Spatial patterns of intestinal parasite infections among children and adolescents in some indigenous communities in Argentina

Carlos Matías Scavuzzo,[1-4] Micaela Natalia Campero,[1-3] Rosana Elizabeth Maidana,[1] María Georgina Oberto,[1] María Victoria Periago,[3,4] Ximena Porcasi[2]

[1]Human Nutrition Research Center, School of Nutrition, Faculty of Medical Sciences, National University of Córdoba, Córdoba; [2]Mario Gulich Institute for Higher Space Studies, National University of Córdoba, National Commission of Space Activities, Falda del Cañete, Córdoba; [3]National Council for Scientific and Technical Research, Buenos Aires; [4]Mundo Sano Foundation, Buonos Aires, Argentina

Correspondence: Ximena Porcasi, Instituto de Altos Estudios Espaciales Mario Gulich, Comisión Nacional de Actividades Espaciales, Falda del Cañete, Córdoba, Argentina.
Tel.: +54.9.351.3331600.
E-mail: ximena.porcasi@conae.gov.ar.

Key words: machine learning, intestinal parasites, spatial analysis, Argentina.

Contributions: SCM, CMN, MRE, conceptualization, investigation, methodology, formal analysis, writing - review and editing, visualization; OMG, conceptualization, investigation, methodology, writing - review and editing, visualization; PX, PMV, conceptualization, investigation, methodology, writing - review, and editing.

Conflict of interest: the authors declare that they have no conflict of interest.

Ethics approval and consent to participate: approval was obtained from the Ethics Committee of the Provincial Commission for Research in Health Sciences, Teaching and Research Programme, Human Resources Directorate of the Ministry of Health of Salta Province, Resolution 1480/2011.

Informed consent: informed consent was obtained from the participants' parents or guardians.

Funding: none.

Availability of data and materials: available from the corresponding author upon request.

Received: 27 November 2023.
Accepted: 3 May 2024.

Geospatial Health 2024; 19:1279
doi:10.4081/gh.2024.1279

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Abstract

Argentina has a heterogeneous prevalence of infections by intestinal parasites (IPs), with the north in the endemic area, especially for soil-transmitted helminths (STHs). We analyzed the spatial patterns of these infections in the city of Tartagal, Salta province, by an observational, correlational, and cross-sectional study in children and adolescents aged 1 to 15 years from native communities. One fecal sample per individual was collected to detect IPs using various diagnostic techniques: Telemann sedimentation, Baermann culture, and Kato-Katz. Moran's global and local indices were applied together with SaTScan to assess the spatial distribution, with a focus on cluster detection. The extreme gradient boosting (XGBoost) machine-learning model was used to predict the presence of IPs and their transmission pathways. Based on the analysis of 572 fecal samples, a prevalence of 78.3% was found. The most frequent parasite was *Giardia lamblia* (30.9%). High- and low-risk clusters were observed for most species, distributed in an east-west direction and polarized in two large foci, one near the city of Tartagal and the other in the km 6 community. Spatial XGBoost models were obtained based on distances with a minimum median accuracy of 0.69. Different spatial patterns reflecting the mechanisms of transmission were noted. The distribution of the majority of the parasites studied was aligned in a westerly direction close to the city, but the STH presence was higher in the km 6 community, toward the east. The purely spatial analysis provides a different and complementary overview for the detection of vulnerable hotspots and strategic intervention. Machine-learning models based on spatial variables explain a large percentage of the variability of the IPs.

Introduction

Neglected tropical diseases (NTDs), induced by bacteria, viruses, parasites, or fungi, are transmitted either directly from person to person or indirectly through intermediate vectors/hosts. These are a group of 20 health conditions that predominantly affect populations residing in socioeconomically disadvantaged environments, particularly in remote rural areas and marginalized urban neighborhoods. These populations often have scarce access to healthcare, education, clean water, basic sanitation, and hygiene (Ault *et al*., 2014; Brindha *et al.*, 2021; Dueñas *et al*., 2021; Iomini *et al*., 2021). The prevention, elimination, and eradication of NTDs need a holistic approach that addresses environmental and social determinants of health. The control of these diseases has been incorporated into global health agendas, including the Millennium Development Goals, the Sustainable Development Goals 2030, the Sustainable Health Agenda for the Americas 2018-2030, and the UN's roadmap on NTDs 2021-2030 (WHO, 2010, 2021, 2023). Within these 20 conditions, soil-transmitted helminths (STHs) are the only intestinal parasites (IPs) included. These parasites require a passage through the soil to become infective and for transmission to occur, thereby linking their life cycle to environmental conditions (Juárez & Rajala, 2013). The parasites included within this group comprise five species that are exclusive to humans, such as *Ascaris lumbricoides, Trichuris trichiura, Strongyloides stercoralis*, and the hookworms *Necator americanus and Ancylostoma duodenale* (Engels & Zhou, 2020; Cuenca-León *et al*., 2021; Romero-Ramírez, 2022).

Adverse socioeconomic conditions, including overcrowding, limited access to clean water, and inadequate environmental sanitation, contribute to the heightened vulnerability of disadvantaged populations to IP infection (Bouzid *et al*., 2018; Anegagrie *et al.*, 2021; Candela *et al.*, 2023; Rivero *et al*., 2022; Tapia-Veloz *et al.*, 2023). These disparities, closely linked to poverty, disproportionately impact women and children in indigenous communities, leading to increased malnutrition and higher rates of infant and maternal mortality (Del Popolo *et al.*, 2014; Müller *et al.*, 2017; De Bourmont *et al*., 2020). Poverty maps in Argentina corroborate this, revealing a correlation between high poverty levels and regions inhabited by indigenous people (Echagüe *et al*., 2015).

Effective NTDs control strategies demand accurate and comprehensive data to target interventions and maximize resources. Knowing which population is most affected, where they live, and having information on the range of causal factors, or those most closely associated with the presence of a disease, including environmental aspects, is key for the development of useful tools. Geospatial technologies offer innovative tools for monitoring NTDs and parasitic infections, facilitating the mapping of parasite distribution, identifying high-prevalence zones, and establishing links between geographic distribution, environmental conditions, and associated risk factors (Anegagrie *et al*., 2021; Álvarez Di Fino *et al.*, 2022; Scavuzzo *et al.*, 2022; Candela *et al*., 2023).

In developing countries, where health needs are escalating and resources to address them are limited, spatial analyses and data visualization in geographic information systems (GIS) can aid efficient decision-making, particularly in challenging-to-access areas (Assaré *et al*., 2015). In Argentina, the application of GIS has been recognized as a powerful tool for studying IP infection, though studies identifying spatial patterns of IP infection in endemic areas remain scarce (Gamboa *et al*., 2014; Cociancic, 2019; Cociancic *et al*., 2019; Álvarez Di Fino *et al.*, 2020; Candela *et al*., 2023).

Additionally, traditional statistical models used to study the association between the presence of an infection and other factors are generally based on linear or generalized linear statistical approaches.

