

(b) Hotspots (HH), coldspots (LL), and spatial outliers (HL, LH)

Fig. 3: Spatial distribution of malnutrition in children

High-Low (Low-High) outliers represent municipalities that have a high (low) value in the variable being analyzed, but their geographical neighbours have low (high) values. In other words, those are high (low) performers surrounded by low (high) performers. For example, in the bottom-right quadrant of Figure 3, the yellow dots indicate statistically significant municipalities with high levels of malnutrition in children surrounded by municipalities with low levels of malnutrition. Conversely, in the upper-left quadrant, the sky-blue dots indicate statistically significant municipalities with low levels of malnutrition surrounded by municipalities with high levels of malnutrition. Finally, regions in gray indicate municipalities where the results are not statistically significant at the conventional significance level of five percent.

The rest of the figures showing a Moran scatter plot with its spatial distribution are interpreted in the same way in terms of clustering patterns and spatial outliers. In Figure 4, populations where Spanish is the mother tongue and dominant language are mainly located in regions of the north and south of the country. An interesting feature is that the hot spots (municipalities with high non-Spanish speaking populations) seem to predominantly correspond to Quechua speaking populations located in the center of the country. There is one high-low spatial outlier that corresponds to a Guarayu (from the Tupí-Guaraní language family) speaking region. There are three low-high spatial outliers located around the main hot spot.



(a) Spatial autocorrelation

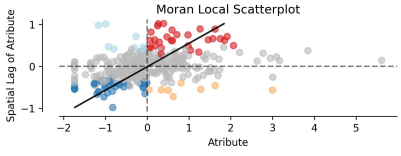
(b) Hotspots (HH), coldspots (LL), and spatial outliers (HL, LH)

Fig. 4: Spatial distribution of non-Spanish speaking population

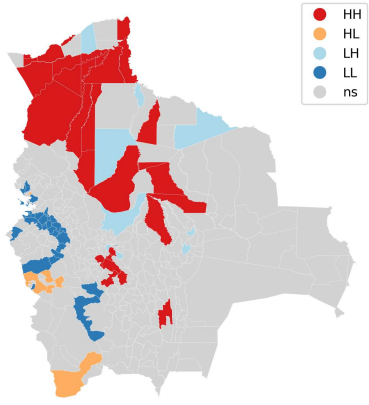
The spatial distribution for secondary dropout rates in Figures 5 and 6 display largely similar messages overall. For both variables, there are regions with higher dropout rates in the north of the country. For the secondary dropout rate for females, however, there is a larger cold spot in the south east part of the country (overlapping in part with the department of Potosí). This cold spot indicates that the secondary dropout rate for females is substantially lower in that part of the country.

Lastly, in Figure 7 there is a large hot spot in the middle-south of the country. This spatial cluster indicates a high level of inequality in the years of education. Interestingly, this cluster tends to overlap with the hot spots clusters of malnutrition and non-Spanish speaking populations. Regions with low levels of inequality in the years of schooling (cold spots) conform multiple large clusters. They are distributed in the north, west, and east of the country. In terms of spatial outliers, it calls the attention the existence of a large region in the south of Santa Cruz. This municipality shows a high level of education inequality and it is surrounded by municipalities with low levels of inequality.

Compared to the uni-dimensional clusters of Figure 1, the Moran scatter plot has provided a mechanism to identify bi-dimensional clusters. The first dimension is based on a human capital attribute (or constraint) and the second dimension is based on the geographic proximity (spatial contiguity)

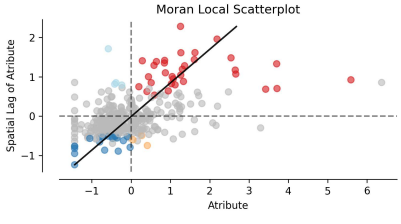


(a) Spatial autocorrelation

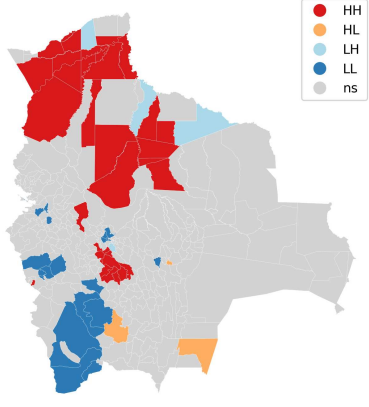


(b) Hotspots (HH), coldspots (LL), and spatial outliers (HL, LH)

