Exploring the diversity of informal settlements in Brazil: a clustering approach to qualitative field survey data

Luis Felipe Bortolatto da Cunha[1], Flávia da Fonseca Feitosa[1], Antonio Paez[2], Carolina Moutinho Duque de Pinho[1]

[1]Laboratório de Estudos e Projetos Urbanos e Regionais (LEPUR) – Universidade Federal do ABC (UFABC)
São Bernardo do Campo, SP, Brasil

[2] School of Earth, Environment & Society – McMaster University
Hamilton, ON, Canada

luis.cunha@ufabc.edu.br, flavia.feitosa@ufabc.edu.br, paezha@mcmaster.com, carolina.pinho@ufabc.edu.br

Abstract. *Upgrading informal settlements and improving the living conditions of their dwellers is still a challenge. In addition, it is important to produce up-to-date information about them. This exploratory research aimed at expanding the knowledge on informal settlements through the identification, description and comparison of typologies present in Brazil. It relied on an extensive qualitative field survey applied to 157 municipalities in six regions. The territorial scope and format of data introduced methodological challenges that were addressed by using an ensemble k-modes clustering approach. The results illustrate the diversity of informal settlements through four typologies with distinctive characteristics.*

1. Introduction

UN-Habitat [2020] estimates that more than 1.6 billion people or 20% of the world's population live in inadequate housing, of which one billion reside in slums or informal settlements. Therefore, upgrading these territories and the living conditions of their dwellers remains a challenge to be faced. But in addition to the need for public policies and interventions, there is a need for accurate, consistent, and up-to-date spatial information on informal settlements that can support planning [Thomson et al. 2020, Feitosa et al. 2021, Abascal et al., 2022, dos Santos et al. 2022]. Its characteristics are diverse, varying between countries, regions, municipalities and even within the settlement. The identification of typologies helps in understanding this diversity and, consequently, in planning better strategies for mapping informal settlements.

Many studies have been developed aiming at identifying and describing typologies of deprived urban settlements in Brazil. Some of them focus on policy and/or intervention needs, such as the typologies from the National Housing Policy [Denaldi 2009], the National Policy for Regularization of Urban Land [Almeida 2018] and the Greater ABC Housing Diagnosis [CIGABC 2016]. Denaldi et al. [2018] also defined typologies of urban fabric, which were applied in precarious settlements mapping in the Metropolitan Region of Baixada Santista [Feitosa et al. 2021]. All these typologies were defined beforehand, relying on technical expertise on the subject. Other studies were based on a data-driven approach to identify the typologies, which is more similar to the perspective embraced in this paper: Oliveira [2021] used machine learning and open data to distinguish four typologies of deprived settlements in São Paulo city, and Feitosa et al. [2022] applied a k-means clustering methodology to find 3 to 4 local typologies of informal settlements in six Brazilian regions, which showed local variations in terms of density, infrastructure precariousness, income and insertion in urban fabric.

Population Census data is often used to identify and characterize deprived urban settlements, but they have poor temporal resolution (usually 10 years) and have shown to be insufficient in capturing the diversity of informal settlements [Denaldi 2022]. Remote sensing data is also being widely used and can be updated more frequently, but they are limited to physical-territorial characteristics observable from space [Kohli et al. 2012, Kuffer et al. 2016]. Employing field survey data presents both the advantage of adopting information that is not currently present in the Census and of assimilating regional differences indirectly included from the field researcher's perspective. To the best of our knowledge, no study relied on a data-driven approach using qualitative data and the same territorial scope as the analysis here proposed.

Therefore, this research focus at exploring the diversity of informal settlements in Brazil, aiming at expanding the knowledge on its characteristics. To achieve this goal, general typologies of informal settlements are identified, described, and compared highlighting their spatial attributes. The analysis relies on an extensive qualitative field survey applied to six Brazilian regions that imposes several challenges due to its qualitative nature and magnitude. An ensemble k-modes clustering approach was adopted to identify the typologies, followed by an exploratory data analysis of the results.