Given the complexity of some relationships between variables and the need to consider the problem as a whole interrelated process, these classical models are recently being complemented by machine learning (ML) techniques. ML is an emerging, promising field and offers a wide set of empirical tools to address both linear and non-linear aspects, fitting the complexities of big datasets with multiple variables and dependent dimensions (Weatherhead *et al.*, 1998). It is precious in scenarios where theoretical knowledge is limited but observational data are available for model training. ML has found applications in various fields, such as earth sciences and the development of biogeophysical information extraction algorithms (Brown *et al.*, 2008; Azamathulla *et al*., 2012), and its use in practical applications shows great promise and prospects (Lundberg & Lee, 2017). Among the most widely used ML algorithms are artificial neural networks, support vector machines, decision trees, random forests, and extreme gradient boosting (XGBoost) (Lary *et al*., 2016).

Given the aforementioned, this study aims to analyze the spatial patterns of IPs prevalence in children and adolescents from indigenous populations along National Route 86 in Tartagal (Salta), Argentina.

Materials and Methods

Study design

An observational, explanatory, and cross-sectional study was conducted. The population comprised boys, girls, and adolescents (BGA) aged 1 to 15 years from indigenous communities in the peri-urban neighborhoods and rural areas of the General José de San Martín Department in Salta Province, Argentina, from October 2021 to November 2022. The evaluated settlements are distributed in the NE area of the city of Tartagal (22°30'58.9'' S 63°48.079' W) and along National Route 86. Through non-probabilistic convenience sampling, participants with willingness to participate in the study and informed consent from their parents or guardians were included. Approval was obtained from the Ethics Committee of the Provincial Commission for Research in Health Sciences, Teaching and Research Program, Human Resources Directorate of the Ministry of Health of Salta Province, Resolution 1480/2011.

These indigenous communities are comprised of different ethnic groups, such as Wichí, Toba, Chorote, and Guaraní. Some communities live in the forest and are more isolated, while others are located on the outskirts of the city of Tartagal and have a higher population density. Tartagal is characterized by cultural diversity due to the presence of several native ethnicities and the population's continuous migration from the neighboring country of Bolivia. This characteristic produces a significant impact on the cultural, social, and economic profile of this community (Taranto *et al.*, 2003).

Data collection

For the detection of IPs, a single stool sample was collected from each individual. Wide-lid, sterile, and airtight containers were distributed. Instructions were provided on how to collect the stool samples: defecating onto a clean surface (bag or paper) without contact with water, urine, or dirt, and then using a wooden spatula to deposit a sufficient amount of the sample into the container. The samples were transported in a refrigerated box to the private laboratory of the Mundo Sano Foundation, Tartagal branch, where they were processed using the Telemann sedimentation technique, Baermann culture, and for STHs, egg counts per gram of feces using the Kato-Katz technique (Ministerio de Salud y Ambiente de la Nación, 2004; Gabrie *et al*., 2012). Processed fecal samples were then evaluated through optical microscopy for the identification of parasitic forms. For georeferencing of the homes, coordinate points were recorded in latitude and longitude form inside the dwelling using a Garmin GPS device.

Spatial analysis and modeling

In the first instance, to detect the presence of spatial patterns, Moran's global index was used (Wetchayont & Waiyasusri, 2021; Esri, 2023a). The characteristics of the detected clusters were evaluated employing a Bernoulli distribution model with the SaTscan program (Kulldorff, 1997). To corroborate and validate the presence of the clusters, Moran's local index and Local Indicators of Spatial Association (LISA) were applied, evaluating spatial coincidences between both tools and the behavior at the neighborhood level of the distribution (Esri, 2023b).

Additionally, a spatial ML model from the random forest family (XGBoost) was implemented to predict infection by IPs in general and infection by transmission route, through water, person-to-person contact, or the soil (STH) (Esri, 2023c). Only spatial characteristics were considered as regressor variables: the closest distance of each individual to the nearest health center, dense native vegetation, distance to Route 86, to the city, and extensive agricultural fields. A proxy for the spatial density of houses was also added. The average of the normalized difference buildings index (NDBI), obtained from Sentinel 2 imagery, was used in a 50-m buffer of the peri domicile. The procedure was carried out on the Google Earth Engine platform. Multiple XGBoost models were constructed to evaluate the role of space and different space-relative positions and distances on the pattern of infected individuals by IPs using spatially derived regressor variables as a proxy for stronger spatial and biological relationships. The objective was to examine variability based on spatial patterns and proxy distances, which can be potential explanations of underlying biological or transmission cycle nature. To adjust the model, all default parameters were used, namely: 100 trees with a leaf size of 1 and an average depth of 5. L2 regularization (λ) of 1.00, minimum loss reduction for splits (γ) of 0.00, and learning rate of 0.30 were tuned. Additionally, in cross-validation, the test percentage was modified to 30% of the total and to train with 70%. The validation split was 5. The mean squared error and the confusion matrix were considered for adjustment and validation. As a result of the analysis, the sensitivity and precision of each model (in the test) and the order of importance of the calculated spatial variables are reported. Also, the model outputs are included in case distribution maps. Descriptive statistics were performed with Stata 15 software (https://www.stata.com/). Spatial analysis and thematic cartography were conducted using Qgis 3.28 software (https://qgis.org/en/site/). Finally, the XGBoost model was implemented through ArcGIS 3.2 software (https://www.arcgis.com/).

Results

Table 1 presents the results from the analysis of the fecal samples. Of the total population, 572 fecal samples were obtained, with 78.3% testing positive for some species of IPs. The specific richness included 14 species of protozoans and helminths. The most prevalent species was *Giardia lamblia* (30.9%), followed by *Entamoeba coli* (29.3%) and *Hymenolepis nana* (25.9%). Regarding STH, hookworms were the most prevalent (20.7%). The study encompassed 717 BGA, residing in 202 households. Of these participants, 49.4% were identified as female and 50.6% as male, with an average age of 7.2±4.0 years. Regarding ethnicities, 70.1% were of Wichí origin, 9.3% Chorote, 9.3% Guaraní, 5.0% Criollo, and 5.7% were other ethnicities (Toba, Weenhayek, and mixed). One non-pathogenic species for which the patient is clinically evaluated to determine if treatment is needed, depending on medical judgment. Figure 1 illustrates the geospatial distribution of the homes of the participating BGA with IPs infection. In Figure 1, white dots indicate homes with at least one participant infected by IPs, while black dots denote uninfected households. Figure 1, where the circle's color around each house represents ethnicity. Furthermore, a concentration of Criollo and Guaraní children can be observed in the northwest of the area, tending to settle closer to the urban zone. It is also noted that the Chorote and Toba ethnicities are predominantly found within the high-risk cluster for STHs. Similarly, the Guaraní and Criollo ethnicities are primarily located in the low-risk cluster for STHs. Conversely, the Wennayeck ethnicity is situated within the high-risk cluster for water-transmitted species. In Table 2, the results of the spatial analysis are presented. Statistically significant spatial clusters of IPs, as well as STHs and waterborne species, are evident. Conversely, species transmitted through direct contact exhibited a random spatial distribution. Considering all parasites in general, the spatial prevalence across the entire area was 77.9%. Considering the relative risk of each cluster detected, STHs constituted the cluster with the highest risk and also the highest level of protection. In Figure 2A, clusters for all grouped species of IPs are presented. A high-risk cluster of larger size and a smaller low-risk area were both observed in the km 6 community. Regarding STHs, Figure 2B reveals two statistically significant groupings with an east-west pattern, being larger towards the west and smaller towards the east. The low-risk cluster (cluster 2 in blue) is adjacent to the urban area with a relatively low risk. On the other hand, the high-risk cluster 1 (in red) is located south of the km 6 community. A considerable number of cases were found in the high-risk cluster for water-transmitted species, which

Table 1. Prevalence of intestinal parasites in children from Tartagal, Salta (Argentina), 2021-2022.