Fig. 5: Spatial distribution of secondary male dropout of males



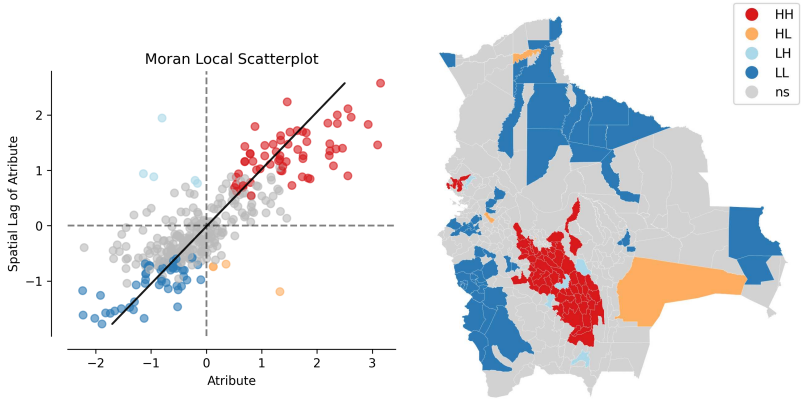
(a) Spatial autocorrelation



(b) Hotspots (HH), coldspots (LL), and spatial outliers (HL, LH)

Fig. 6: Spatial distribution of secondary dropout of females

of the municipalities. Nonetheless, one may still ask the following question: Is there a way to cluster those non-statistically significant (grey) regions? 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).



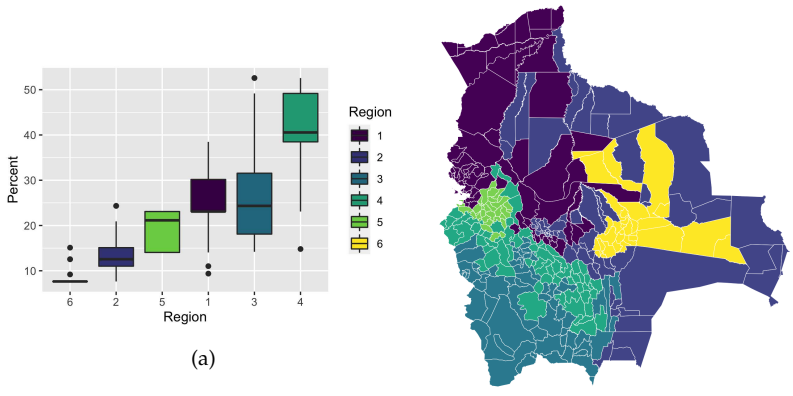
(a) Spatial autocorrelation

(b) Hotspots (HH), coldspots (LL), and spatial outliers (HL, LH)

Fig. 7: Spatial distribution of inequality of years of education

4.2 Regionalization

The division of Bolivia into nine departments has historic and administrative reasons (some of them going back to the colonial administration indicating that their current existence predates the nation itself). At the same time, Kuscevic and del Río Rivera (2013) underlined that often times municipalities in each department are not only abstract administrative divisions but that roads, local elites, and diversity gives them distinctive economic characteristics. Weak integration in terms of transportation gives way to relatively isolated municipalities where the role of local elites is strengthened. These circumstances are combined with the inherent geographic diversity of the country which, in turn, provides comparative advantages to their productive activities. While considering the case for distinctive attributes of several regions, it is important to recognize as well that from a spatial perspective all regions are also related. Furthermore, in practice, some municipalities will be even more related to other municipalities in close physical proximity, including those that belong to other departments. Therefore, even though the current administrative division of Bolivia has its uses, the identification of new regions could be more informative when it comes to the analysis, design, and implementation of public policies.

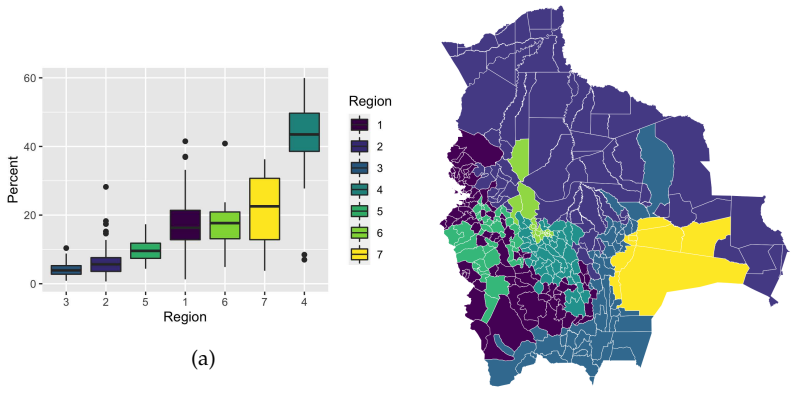


(a) (b)

Fig. 8: Regionalization of chronic malnutrition in children

The regional clusters proposed by the results of this article can inform those policy-making processes with more accurate divisions based on municipalities' performance for the issues at hand. One reason behind the use of regional clusters is that spatial similarities are a proxy for interaction. The probability of a municipality to interact with others increases with physical proximity. Neighbours are important because the success of a municipality can influence the success on its neighbours and vice-versa. The same happens with municipalities that fail.