2. Data

The analysis relied on a spatial dataset produced in the scope of the project 'Informal Urban Settlements' (NUI) (*Pesquisa de Núcleos Urbanos Informais no Brasil*), developed by the Institute for Applied Economic Research (Ipea) in partnership with the National Housing Secretariat (SNH) [Krause & Denaldi 2022]. The project's extent included six regions in Brazil (*Polos – Belo Horizonte, Brasília, Juazeiro do Norte, Marabá, Porto Alegre and Recife*) as an effort to represent Brazilian urban heterogeneity. The study area (Figure 1) comprised 157 municipalities, although only 137 municipalities had informal settlements reported.

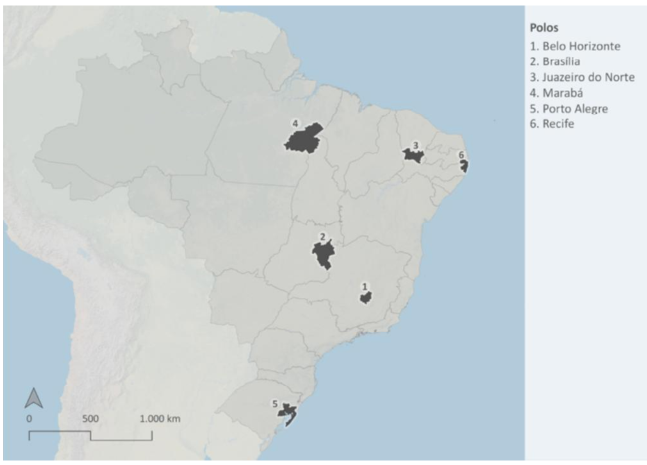

Figure 1. Study area map [Petrarolli et al. 2021]

In addition to mapping informal settlements, field researchers collected qualitative information on each of them, detailed in Table 1. The choice of variables considered the objective of the NUI project, which was to build an initial reference of settlements whose characteristics could make them eligible for federal land tenure regularization programs, and the possibility of collecting information through field survey [Petrarolli et al. 2021].

Table 1. Description of the dataset

Variable	Categories
Urban contiguity: Location in relation to the urban network of the city	**Central:** completely inserted in the central urban network of the city
	Periphery or remote: On the periphery of the urban network or completely isolated from the urban network of the city
Establishment time: Approximate time of occupation	**Less than 10 years**
	More than 10 years
	No information
Real estate dynamics: Variation in the number of dwellings in the recent period (2 years)	**Increase:** Rapid or slow emergence of new dwellings
	Stable
	Decrease: Rapid or slow decrease in the number of dwellings
	No information
Special Zones of Social Interest (ZEIS) intersection	**Yes**
	No
	No information
Environmental Conservation Areas intersection	**Yes**
	No
Permanent Preservation Areas of water resources intersection	**Yes**
	No
Risk situation: Has there ever been an occurrence of landslide, flooding, or fire? Is it located over a contaminated area (landfill, dump etc.), gas pipeline, water mains, or under a high voltage line?	**Yes**
	No
	No information
Risk susceptibility: Is it, or part of it, inside an area of medium or high susceptibility to risk?	**Yes**
	No
	No information
Street layout and access to parcels	**Mostly present:** Predominance of adequate vehicular routes that structure the occupation and most parcels have direct access to the street
	Mostly absent: Few or no blocks are structured by vehicular routes and most of the parcels are accessed through narrow roads, alleys, or staircases
	No information
Parcel subdivision: Are the parcels ordered and well-defined?	**Well-defined** parcels subdivision predominates
	Undefined parcels subdivision predominates
	No information
Buildings distance: Is there distance between the buildings?	**Present:** there is a distance between the buildings (retreats and spaces between the dwellings)
	Absent: little or no distance between the buildings (retreats and spaces between the dwellings)
	No information
Buildings quality	**Adequate:** popular standard dwellings predominate (consolidated and with acceptable dimensions and quality)
	Partial: Dwellings in different stages of consolidation and precariousness (popular standard, improvised materials, precarious constructions etc.)
	Inadequate: Precarious and/or improvised dwellings predominate
	No information
Infrastructure condition: How is the access to basic infrastructure?	**Adequate** infrastructure condition
	Partial: infrastructure is partially adequate
	Inadequate: Lack of basic infrastructure or very precarious existing infrastructure
	No information