Parasitological description (n=572 fecal samples) n (%)	
Positive	448 (78.3)
Negative	124 (21.7)
Prevalence by species	
Giardia lamblia	170 (30.9)
Entamoeba coli	169 (29.3)
Hymenolepis nana	149 (25.9)
Uncinarias	119 (20.7)
Blastocystis hominis	114 (20.0)
Entamoeba hartmanni	78 (13.7)
Endolimax nana	61 (10.6)
Cryptosporidium spp.	40 (6.9)
Entamoeba histolytica/dispar	40 (6.9)
Strongyloides stercoralis	34 (5.9)
Chilomastix mesnili	26 (4.6)
Enterobius vermicularis	16 (2.8)
Ascaris lumbricoides	15 (2.6)
Trichuris trichiura	1 (0.17)
Prevalence by transmission routes	
Waterborne transmission	366 (64.0)
Direct contact transmission	169 (29.6)
Soil transmission	140 (24.5)

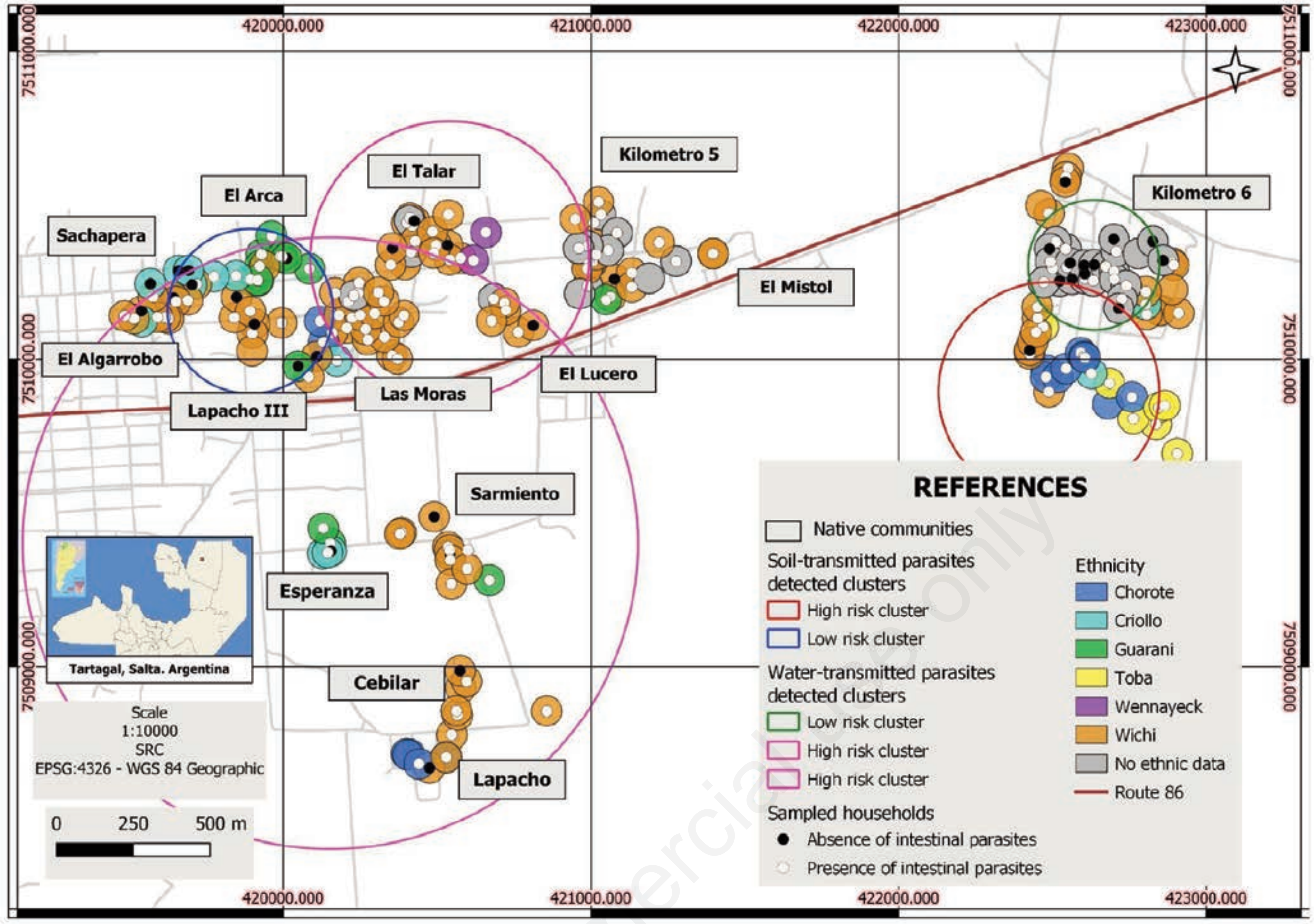

Figure 1. Geospatial distribution of households visited in the city of Tartagal and surroundings Salta (Argentina), 2021-2022. White dots, households with at least one case; black dots, negative households. The colors of the circles correspond to the ethnic groups. Map data © 2023 Google. Base map obtained through Quick Map Services QGIS plugin - QGIS Geographic Information System. Open source: Geospatial Foundation Project. http://qgis.osgeo.org.

Table 2. Characteristics of the detected clusters for intestinal parasites, soil-transmitted helminths, and water-transmitted parasites in children aged 1 to 15 years in Tartagal (Salta, Argentina), 2021-2022.