Taking Figure 8 as first example, the max-p method suggests that Bolivia could be classified into six regions from the perspective of the prevalence of chronic malnutrition in children as well as geographic proximity across municipalities. As it can be seen on the right side of the figure, these new regions do not necessarily coincide with the geographical borders currently established for the nine departments of Bolivia. This is rightly so because in terms of malnutrition, as well as for other variables, municipalities could benefit from coordinating with other municipalities that may not be in the same department but face similar constraints in the analyze attribute and are geographically close (despite belonging to a different department). The results show municipalities clustered in regions that share similar challenges (or advantages) and could reap benefits from coordinating and being analyzed as a group when thinking about policies, in particular for those municipalities



(a) (b)

Fig. 9: Regionalization of non-Spanish speaking population

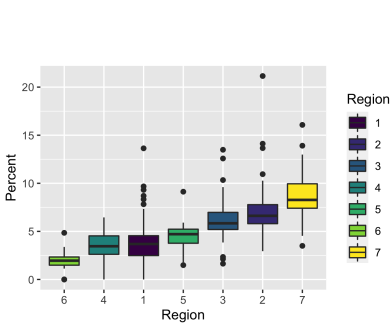
that are being left behind in the journey to reach the SDGs. From this perspective, it is clear that the south-west of the country is falling behind in terms of malnutrition in comparison to a much better performance by municipalities in the east where the best performers are also clustered (see yellow cluster). The fact that six different regions are identified may support the results of Miranda et al (2020) suggesting that Bolivia is a transitional country where both under-nutrition as well as over-nutrition coexist and may display profound inequalities depending on the level of education among other socioeconomic variables. Furthermore, they also found a statistically significant relationship between mothers' level of education (particularly for mothers with 7 to 12 years or more than 12 years of education) and levels of stunting or short stature among children less than five years old. Less constraints to education might create opportunities to break the cycle.

There is a less clear picture in terms of language as depicted in Figure 9. There are seven regions that can be identified (one more than for the case of children malnutrition). The west part of the country contains Region 3 and Region 2 with the lowest percentages of non-Spanish speakers, but it also hosts Region 7 containing the second highest percentage of non-Spanish speakers (in line with results of Figure 4 highlighting the role of Tupí-Guaraní language). Regions 1 and 6 are somewhat similar to 7, but mainly for the case of Quechua speaking populations. Region 5 may be suggesting regional clus-

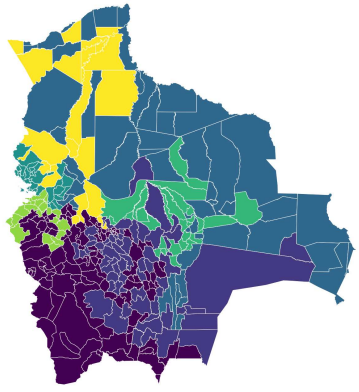
ters of predominantly Aymara speaking populations. There is evidence for the case of Bolivia indicating that students whose mother tongue is not Spanish tend to have lower attendance and performance at school (Patrinos and Psacharopoulos 1993). In an international context, Keller (2002) underlined the role of language for the diffusion of knowledge and technology. It is possible that children living in municipalities that are clustered in Regions 4 and Region 7 (and to some extent in Region 6 and Region 1) may benefit from educational experiences that also account for their bilingualism.

Figures 10 and 11 suggest that constraints to human development such as those posed by school dropout rates at the secondary level have some differences for males and females. Overall, males experience higher dropout rates. Regions with the highest dropout rates are located at the north of the country where an additional region appears displaying the highest dropout rates for the case of males. In terms of regions with the lowest dropout rates for females, Regions 7, 5, 1 and 4 include a combination of urban and rural municipalities located near the lower-half of the country (there are high dropout rates in Region 3 in the west and Region 6 in the east). Also, there seems to be a dual dynamic. On the one hand, in urban centers, educated women have increased labor market participation. For instance, Patrinos and Hurst (2007) find that women in La Paz have around 16 percent higher earnings than men. This is an incentive for completing school. On the other hand, Godoy et al (2005) state that in rural municipalities women do not usually participate in the labor market, so there is less pressure to dropout from school. Although they are different mechanisms, they both can help explaining lower dropout rates for females in general. For the case of males, the east part of the country seems to pose greater restrictions to human capital development as expressed by secondary school dropout rates (Regions 7, 2, 3, and 5). A complex challenge emerges for the department of La Paz as it hosts five different regional clusters (similar situation for Cochabamba hosting at least four types of regional clusters). Nevertheless, this is also an opportunity because this department hosts Region 6, which can provide useful lessons as it has the lowest dropout rates for males in the whole country.

Considering prevalent disparities within single departments, another constraint to human capital development could be studied through inequalities in the years of education. Figure 12 classifies municipalities into six regions. Five of them are located in the north-west and south-east of the country. The other, Region 5, shows the highest level of inequality (Gini coefficient higher

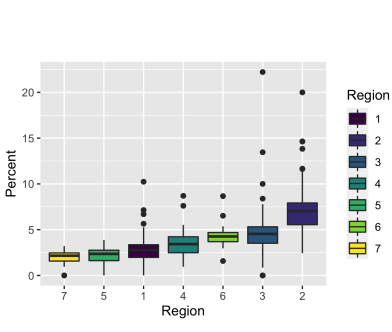


(a)