3. Methodology

Identifying typologies, from a data mining perspective, consists in assigning labels to observations in such a way that those with the same label are similar to each other and those with different labels are dissimilar in some sense. Since the labels are not available beforehand, this process is known as unsupervised classification or clustering – a well-established statistical method for finding groups in data, used in many contexts and in many disciplines. It is essentially an exploratory analysis, which can uncover hidden patterns in data [Jain et al. 1999, Bandyopadhyay & Saha 2013, Hennig et al. 2015].

The k-means algorithm is arguably the most popular clustering method because it is computationally easy, fast, and memory-efficient. But on its original formulation, it employs the Euclidean distance as a measure of dissimilarity and minimizes the cost function by changing the means of clusters, being only applicable to numeric values [MacQueen 1967]. Based on the k-means, Huang [1998] proposed the k-modes algorithm, designed to derive clusters from categorical data by using the simple matching distance as dissimilarity measure and minimizing the cost function by changing the modes of clusters. The simple matching distance between two vectors can be defined by the total mismatches of the corresponding attribute categories of the two objects.

As any machine learning algorithm, it has several limitations. The most relevant one is that it may not converge to an optimal solution because it is dependent of the choice of the initial cluster (seed). For this reason, the k-modes algorithm was placed into a cluster ensemble, which is a methodology for combining different clustering results into a single consensus cluster. Cluster ensembles are highly recommended as they have shown to be better than any standard clustering algorithm at improving the accuracy and robustness of clustering and they can be applied to the output of the k-modes algorithm – in fact, the cluster ensemble can be treated as a problem of clustering categorical data (using the output of several clusters as features) [Strehl & Ghosh 2002, Ayad & Kamel 2010, Ghosh & Acharya 2011, Boongoen & Iam-On 2018].

Figure 2 shows the methodology steps, which included: (1) Exploratory data analysis to choose the number of clusters, (2) Clustering ensemble model (k-modes and consensus clustering), and (3) Descriptive and comparative analysis of the output.

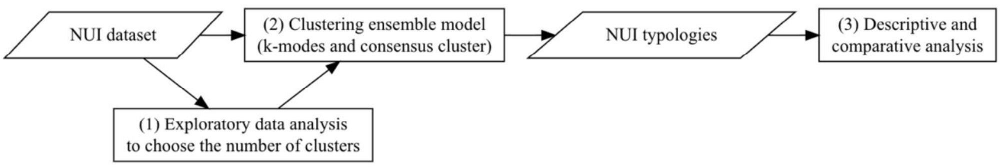

Figure 2. Methodology steps

Since the number of clusters (k) in the k-modes algorithm must be defined beforehand, an attempt of addressing this problem was made by creating an elbow plot showing the dissimilarity decay by k (measured as the mean and standard deviation simple matching distance across 100 models with different seeds). Using this plot as reference, some alternatives were tested and analyzed subjectively for the definition of the final number of clusters. Afterwards, the k-modes algorithm [Weihs et al. 2005] was applied 1000 times with different seeds to a random subsample of 80% of the NUI dataset. The results were introduced into a consensus clustering function [Chiu & Talhouk 2018], that relabels the clusters to make them comparable and find the majority voting across each observation, resulting in the final NUI typologies. The NUI typologies were then analyzed, described, and compared. All computation was done with R statistical computing software and the code is available on GitHub[1].

4. Results and discussion

A first assessment of the dataset aimed at choosing the number of typologies (or clusters). Figure 3 shows the dissimilarity decay as the number of clusters increase, indicating that the clustering methodology can reduce the within-cluster dissimilarity. Nonetheless, it does not highlight one unique solution. Therefore, a subjective analysis of the output considering 3 to 6 clusters was made, leading to the choice of k = 4 for its higher coherence and interpretability.