	Radius (m)	p	Relative risk	Number of people within the cluster	Positive observed cases (expected cases)
Intestinal parasites infected individuals (prevalence for the area: 77.9%)					
Cluster 1 (low risk in blue)	88	0.0003	0.5	58	28 (45.16)
Cluster 2 (high risk in red)	480	0.005	1.2	90	85 (70.07)
Soil-transmitted parasites infected individuals (prevalence for the area: 24.6%)					
Cluster 1 (high risk in red)	360	<0.001	3.7	64	45 (15.75)
Cluster 2 (low risk in blue)	270	<0.001	0.03	110	1 (27.07)
Water-transmitted parasites infected individuals (prevalence for the area: 43.9%)					
Cluster 1 (low risk in blue)	210	<0.001	0.3	106	27 (67.5)
Cluster 2 (high risk in red)	450	<0.001	1.5	121	105 (76.54)
Cluster 3 (high risk in red)	1	0.008	1.3	279	202 (176.49)

also covers a significantly larger area than the other groupings (Figure 2C). As for the water-transmitted IP species (Figure 2C), a low-risk cluster (cluster 1 in blue) was observed northwest of the km 6 community. In the area of the El Talar and Las Moras communities, the high-risk cluster (cluster 2 in red) was noted. Another high-risk cluster (cluster 3 in red), encompassing all communities in the southwest quadrant of the study area (Sarmiento, Esperanza, Cebilar, among others), was identified. In this instance, the pattern is inverse, being east-west but larger towards the east and smaller towards the west. Concerning the clusters by species (Table 3), hookworms exhibited clusters with higher relative risks. The highest cases were observed in the high-risk cluster for *G. lamblia*. It

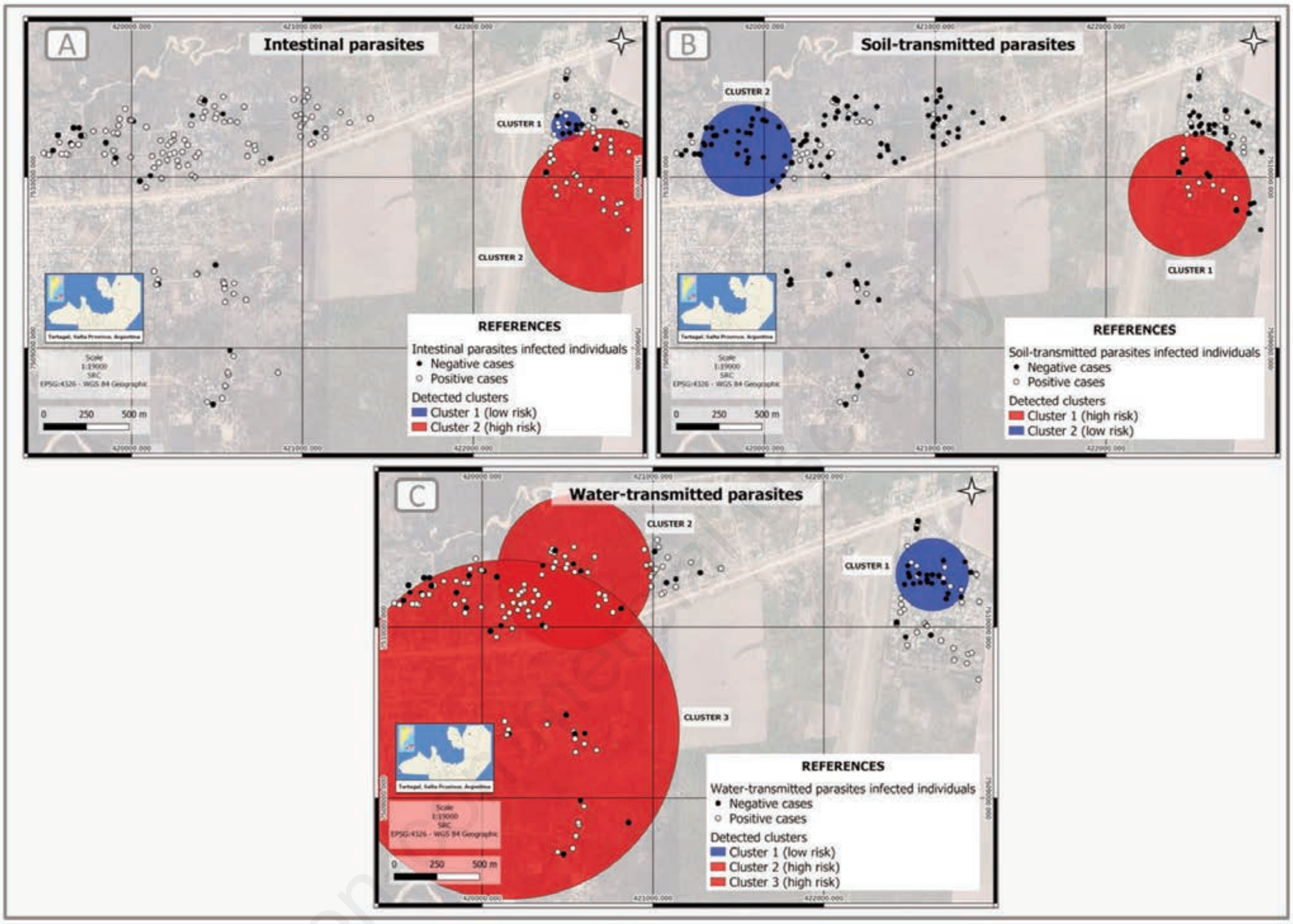

Figure 2. Spatial analysis of intestinal parasites, waterborne and soil-transmitted species, in children aged 1-15 years, Tartagal (Salta, Argentina), 2021-2022. Blue circles are clusters identified as low-risk, and red circles are high-risk clusters.

Table 3. Characteristics of the detected clusters for *Giardia lamblia*, Hookworm, and *Strongyloides stercoralis* in children aged 1 to 15 years in Tartagal (Salta, Argentina), 2021-2022.

	Radius (m)	p	Relative risk	Number of people within the cluster	Positive observed cases (expected cases)
***Giardia lamblia* infected individuals (prevalence for the area: 30.5%)**					
Cluster 1 (high risk in red)	600	<0.001	2.03	168	80 (51.8)
Cluster 2 (low risk in blue)	540	<0.001	0.42	180	28 (54.8)
Hookworm-infected individuals (prevalence for the area: 20.8%)					
Cluster 1 (high risk in red)	360	<0.001	4.5	64	43 (13.3)
Cluster 2 (low risk in blue)	290	<0.001	0.03	119	1 (24.7)
***Blastocystis hominis* infected individuals (prevalence for the area:19.8%)**					
Cluster 1 (zero risk in blue)	240	<0.001	0	115	0 (22.7)
Cluster 2 (high risk in red)	960	0.002	2.22	230	68 (45.5)

can be seen that *Blastocystis hominis* has the highest risk cluster 2 with the largest surface area with a diameter of 1.92 km. The distribution of *G. lamblia* clusters (Figure 3A) exhibits a pattern with a tendency corresponding to water-transmitted species, indicating a higher risk towards the east and a lower risk towards the west. For hookworms (Figure 3B), a high-risk cluster (cluster 1 in red) is observed in the southwest of the km 6 community. Towards the east and adjacent to the city, the low-risk cluster 2 (in blue) is established, also displaying a dispersion pattern typical of soil-transmitted species. Regarding *B. hominis* (Figure 3C), a high-risk cluster (cluster 1 in red) is arranged in a south-western direction and close to the urban area, encompassing communities such as El Lucero, Las Moras, Sarmiento, Esperanza, Cebilar, Lapacho I and II. In the area of the km 6 community, a cluster of zero risk was observed (cluster 2 in blue). This also coincides with the dispersal pattern of waterborne species. In Figure 4A, two partially overlap-