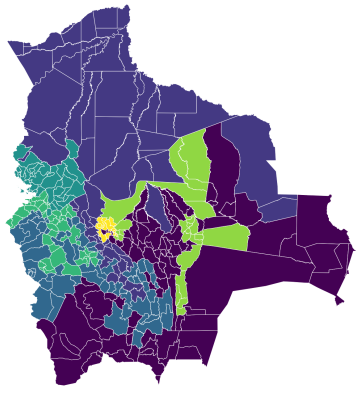


(b)

Fig. 10: Regionalization of secondary dropout of males



(a)

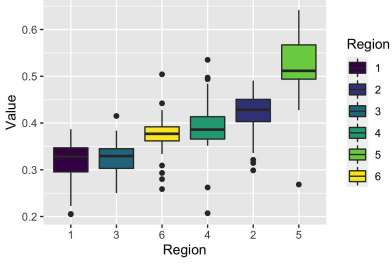


(b)

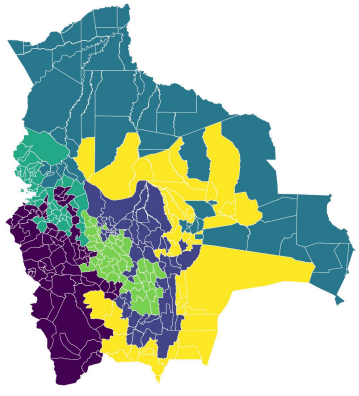
Fig. 11: Regionalization of secondary dropout of females

than 0.5) and is located around the lower middle of the country. This region includes the municipalities from four departments: Chuquisaca, Potosí, Oruro, and Cochabamba. Region 2 is the second cluster with the highest inequality and it is located right next to Region 5. A noteworthy fact is that the department of Cochabamba could be considered the most unequal in terms of years of education because the whole department overlaps with the two

most unequal clusters: Region 2 and Region 5. Cetrángolo et al (2017) noted that education is in itself a key component to reduce income inequality, foster growth, and strengthen democracy. Furthermore, as Kelley (1988) pointed out for the Bolivian context, closing education gaps could have the potential to reduce ethnic inequality as well. This is key to reach SDG10.



(a)

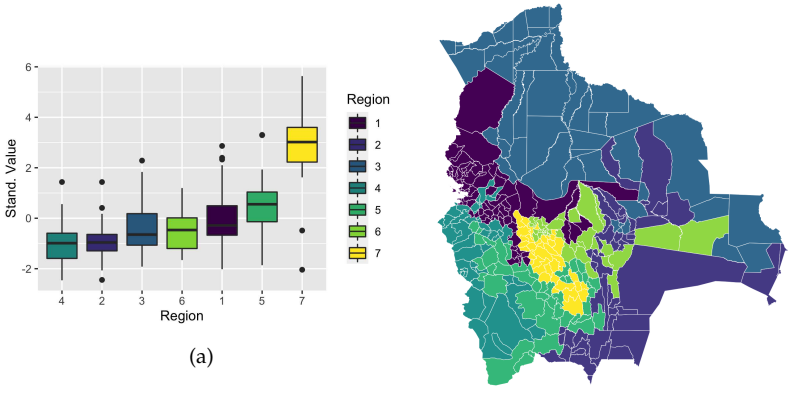


(b)

Fig. 12: Regionalization of inequality of years of education

Lastly, a multivariate Max-p approach was implemented based on all the previous variables (malnutrition, language, dropout rates for men and women, and inequality in years of education). The intention is to include more information that can be simultaneously considered for the identification of the regions. Figure 13 shows the results of this multivariate approach. Region 7 contains municipalities facing the largest challenges in terms of human capital constraints. Therefore, it is in this region where public policy at the national, department, and municipal levels need to be streamlined to address shared obstacles. If the municipalities of Region 7 continue under-performing, these results may affect the average of the entire country and constraint the overall achievement of national SDGs. Comparing the results of Figure 13 to all the previous results, it can be observed that Region 7 overlaps with Region 5 of Figure 12 (inequality in years of education), Region 4 of Figure 9 (non-Spanish speaking population), and Region 4 of Figure 8 (chronic malnutrition

of children). Indeed, this geographical overlapping reinforces the notion that human capital is a multidimensional concept.



(a) (b)

Fig. 13: Regionalization of integrated human capital constraints

Besides helping monitor regional progress in human capital, the clusters identified in this article help us pin down other issues to be explored in further studies. Looking at the results of Figure 13, it can be understood that municipalities in Region 4 have less obstacles to reach the SDGs in terms of human capital development. Nevertheless, little can be said about the actual returns to human capital investment. Likewise, it would be helpful to count with more data in terms of the quality of education (cognitive capacity, teachers' quality, etc.)

A better understanding of the spatial interactions between multidimensional indicators can inform public and private investment. 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. Unfortunately, lack of evidence is not the only problem. Aidt and Dutta (2007); Bonfiglioli and Gancia (2013); Atolia et al (2019); Acosta-Ormaechea and Morozumi (2017) presented convincing results noting that public investment in education is

often replaced by other types of investment.[15] This is because investment in the provision of determined public goods that diminish constraints to education may show benefits only after a couple of decades. Conversely, other types of investment may bring more perceptible benefits in the shorter term (such as roads for the first 9 to 10 years) while political incumbents hold office. Cetrángolo et al (2017) noted that spending per student is still very low and increases in investment are not always translated into better educational results. This is yet another reason for more detailed as well as integrated analysis to understand human capital development.