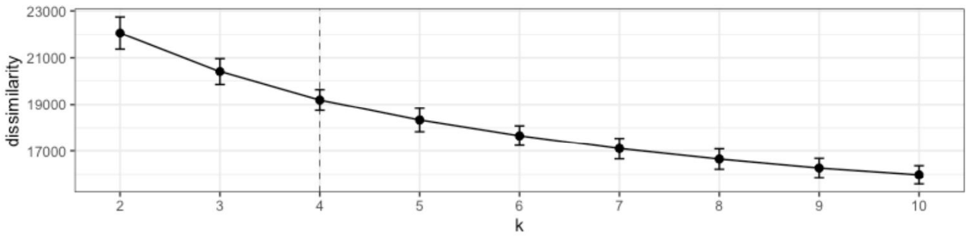

Figure 3. Elbow plot

Table 2 shows the proportion of typologies by region, Table 3 presents a summary of the characteristics of each typology, Figure 4 illustrates a settlement of each typology, including an aerial view and a street view, and Figure 5 shows maps with the spatial distribution of the typologies by region. The analysis of this information enabled a preliminary description of each typology:

- **Typology 1** (T1) includes informal settlements that usually have appropriate street layout (95%), well-defined parcel subdivision (89%), distance between buildings (86%), and adequate buildings quality (73%). They are usually located at peripheric locations (71%), and few had risk situation reported (19%) or risk susceptibility mapped (12%), even though only 18% have adequate infrastructure. This is the most common typology (40% of all settlements mapped), and it is predominant in Brasília (93%), Juazeiro do Norte (75%), Marabá (52%), and Porto Alegre (48%), and significantly present in Belo Horizonte (30%) and Recife (16%) as well. Figure 4a shows a settlement at the periphery of the urban network (aerial view and street view), which has a regular street layout but lacks basic infrastructure (paved roads) in part of it. Figure 5 illustrates that it is widespread in Brasília and Juazeiro do Norte, and present mostly as isolated settlements in other regions.

[1] https://github.com/luisfelipebr/geoinfo2022

- **Typology 2** (T2) informal settlements are characterized by their central location (90%), lack of distance between buildings (85%), being often mapped as special zone of social interest in local plans (76%), and presence of risk susceptibility (59%). The condition of buildings (10% inadequate) and infrastructure (9% inadequate) are comparatively better than T3 and T4. This typology is more common in Belo Horizonte (35%), Porto Alegre (34%), and Recife (43%) regions. Figure 4b displays a very high-density settlement, in which several buildings have more than one floor built, and the terrain is steep, which can lead to risk situation and make access to the parcels difficult (made by ramps and stairs). Figure 5 highlights the location of these settlements, usually at the central municipality.

- **Typology 3** (T3) main characteristics are the presence of permanent preservation areas of water resources (97%) and incidence of risk situation (86%), which are often associated. They are mostly located in central areas (74%) and more than half of them had an increase in the number of dwellings in the recent period (58%). This typology stands out in Belo Horizonte (31%) and Marabá (23%) regions. Figure 4c demonstrates a settlement close to water resources, which intersects with environmental conservation areas and permanent preservation areas of water resources, and with risk situation and lack of basic infrastructure reported. This typology is not as usual as T1 and T2 (15% of all informal settlements mapped) and is mostly visible in Belo Horizonte region (Figure 5).

- **Typology 4** (T4) settlements have the worst basic infrastructure condition (68% are inadequate) and quality of buildings (only 3% are adequate). Its parcel subdivision is not well-defined (15%) and there is a lack of information about risk situation (63%) and risk susceptibility (46%), which might be related to the fact that a lot of these settlements were recently established (29%) and the number of dwellings in recent period increased in more than half of the settlements in this typology (66%). This typology appears mostly in Recife region, where 33% of the settlements are of Typology 4, and they are usually not located at the central municipality (Figure 5). Figure 4d shows an isolated occupation in Marabá region, with inadequate buildings quality and infrastructure condition.