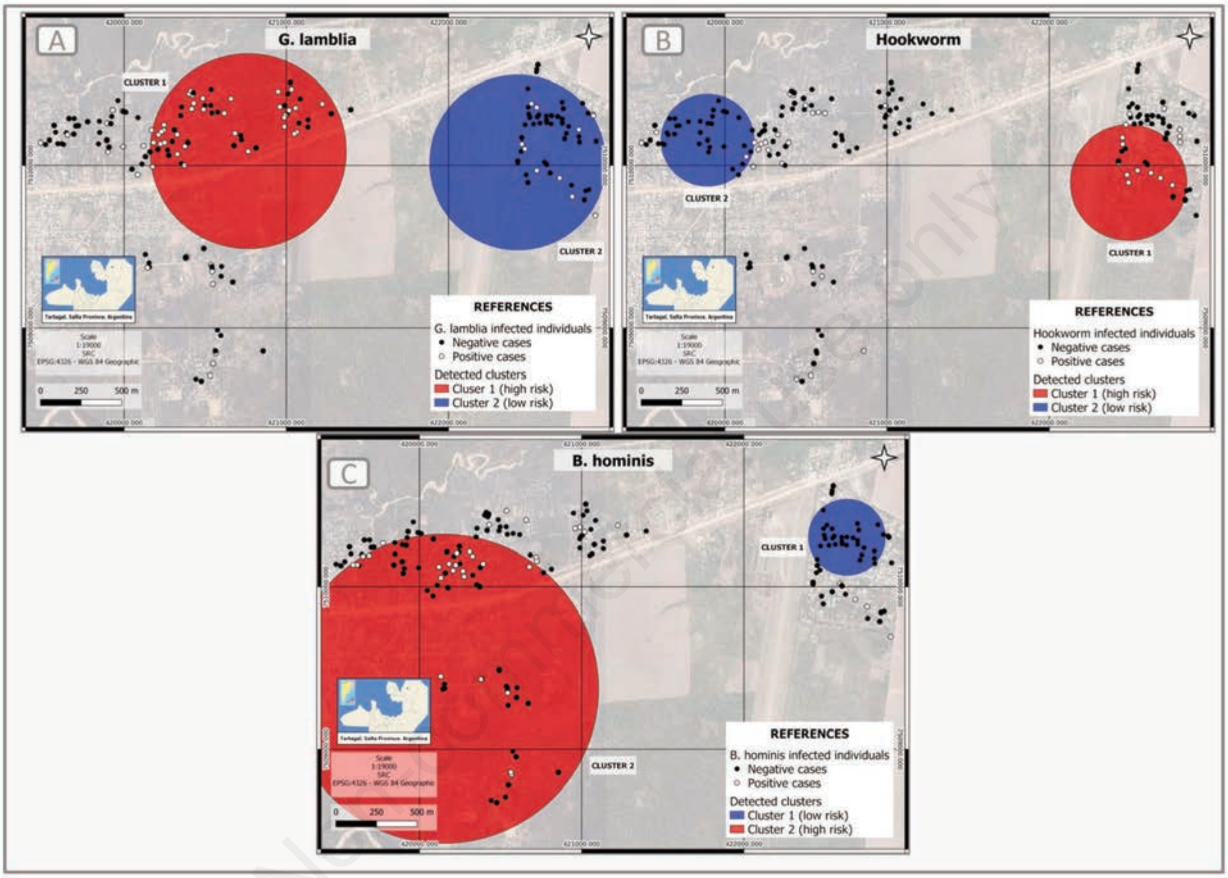

Figure 3. Spatial analysis of Giardia lamblia, hookworm, and *Blastocystis hominis* in children aged 1-15 years, Tartagal (Salta, Argentina), 2021-2022. The area included in this study was the area corresponding to the city of Tartagal, Salta province.

Table 4. Characteristics of clusters detected for *Entamoeba coli*, *Entamoeba hartmanni,* and *Endolimax* nana in children aged 1-15 years in Tartagal (Salta, Argentina), 2021-2022.

	Radius (m)	p	Relative risk	Number of people within the cluster	Positive observed cases (expected cases)
***Entamoeba coli* infected individuals (prevalence for the area: 28.9%)**					
Cluster 1 (low risk in blue)	210	<0.001	0.19	80	5 (23.1)
Cluster 2 (low risk in blue)	280	<0.001	0.11	58	2 (16.7)
***Entamoeba hartmanni* infected individuals (prevalence for the area: 13.2%)**					
Cluster 1 (zero risk in blue)	230	<0.001	0	114	0 (15.0)
Cluster 2 (high risk in red)	190	0.001	3.8	44	18 (5.81)
***Endolimax* nana infected individuals (prevalence for the area: 10.0%)**					
Cluster 1 (zero risk in blue)	240	<0.001	0	116	0 (11.5)

ping high-risk red clusters can be visualized in the km 6 community. The pattern is consistent with the spatial dispersion of higher-risk STH towards the east. For *C. mesnilli* (Figure 4C), cluster 1 (in red) of high risk includes the community of El Talar. Also, north of km 6 there is a blue cluster of zero risk. Table 4 shows the characteristics of the significant clusters for *E. coli, Entamoeba hartmanni,* and *Endolimax nana.* The cluster with the largest surface area was low risk for *E. coli.*

Figure 4A shows two statistically significant low-risk clusters for *E. coli.* They are located in the km community and are close and partially overlapping. For *E. hartmanni* (Figure 4B) one null risk cluster (in blue) is observed in the center of the km 6 community. Adjacent to the previous one but towards the outskirts of the community and in a southerly direction, a high-risk cluster is observed. Concerning *E. nana* (Figure 4C), a single cluster of zero risk is observed in community km 6. Likewise, for the species *E. vermicularis, H. nana, T. trichiura,* and those transmitted by direct contact, no statistically significant clusters were found, so it is assumed that they are randomly distributed in space.

The result of the spatial autocorrelation analysis through the global Moran index yielded a value of 0.099432 indicating a slight positive spatial autocorrelation. This suggests that there is a tendency towards clustering of areas with similar characteristics (presence or absence of parasites) more than would be expected if the pattern were completely random. The variance was 0.001084 and the Z-score was 3.072, suggesting that the observed Moran's index is approximately 3.07 standard deviations above that expected under the null hypothesis of random distribution. This is significant and strongly suggests that the observed pattern is not random. In turn, a p-value of 0.002121 was observed, indicating that there is less than a 0.21% chance of observing a Moran index as high or higher by chance if the data were truly random. This confirms the presence of significant spatial autocorrelation. Furthermore, the results of the local Moran's index analysis, or LISA, mirrored the clusters observed in the SatScan analysis, thus validating the findings through two different methodologies. The XGBoost spatial model that achieved the highest test sensitivity (0.86) was the one that predicted IPs; however, the one that achieved the highest test accuracy was the one that predicted STH (0.79). The median accuracy for all models ranged between 0.69 and 0.73. In training, the metrics were superior.