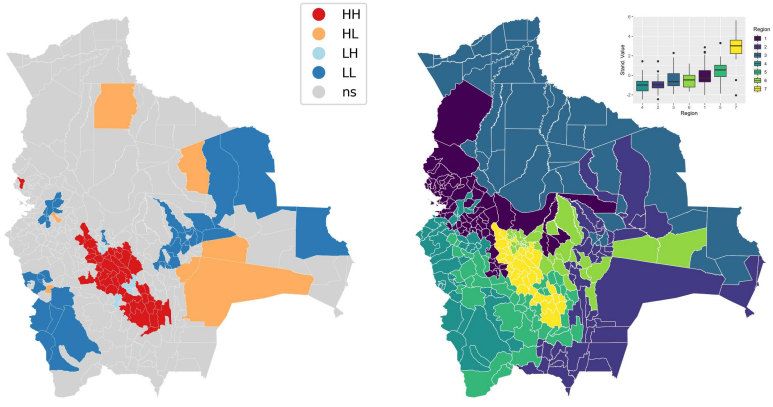
5 Discussion

5.1 Local Moran clusters vs Max-p clusters

Based on the integrated human capital constraints indicator, Figure 14 presents a comparison of the spatial clusters that are derived from the local Moran and the Max-p frameworks. Before focusing on this comparison, it is important to point out a central methodological difference between the frameworks. In contrast to the local Moran framework, the Max-p algorithm can easily accommodate multiple variables to identify spatial clusters. When multiple variables need to be considered in the local Moran approach, a common methodological choice is to use dimensionality reduction methods such as PCA (Anselin et al 2007). For instance, in Figure 14a, PCA is used to integrate five human capital constraints into one index (that is, the first component of the PCA). Then, based on this univariate index, local Moran clusters are compared to the clusters of the Max-p approach.

The main message of this comparison is that both methodologies should be considered as complements. In a first stage, the local Moran approach could be used to identify extreme cases. Then, in a second stage, the Max-p approach could be used to classify all the remaining cases into spatially contiguous areas. Intuitively, the relationship between the two approaches is that local Moran clusters tend to function as core centers of attraction, which largely affect the composition of the Max-p clusters. For instance, in Figure 14, it is clear that the high-high (HH) cluster of the local Moran approach is largely represented by region 7 of the Max-p approach. Similarly, the south-

[15] This is despite the fact that public investment in education, as compared to other type of expenditure, is found to significantly be associated with economic growth (Bose et al 2007).



(a) Local Moran clusters (b) Max-P clusters

Fig. 14: Local Moran clusters vs Max-P clusters

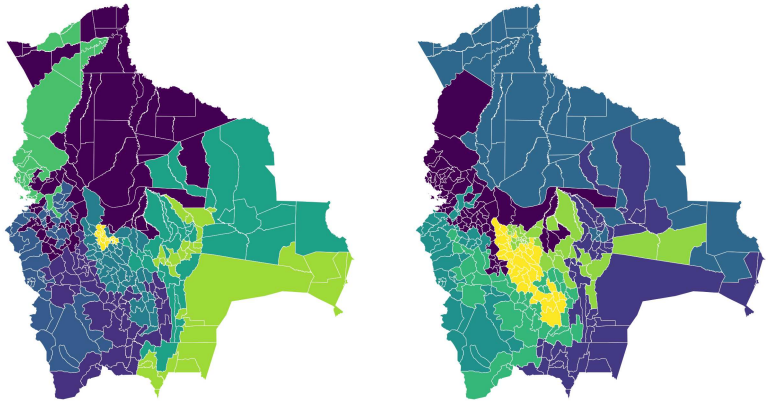
west low-low (LL) cluster is largely represented by region 4. It is worth noticing, however, that the composition and shapes of the clusters in both frameworks is not identical. This is due to the additional constraints of the Max-p framework, which are the spatial contiguity of regions and full classification.

A final caveat regarding the comparability of the two frameworks has to do with the treatment of spatial outliers. Besides identifying spatial clusters (HH and LL in Figure 14a), the local Moran framework identifies spatial outliers (HL and LH in Figure 14a). The Max-p framework, however, does not identify spatial outliers. This methodological difference emphasizes the notion that the local Moran approach and Max-p approach should be treated as complements rather than substitutes.

5.2 Univariate Max-p vs Multivariate Max-p

Similar to the local Moran clusters of Figure 14a, one could use a PCA to integrate the five human capital constraints into one index, and then apply the Max-p algorithm to identify univariate spatial clusters. Figure 15a shows the results of this exercise. Compared to the multivariate Max-p results (Figure 15b), the univariate Max-p identifies eight clusters instead of seven. In spite of this difference, the layout, size, and composition of the clusters are largely

similar. Thus, to a large extent, results from the multivariate Max-p approach tend to be robust to the dimensionality reduction procedure of the PCA.