Table 2. Proportion of typologies by region

Region	T1 (n = 1995)	T2 (n = 1567)	T3 (n = 733)	T4 (n = 673)
Belo Horizonte	0,30	0,35	0,31	0,04
Brasília	0,93	0,03	0,03	0,01
Juazeiro do Norte	0,75	0,05	0,13	0,07
Marabá	0,52	0,10	0,23	0,15
Porto Alegre	0,48	0,34	0,07	0,11
Recife	0,16	0,43	0,08	0,33

Table 3. Proportion of settlements in each category by variable and typology

Variable	Categories	All settlements (n = 4968)	T1 (n = 1995)	T2 (n = 1567)	T3 (n = 733)	T4 (n = 673)
Urban contiguity	Central	**0,56**	0,29	**0,90**	**0,74**	0,41
	Periphery or remote	0,44	**0,71**	0,10	0,26	**0,59**
Establishment time	Less than 10 years	0,13	0,14	0,06	0,10	0,29
	More than 10 years	**0,86**	**0,85**	**0,94**	**0,89**	**0,69**
	N.A.	0,01	0,01	0,00	0,01	0,02
Real estate dynamics	Increase	0,33	0,32	0,10	**0,58**	**0,66**
	Stable	**0,66**	**0,68**	**0,89**	0,40	0,33
	Decrease	0,01	0,01	0,02	0,02	0,01
	N.A.	0,00	0,00	0,00	0,00	0,00
Special Zone of Social Interest	Yes	**0,42**	0,24	**0,76**	0,40	0,16
	No	**0,42**	**0,52**	0,17	**0,50**	**0,63**
	N.A.	0,16	0,24	0,07	0,10	0,21
Environmental Conservation Areas	Yes	0,22	0,31	0,10	0,24	0,19
	No	**0,78**	**0,69**	**0,90**	**0,76**	**0,81**
Permanent Preservation Areas	Yes	0,49	0,37	0,41	**0,97**	0,47
	No	**0,51**	**0,63**	**0,59**	0,03	**0,53**
Risk situation	Yes	0,33	0,19	0,34	**0,86**	0,16
	No	**0,43**	**0,58**	**0,53**	0,04	0,21
	N.A.	0,24	0,23	0,13	0,10	**0,63**
Risk susceptibility	Yes	0,34	0,12	**0,59**	0,35	0,37
	No	0,18	0,11	0,28	0,16	0,18
	N.A.	**0,49**	**0,77**	0,14	**0,48**	**0,46**
Street layout	Mostly present	**0,76**	**0,95**	**0,59**	**0,89**	0,46
	Mostly absent	0,24	0,05	0,41	0,10	**0,54**
	N.A.	0,00	0,00	0,00	0,00	0,00
Parcel subdivision	Well-defined	**0,66**	**0,89**	**0,59**	**0,67**	0,15
	Undefined	0,33	0,11	0,41	0,32	**0,85**
	N.A.	0,00	0,01	0,00	0,00	0,00
Buildings distance	Present	**0,57**	**0,86**	0,14	**0,60**	**0,63**
	Absent	0,43	0,13	**0,85**	0,40	0,36
	N.A.	0,00	0,01	0,00	0,00	0,00
Buildings quality	Adequate	**0,47**	**0,73**	0,41	0,28	0,03
	Partial	0,32	0,13	**0,42**	**0,40**	**0,60**
	Inadequate	0,12	0,05	0,10	0,21	0,25
	N.A.	0,09	0,08	0,08	0,11	0,13
Infrastructure condition	Adequate	0,20	0,18	0,32	0,11	0,03
	Partial	**0,55**	**0,68**	**0,52**	**0,65**	0,16
	Inadequate	0,18	0,09	0,09	0,18	**0,68**
	N.A.	0,07	0,06	0,07	0,06	0,13