As shown in Table 5, in the prediction models for IPs, STH, and waterborne species, distance to extensive crops was the most important predictor, contributing between 24-27% of the prediction. However, in the direct contact-transmitted species model, NDBI was observed as the most important predictor with a 19% contribution to the model.

From the second to the fifth place, different orderings are observed among the rest of the variables, presenting different combinations but with similar contributions to the model prediction. In

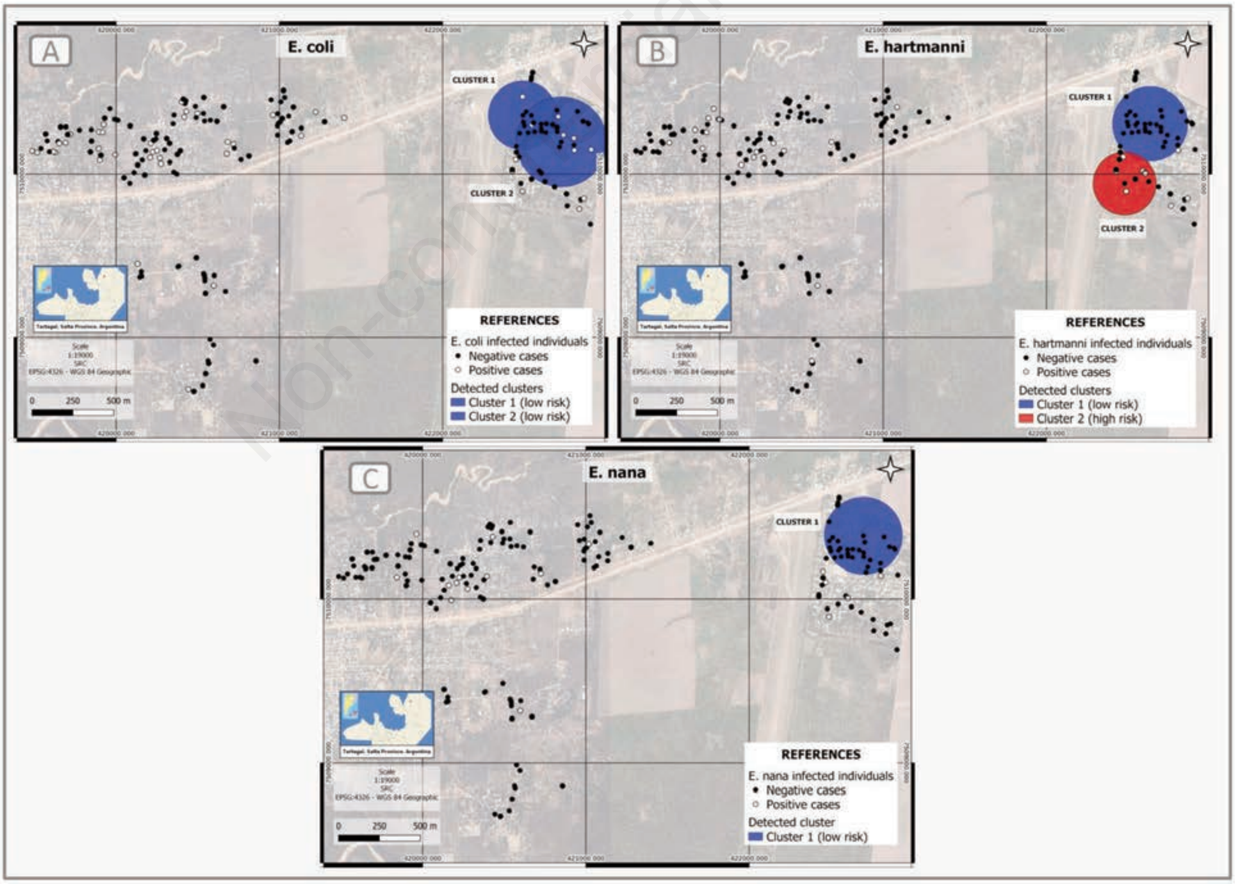

Figure 4. Purely spatial analysis of the presence of Entamoeba coli, *Entamoeba*. hartmanni, *Entamoeba and Endolimax nana,* in children aged 1-15 years, Tartagal (Salta, Argentina), 2021-2022. The area included in this study was the city of Tartagal, Salta province. Map data © 2024 Google.

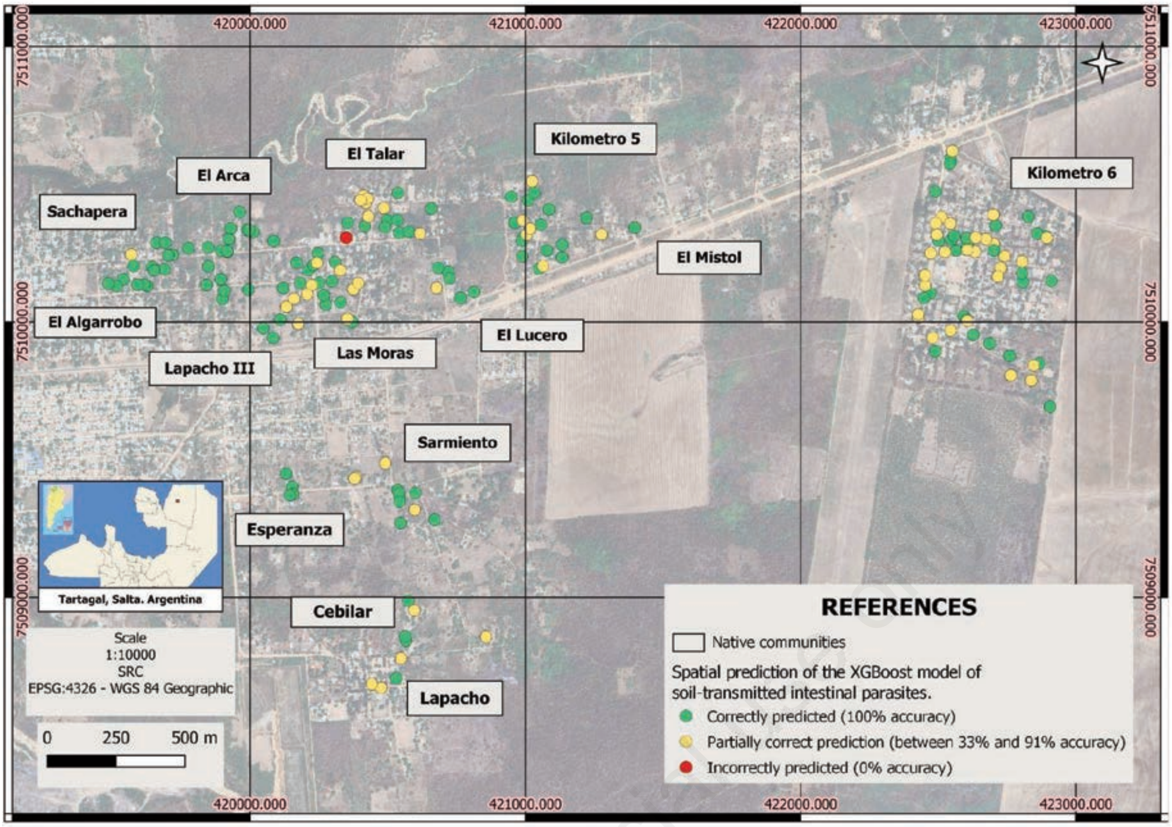
420000.000
421000.000
422000.000
423000.000
7511000.000
7510000.000
7509000.000
Kilometro 5
El Talar
El Arca
Sachapera
Kilometro 6
El Mistol
El Algarrobo
Lapacho III
Las Moras
El Lucero
Sarmiento
Esperanza
Cebilar
Lapacho
Tartagal, Salta. Argentina
Scale
1:10000
SRC
EPSG:4326 - WGS 84 Geographic
0
250
500 m
REFERENCES
Native communities
Spatial prediction of the XGBoost model of soil-transmitted intestinal parasites.
Correctly predicted (100% accuracy)
Partially correct prediction (between 33% and 91% accuracy)
Incorrectly predicted (0% accuracy)