(a) Univariate Max-p: 8 clusters (b) Multivariate Max-p: 7 clusters

Fig. 15: Univariate Max-p vs Multivariate Max-p

Although there are more similarities than differences when comparing the univariate and multivariate approach, it is often the case that policy makers may need to prioritize only one classification to monitor regional development. In this case, the multivariate approach should be prioritized as it includes more information to identify the clusters. Comparatively, the univariate approach suffers from less information content due to the application of the PCA methodology. Specifically, the univariate approach is based on the first principal component of the PCA, and although this component explains most of the variation of the data, an analysis based on the the entire variation of the data is preferred.

5.3 Max-p clusters based on alternative connectivity structures

As explained in Section 3, 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 alternative connectivity structures based on distance and k-nearest neighbor criteria.

Figure 16 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.

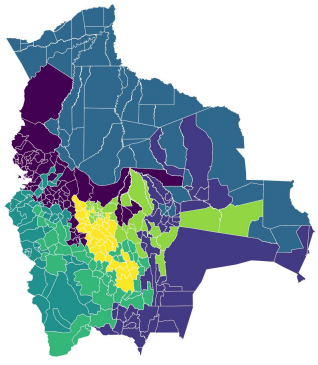
Panels (c) and (d) are based on two alternative definitions of distance: minimum distance band and inverse distance squared. In both cases, the number, size, and composition of the clusters are the same. Regional contiguity, however, is a missing feature. The main result of this comparison is that regions 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.

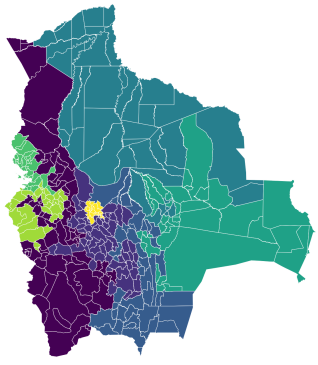
Taken together, the results of Figure 16 suggest that the identification of Max-p spatial clusters is sensitive to alternative connectivity structures and regional design objectives. On the one hand, if the objective is to achieve both spatial contiguity and high regional compactness, the rook contiguity 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.

6 Concluding remarks

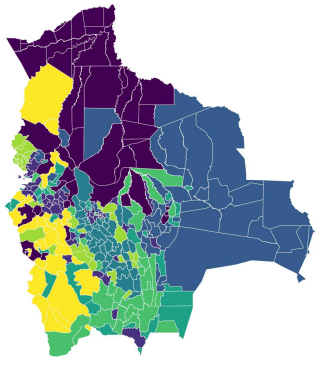
In this article, through the lens of modern geospatial analytical methods, we identify clusters of regions facing similar human capital constraints. Specifi-



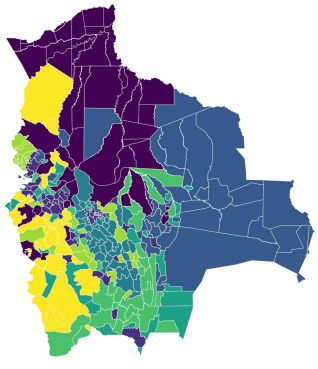
(a) Queen contiguity: 7 clusters



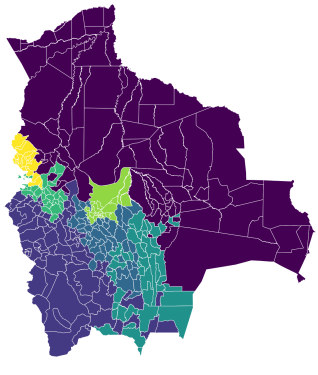
(b) Rook contiguity: 8 clusters



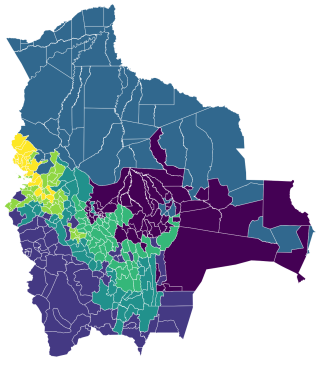
(c) Distance band: 8 clusters



(d) Inverse distance squared: 8 clusters



(e) Six nearest neighbors: 7 clusters



(f) Eight nearest neighbors: 7 clusters

Fig. 16: Max-p clusters based on alternative connectivity structures

cally, using a novel dataset of 339 municipalities in Bolivia, 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. In addition to value similarity in each of the previously listed constraints, the clusters identified in this article are characterized by spatial contiguity.

Two methodological approaches are implemented to identify geographically contiguous clusters. On the one hand, we use the spatial dependence framework of Anselin (1995) to identify regional hot spots (high-value clusters), cold spots (low-value clusters), and spatial outliers. On the other, we use the integer programming approach of 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 spatial dependence analysis are four fold. Chronic malnutrition is mostly located in the lower center and lower west part of the country. Non-Spanish speaking populations are located in the lower center of the country. Secondary dropout rates (both male and female) are located in the north of the country. High education inequality is located in the lower center of the country.