(a) T1 - ARIS Pôr do Sol, Brasília

(b) T2 - Alto Vera Cruz, Belo Horizonte

(c) T3 - Vila dos Sargentos, Porto Alegre

(d) T4 - Vila Café, Marabá

Figure 4. Typologies examples (Google Satellite and Google Street View)

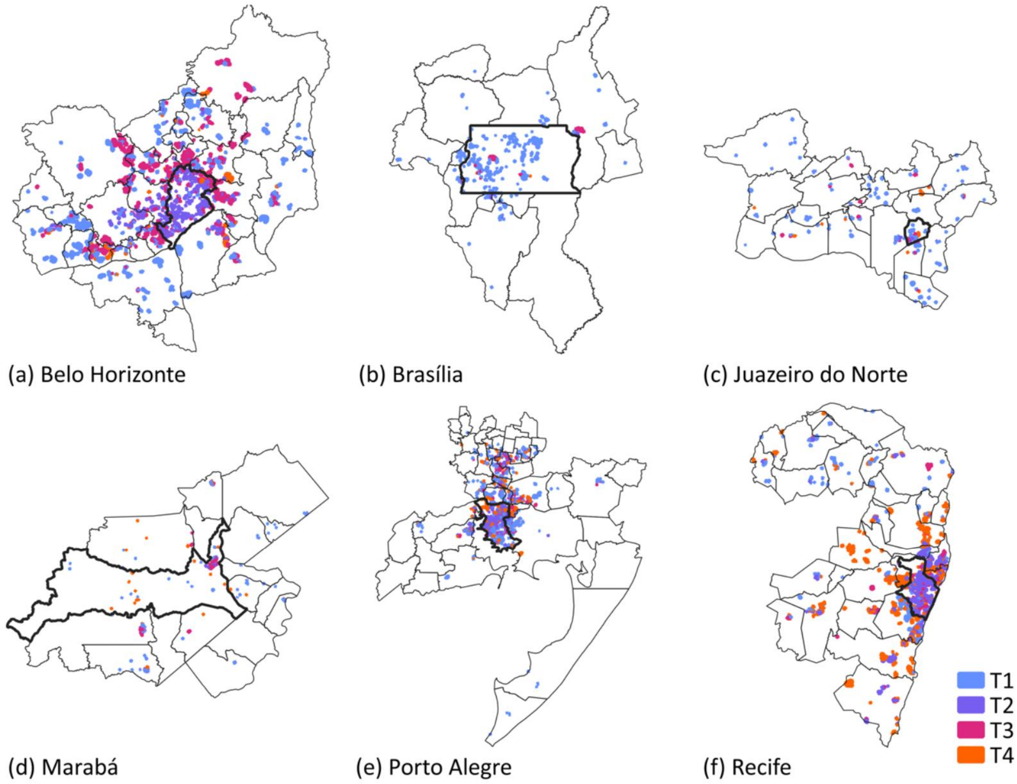

Figure 5. Spatial distribution of typologies by region

The typologies identified through the clustering approach are quite distinct and display four different facets of the precariousness present in informal settlements in Brazil. The analysis benefited from the subjective perception of the mapping experts since the definition of adequate/inadequate varies according to the context of each region. The results show how location is a relevant characteristic of these settlements, expressed through the urban contiguity variable, and highlight regional differences, with some regions being less diverse and others more diverse.

5. Conclusion

The exploratory analysis presented helped to emphasize the diversity of informal settlements in Brazil. The results show four typologies with distinctive characteristics, which were further described and compared. These typologies are widespread across the six regions, but in varying proportions. By mapping the typologies, it was also possible to understand the geographical nature of precariousness in each region.

The results can feed into the mapping of informal settlements and the process of planning public policies and interventions. Further research may explore other methods for choosing the number of clusters, arriving at a consensus cluster, or deriving the clusters altogether. Acknowledging the limitations of the clustering technique adopted and of qualitative data, the analysis could be refined by adopting mixed data from multiple data sources. The methods and results are openly available, which may lead to additional insights on informal settlements diversity.

Acknowledgements

Ipea and SNH for providing the dataset used in the analysis. The first author was partially funded with a scholarship provided with the support of the Government of Canada.