Figure 5. Spatial output of the extreme gradient boosting model according to the accuracy of prediction. The area included in this study corresponds to the City of Tartagal (Salta, Argentina) and its surroundings. Map data © 2024 Google. Base map obtained through the ArcGIS Online map service - ArcGIS Geographic Information System. Owner: Esri, Inc. http://www.arcgis.com.

Table 5. Characteristics of spatial extreme gradient boosting models for intestinal parasites and their transmission pathways from distances in children aged 1-15 years in Tartagal (Salta, Argentina), 2021-2022.

Dependent variable	Sensitivity in test	Accuracy in test	Spatial feature importance and % contribution to the prediction
Intestinal parasites	0.86	0.70	1 - Distance to extensive agricultural crops (20%) 2 - NDBI (18%) 3 - Distance to route “86” (17%) 4 - Distance to city (17%) 5 - Distance to the dense native forest (16%) 6 - Distance to health centres (13%)
Species transmitted by direct human -to -human contact	0.85	0.63	1 - NDBI (19%) 2 - Distance to the dense native forest (18%) 3 - Distance to extensive agricultural crops (18%) 4 - Distance to route 86 (16%) 5 - Distance to health centres (16%) 6 - Distance to city (13%)
Water-transmitted species	0.68	0.60	1 - Distance to extensive agricultural crops (20%) 2 - Distance to city (19%) 3 - Distance to the dense native forest (18%) 4 - NDBI (17%) 5 - Distance to health centres (13%) 6 - Distance to route 86 (13%)
Soil -transmitted species	0.58	0.79	1 - Distance to extensive agricultural crops (22%) 2 - Distance to health centres (17%) 3 - Distance to city (16%) 4 - NDBI (15%) 5 - Distance to the dense native forest (15%) 6 - Distance to route 86 (14%)

NDBI, normalized difference buildings index.

any case, it is observed that the distance to Route 86 is placed as the last predictor in the STH and water transmission models, but it is observed that the distance to the city becomes more important.

In addition, Figure 5 shows the spatial output of the model indicating the classified (predicted) observations. A predominance of correctly classified values is observed, except for communities such as km 6, Las Moras, and El Talar, which present a lower accuracy in predictions.

The model's performance metrics, particularly the F1 scores and Matthew's Correlation Coefficient (MCC), display noteworthy strengths. The F1 score for non-infection (category 0) in training data is impressively high at 0.92, suggesting robustness in identifying negative cases. The slightly lower score of 0.74 for infection cases (category 1) in training still indicates reasonable effectiveness. Comparatively, in the test data, the model maintains a good level of accuracy, though with a slight drop, reflecting its generalisability. The MCC values of 0.66 in training and 0.41 in validation signify a decent overall predictive quality. These results collectively demonstrate the model's potential in accurately categorizing data, an essential attribute for applications in health-related fields.

Discussion

In Argentina, the distribution of IP infections in vulnerable children and adolescents differs according to the geographical region analyzed, with a decreasing prevalence from north to south and from east to west. This is due to wide environmental and socioeconomic variability (Zonta *et al*., 2007; Zonta *et al*., 2020; Cociancic *et al*., 2021). In the present work, in Tartagal City and its surroundings located in the north, the prevalence of IPs was 78.3%. In addition, another investigation carried out in the same study area (Salta) reported a prevalence of 94.6% for helminths and protozoa in similar populations (Menghi *et al*., 2007). It is important to highlight that, influenced by a range of socioeconomic factors, the health indicators of indigenous people tend to be less favorable compared to those of non-indigenous populations (Alfonso-Durruty & Valeggia, 2018). The most frequently identified species were *G. lamblia, E. coli, H. nana,* and hookworm. Similarly, other research conducted in the Wichi community reported a high prevalence of hookworm and *H. nana* (Taranto *et al*., 2003; Menghi *et al*., 2007), considering the area endemic for It is important to highlight here the relevance of the study of spatial patterns in health where the spatial associations of the data collected are revealed, providing a better understanding of the problem, in line with other works at different scales and with different objects of study in health (Celemín *et al.*, 2015; Diez Roux, 2015; Celemín & Velázquez, 2017; Longhi *et al*., 2022). In this study, a significant spatial autocorrelation in the distribution of the presence or absence of parasites was found, with a probability of less than 1% that this clustered pattern is the result of random variation. This indicates that it is very likely that underlying factors or processes are influencing the distribution of the presence or absence of parasites in the study area. In this work, only a few approximations are based on features extracted from remote sensing. Geographically, we observed a spatial distribution of IPs with an east-west pattern, polarized in two large foci, one to the east with a nucleus in the community of km 6 and far from the city, and the other pole in the peri-urban area to the west of the city. Regarding the eastern pole in the km 6 community, clusters of high relative risk of the presence of all IP species, including STH, in particular hookworm. This area is inhabited mainly by Wichi, Chorote, and Toba individuals. At the same time, relatively low-risk clusters of waterborne species are observed, particularly *G. lamblia, B. hominis, Chilomastix mesnilli, E. coli, and E. nana.*

When looking at the western pole adjacent to the city, there are clusters of high relative risk of waterborne species, in particular *G. lamblia, B. hominis*, and *C. mesnilli.* In the same area, clusters of low relative risk of STH, but also hookworm and *S. stercoralis* can be observed independently. However, in Argentina, this cluster detection methodology has been used for microorganisms such as *Shigella spp., D. immitis*, and *Leishmania spp* (Stelling *et al*., 2010; Hoyos *et al*., 2011; Esteban Mendoza *et al*., 2020); it is still an underexplored tool in the study of IPs in humans.