Results of the regionalization analysis indicate that Bolivia can be divided into seven to eight geographical regions that face similar constraints in the accumulation of human capital. The borders of these regions are largely different to those indicated by the political map of the country. Although, from a political administration standpoint, Bolivia is divided into nine regions; from a human capital constraints standpoint, Bolivia can be divided into eight regions at most. 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 largely coordinated across multiple local governments and actively supported by the national government.

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 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 depen-

dence and regionalization analyses helps overcome these issues and provides a more comprehensive evaluation of the geographical system being studied.

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 municipal-level data. First, the sensitivity of the Max-p algorithm can also be re-evaluated using alternative initialization and size parameters. Furthermore, the regionalization of Bolivia can be re-evaluated using alternative clustering frameworks. Among them, the spatially constrained clustering approach of Assuncao et al (2006) seems to be the closest alternative. A compelling feature of this framework is that clusters are identified by pruning the minimum spanning tree created from the spatial weights matrix.

Acknowledgements We would like to thank the anonymous referees for their thoughtful comments and suggestions, which have significantly improvement of the manuscript. We also acknowledge the comments and suggestions from the members of the QuaRCS-lab (https://quarcs-lab.org).

References

Acosta-Ormaechea S, Morozumi A (2017) Public Spending Reallocations and Economic Growth Across Different Income Levels. *Economic Inquiry* 55(1):98–114, DOI 10.1111/ecin.12382

Aidt TS, Dutta J (2007) Policy myopia and economic growth. *European Journal of Political Economy* 23(3):734–753, DOI 10.1016/j.ejpoleco.2006.05.002

Anselin L (1995) Local indicators of spatial association—lisa. *Geographical analysis* 27(2):93–115

Anselin L (2017) Cluster Analysis (3): Spatially Constrained Clustering Methods

Anselin L, Sridharan S, Gholston S (2007) Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. *Social Indicators Research* 82(2):287–309

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

Assuncao RM, Neves MC, Camara G, da Costa Freitas 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 BG, Marto R, Melina G (2019) Investing in Public Infrastructure: Roads or Schools? *Macroeconomic Dynamics* pp 1–30, DOI 10.1017/S1365100519000907

Barrenechea Vargas MH (2004) A spatial study about municipal poverty in Bolivia

Barro RJ (2001) Human Capital and Growth. *American Economic Review* 91(2):12–17

Becker GS, Murphy KM, Tamura R (1990) Human Capital , Fertility , and Economic Growth. *Journal of Political Economy* 98(5):12–37

Bivand RS, Wong DW (2018) Comparing implementations of global and local indicators of spatial association. *Test* 27(3):716–748, DOI 10.1007/s11749-018-0599-x, URL `https://doi.org/10.1007/s11749-018-0599-x`

Bonfiglioli A, Gancia G (2013) Uncertainty, Electoral Incentives and Political Myopia. *Economic Journal* 123(568):373–400, DOI 10.1111/ecoj.12029

Bose N, Haque ME, Osborn DR (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. Tech. rep., United States Department of Labor, Washington D.C.

Canavire-Bacarreza G, Duque JC, Urrego JA (2016) Moving citizens and deterring criminals : innovation in public transport facilities

Canelas C, Niño-Zarazúa M (2019) Schooling and Labor Market Impacts of Bolivia's Bono Juancito Pinto Program. *Population and Development Review* 45(S1):155–179, DOI 10.1111/padr.12270

Cetrángolo O, Curcio J, Calligaro F (2017) Evolución reciente del sector educativo en la región de América Latina y el Caribe: los casos de Chile, Colombia y México. Tech. rep., Search Results Web result with site links Economic Commission for Latin America and the Caribbean (ECLAC), Santiago

Chetty R, Hendren N, Katz LF (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, DOI 10.3386/w21156

Church R, Duque JC, Restrepo D (2020) The p-innovation ecosystems model. *arXiv preprint arXiv:200805885*

Cliff AD, Ord JK (1981) *Spatial processes: models & applications*. Taylor & Francis

Collin M, Weil DN (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 Gonzalez LM (2003) *Evolucion reciente de las disparidades economicas territoriales en America Latina: estado del arte, recomendaciones de politica y perspectivas de investigacion*. Economic Commission for Latin America and the Caribbean, Santiago de Chile

Delboy M (2019) Determinants of School Attendance rate for Bolivia: A spatial econometric approach

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

Duque JC, Ramos R, Suriñach J (2007) Supervised regionalization methods: A survey. *International Regional Science Review* 30(3):195–220

Duque JC, Anselin L, Rey SJ (2012) The max-p-regions problem. *Journal of Regional Science* 52(3):397–419, DOI 10.1111/j.1467-9787.2011.00743.x

Duque JC, Patino J, Ruiz LA, Pardo JE (2013) Quantifying Slumness with Remote Sensing Data, DOI 10.2139/ssrn.2390737

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

Fischer M (1980) Regional taxonomy: a comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics* 10(4):503–537