References

Abascal, A., Rothwell, N., Shonowo, A., Thomson, D., Elias, P., Elsey, H., ... & Kuffer, M. (2022). “Domains of deprivation framework” for mapping slums, informal settlements, and other deprived areas in LMICs to improve urban planning and policy: A scoping review. Computers, environment and urban systems, 93, 101770.

Almeida, G. M. J. A. de (2018). Estudo da demanda para regularização fundiária de assentamentos urbanos no âmbito do Programa Papel Passado. Santo André: Ed. UFABC, 2018.

Ayad, H. G., & Kamel, M. S. (2010). On voting-based consensus of cluster ensembles. Pattern Recognition, 43(5), 1943-1953.

Bandyopadhyay, S., & Saha, S. (2013). Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Berlin: Springer.

Boongoen, T., & Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. Computer Science Review, 28, 1-25.

Chiu, D. S., & Talhouk, A. (2018). diceR: an R package for class discovery using an ensemble driven approach. BMC bioinformatics, 19(1), 1-4.

CIGABC – Consórcio Intermunicipal do Grande ABC (2016). Diagnóstico Habitacional Regional do ABC. (Report).

Denaldi, R. (2009). Assentamentos precários: identificação, caracterização e tipologias de intervenção. In Denaldi, R. & Rosa, J. (Eds.). Curso à Distância: Planos Locais de Habitação de Interesse Social. Brasília, Ministério das Cidades/Aliança de Cidades.

Denaldi, R., Petrarolli, J., Gonçalves, G., Moraes, G. (2018). Tecidos Urbanos e a Identificação de Assentamentos Precários na Região Metropolitana da Baixada Santista. Proceedings of III URBFavelas, Salvador, BA, Brasil.

Denaldi, R. (2022). Identificação e caracterização da precariedade habitacional: desafios conceituais e metodológicos. In Krause, C., Denaldi, R. (Eds.) (2022). Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Ipea [Book in press].

dos Santos, B. D., de Pinho, C. M. D., Oliveira, G. E. T., Korting, T. S., Escada, M. I. S., & Amaral, S. (2022). Identifying Precarious Settlements and Urban Fabric Typologies Based on GEOBIA and Data Mining in Brazilian Amazon Cities. Remote Sensing, 14(3), 704.

Feitosa, F., Vasconcelos, V., Pinho, C., Silva, G., Gonçalves, G., Danna, L., & Lisboa, F. (2021). IMMerSe: An integrated methodology for mapping and classifying precarious settlements. Applied geography, 133, 102494.

Feitosa, F., Gonçalves, G., Cunha, L. (2022). Aglomerados Subnormais e Núcleos Urbanos Informais: Uma Análise Comparativa. In Krause, C., Denaldi, R. (Eds.) (2022). Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Ipea [Book in press].

Ghosh, J., & Acharya, A. (2011). Cluster ensembles. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 1(4), 305-315.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). Handbook of cluster analysis. CRC Press.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Krause, C., Denaldi, R. (Eds.) (2022). Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Ipea [Book in press].

Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. Computers, Environment and Urban Systems, 36(2), 154-163.

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space—15 years of slum mapping using remote sensing. Remote Sensing, 8(6), 455.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297).

Oliveira, L. (2021). The diversity of deprived areas: applications of unsupervised machine learning and open geodata (Master's thesis, University of Twente).

Petrarolli, J. et al. (2021). Produto 11 – Relatório de consolidação das pesquisas de campo sobre aspectos físicos e territoriais de núcleos urbanos informais. Pesquisa de Núcleos Urbanos Informais no Brasil (Report).

Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, 3(Dec), 583-617.

Thomson, D. R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H., ... & Kabaria, C. (2020). Need for an integrated deprived area “slum” mapping system (IDEAMAPS) in low-and middle-income countries (LMICs). Social Sciences, 9(5), 80.

UN-Habitat (2020). World Cities Report 2020: The Value of Sustainable Urbanization. UN-Habitat, Nairobi.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.