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

Gaspar V, Amaglobeli D, Garcia-Escribano M, Soto M (2019) Fiscal Policy and Development: Human, Social, and Physical Investment for the SDGs. *IMF Staff Discussion Note* 19(3):3–45

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, DOI 10.1111/j.1468-0084.1996.mp58001002.x

Godoy R, Karlan DS, Rabindran S, 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, DOI 10.1016/j.econedurev.2003.11.008

Hansen P, Jaumard B, Meyer C, Simeone B, Doring V (2003) Maximum split clustering under connectivity constraints. *Journal of Classification* 20(2):143–180

Hanushek EA (2013) Economic growth in developing countries: The role of human capital. *Economics of Education Review* 37:204–212, DOI 10.1016/j.econedurev.2013.04.005

Hanushek EA, Woessmann L (2008) The role of cognitive skills in economic development. *Journal of Economic Literature* 46(3):607–668, DOI 10.1257/jel.46.3.607

Hornberger NH (1992) Literacy in South America. *Annual Review of Applied Linguistics* 12:190–215, DOI 10.1017/s0267190500002221

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6):417–441, DOI 10.1037/h0071325

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

Jollife IT, 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), DOI 10.1098/rsta.2015.0202

Keller W (2002) Geographic localization of international technology diffusion. *American Economic Review* 92(1):120–142, DOI 10.1257/000282802760015630

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

Kuscevic CMM, del Río Rivera MA (2013) Convergencia en Bolivia: Un enfoque espacial con datos de panel dinámicos. *Revista de Economia del Rosario* 16(2):233–256

Law S, Neira M (2019) An unsupervised approach to geographical knowledge discovery using street level and street network images. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, pp 56–65

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

Lefkovitch LP (1980) Conditional clustering. *Biometrics* pp 43–58

Maclsaac DJ, Patrinos HA (1995) Labour Market Discrimination Against Indigenous People in Peru. *The Journal of Development Studies* 32(2):218–233, DOI 10.1080/00220389508422412

Manly BFJ, Navarro Alberto JA (2017) *Multivariate Statistical Methods: A Primer*. CRC Press, Boca Raton

Maravalle M, Simeone B (1995) A spanning tree heuristic for regional clustering. *Communications in statistics-theory and methods* 24(3):625–639

Mardia KV, Kent TJ, Bibby J M (1994) *Multivariate Analysis*. Academic Press, London

Martínez PP (1990) Towards standardization of language for teaching in the Andean countries. *Prospects* 20(3):377–384, DOI 10.1007/BF02195079

McKenzie D, 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(Nov):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. *Revista Latinoamericana de Desarrollo Económico* (31):77–98, DOI 10.35319/lajed.201931347

Mincer J (1984) Human capital and economic growth. *Economics of Education Review* 3(3):195–205, DOI 10.1016/0272-7757(84)90032-3

Miranda M, Bento A, Aguilar AM (2020) Malnutrition in all its forms and socioeconomic status in Bolivia [published online ahead of print, 2020 Mar 11]. *Public Health Nutrition* pp 1–8, DOI 10.1017/S1368980019003896

Morales R, Galoppo E, Jemio LC, Choque MC, Morales N (2000) Bolivia: Geografía y Desarrollo Económico

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

Patrinos HA (1997) Differences in education and earnings across ethnic groups in Guatemala. *Quarterly Review of Economics and Finance* 37(4):809–821, DOI 10.1016/s1062-9769(97)90005-3

Patrinos HA, Hurst ME (2007) Indigenous language skills and the labor market in a developing economy: Bolivia. *The Economics of Language: International Analyses* 48(2):473–489, DOI 10.4324/9780203963159

Patrinos HA, 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, DOI 10.2307/3338733

Pearson K (1901) 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, DOI 10.7748/ns.4.48.17.s35

Psacharopoulos G (1993) Ethnicity, Education, and Earnings in Bolivia and Guatemala. *Comparative Education Review* 37(1):9–20, DOI 10.1086/447161

Psacharopoulos G, Patrinos HA (2018) Returns to investment in education: a decennial review of the global literature. *Education Economics* 26(5):445–458, DOI 10.1080/09645292.2018.1484426

Rey SJ, Sastré-Gutiérrez ML (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

SDSN-Bolivia (2020) Atlas Municipal de los Objetivos de Desarrollo Sostenible en Bolivia

Soruco Carballo CF (2012) Espacio , convergencia y crecimiento regional en Bolivia: 1990 – 2010

Urquiola MS, Andersen L, Antelo E, Evia JL, Nina O (1990) Geography and Development in Bolivia: Migration, Urban and Industrial Concentration, Welfare, and Convergence: 1950-1992, DOI 10.2139/ssrn.1814660

Vernier Fujita LD, Pengo Bagolin I, Fochezatto A (2020) Spatial distribution and dissemination of education in Brazilian municipalities. *The Annals of Regional Science* DOI https://doi.org/10.1007/s00168-020-01020-3

Wise S, Haining R, Ma J (1997) Regionalisation tools for the exploratory spatial analysis of health data. In: Recent developments in spatial analysis, Springer, pp 83–100