This study, by showing a good fit of the model (especially to predict negative individuals), represents an innovative and highly promising approach. As demonstrated in this study, ML tools have proven to be valuable in addressing problems where limited data are available for model training, a situation frequently encountered in the field of epidemiology (Scavuzzo *et al.*, 2022). In the global landscape characterized by increasing needs and constrained resources, the integration of open data science and ML techniques assumes a pivotal role in contributing to the generation of research outputs and facilitating decision-making tailored to local health priorities. In the field of epidemiology, where optimizing resources for field data collection is crucial, ML models in health prove to be of paramount importance in developing efficient models capable of learning from limited datasets. This fosters the widespread adoption of these technologies in entire communities grappling with analogous challenges (Bates *et al*., 2014; Gebreyes *et al*., 2014; Roski *et al*., 2014; Han *et al.*, 2015; Wiens & Shenoy, 2018).

The distance to crops was established as the most influential variable across all analyzed models. This correlation arises from considering such distance as a proxy indicator of a drastic transformation in land use, which represents a spatial pattern predictor and is inherently linked to the transmission route of STH. These parasites require specific conditions for their development and survival, including adequate soil moisture, vegetative cover, sunlight exposure, soil salinity, and soil pH, among other factors. Following a similar line, the NDBI is established as a proxy indicator for the spatial density of dwellings, exploring land use in relation to the transmission pathway of STH and also for IPs transmitted through water, since access to a safe water network, which is present in more built-up areas, defines a spatial pattern that could be explaining part of the variability observed in the model.

The observed modest predictive ability value in our spatial model highlights inherent limitations when relying solely on geographical inputs for predictive accuracy. This outcome underscores the complexities of using spatial data as a stand-alone predictor in epidemiological models, particularly for diseases influenced by a multitude of environmental and socio-economic factors. Despite this limitation, the primary objective of our study was not to forecast individual incidences of parasitic infections but rather to evaluate the critical role of spatial variables as epidemiological predictors. Our findings demonstrate that even isolated spatial variables can reveal significant, interpretable patterns essential for understanding disease transmission dynamics. Furthermore, while incorporating socio-economic dimensions could potentially refine these predictions, our focus was to establish a foundational understanding of the spatial determinants. This approach not only paves the way for integrating more comprehensive models in future research but also substantiates the utility of spatial proxy variables in epidemiology. Thus, our study contributes to the broader discourse in spatial epidemiology by highlighting the need for nuanced approaches that appreciate the spatial context of health phenomena, thereby enhancing the field's methodological arsenal (Collins *et al*., 2015). The relative or future importance of variables in ML models can be effectively evaluated using various techniques that help determine the stability and generalizability of those variables. This approach is particularly useful when one wishes to validate individual predictors within a broader context, even though their isolated predictive capacity may be limited. This underscores the importance of presenting and evaluating unidimensional models to understand their real contribution within the context of more complex and multidimensional predictive models. Although a model may perform poorly predictively when using a single dimension, it can still be useful to highlight the importance of that specific dimension as an epidemiological predictor, justifying its inclusion and assessment in research studies (Collins *et al*., 2015).

This study applied these cutting-edge tools to address a public health issue in a region of Argentina, where the prevalence of STH infection is recognized as one of the NIDs listed by the World Health Organization, highlighting the substantial potential of this methodology (WHO, 2010). However, in terms of strengths and from the perspective of panoramic epidemiology, the use of geomatics and ML in a geographical area of Argentina endemic to IP infection is highlighted for the analysis of spatial patterns of distribution. In this sense, space becomes another epidemiological variable to be analyzed for the detection of high- and low-risk areas.

Conclusions

This study substantiates the use of spatial analysis as a crucial methodological tool to detect clustering patterns of IP infections, revealing areas of high and low risk that would benefit from targeted socio-health strategies. We observed waterborne species predominantly distributed towards the west and near urban areas, while a higher prevalence of STH was noted in the km6 community, located eastward. This spatially focused analysis not only enhances our understanding of the geographical pattern of IP infections but also complements existing data, facilitating the identification of vulnerable zones for strategic interventions. This provides an important informed-decision tool for government and health authorities for optimizing resources and tailoring prevention strategies to approach the health problems of their population.

References

Alfonso-Durruty MP, Valeggia CR, 2018. Talla, peso e índice de masa corporal en niños y niñas wichí de Formosa, Argentina. Arch Argent Pediatr 116:359-64. [Article in English, Spanish].

Álvarez Di Fino EM, 2020. Aplicación de tecnologías geoespaciales para el análisis de la seguridad alimentaria y nutricional en la ciudad de Córdoba. Available from: http://hdl.handle.net/11086/17133. [Article in Spanish].

Álvarez Di Fino EM, Scavuzzo CM, Campero MN, Scavuzzo CM, Defagó MD, 2022. Explorando el uso de herramientas de sensores remotos y tecnologías geoespaciales aplicadas al problema multidimensional de la seguridad alimentaria. Uniciencia 36:1-15. [Article in Spanish].

Anegagrie M, Lanfri S, Aramendia AA, Scavuzzo CM, Herrador Z, Benito A, Periago MV, 2021. Environmental characteristics around the household and their association with hookworm infection in rural communities from Bahir Dar, Amhara Region, Ethiopia. PLoS Negl Trop Dis 15:e0009466.

Assaré RK, Lai YS, Yapi A, Tian-Bi YNT, Ouattara M, Yao PK, Knopp S, Vounatsou P, Utzinger J, N'Goran EK, 2015. The spatial distribution of Schistosoma mansoni infection in four regions of western Côte d'Ivoire. Geospat Health 10:345.

Ault SK, Catalá Pascual L, Grados-Zavala ME, Gonzálvez García G, Castellanos LG, 2014. El camino a la eliminación: un panorama de las enfermedades infecciosas desatendidas en América Latina y el Caribe. Rev Peru Med Exp Salud Pública 31:319-25. [Article in Spanish].

Azamathulla HM, Ab Ghani A, Fei SY, 2012. ANFIS-based approach for predicting sediment transport in clean sewer. Appl Soft Comput J 12:1227-30.

Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G, 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 33:1123-31.

Bouzid M, Kintz E, Hunter PR, 2018. Risk factors for Cryptosporidium infection in low and middle income countries: A systematic review and meta-analysis. PLoS Negl Trop Dis 12:e0006553.

Brindha J, Balamurali MM, Chanda K, 2021. An overview on the therapeutics of neglected infectious diseases—Leishmaniasis and Chagas diseases. Front Chem 9:622286.

Brown ME, Lary DJ, Vrieling A, Stathakis D, Mussa H, 2008. Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. Int J Remote Sens 29:7141-58.

Candela E, Goizueta C, Sandon L, Muñoz-Antoli C, Periago MV. The relationship between soil-transmitted helminth infections and environmental factors in Puerto Iguazú, Argentina: cross-sectional study. JMIR Public Health Surveill 2023;9:e41568.

Celemín JP, Mikkelsen C. Velázquez G, 2015. La calidad de vida desde una perspectiva geográfica: Integración de indicadores objetivos y subjetivos. Rev Univ Geogr 24:63-84.

Celemín JP, Velázquez GA, 2017. Spatial analysis of the relationship between a life quality index, hdi and poverty in the province of Buenos Aires and the autonomous city of Buenos Aires, Argentina. Soc Indic Res 134: 1-21.

Cociancic P, 2019. Evaluación del riesgo de infecciones parasitarias intestinales en poblaciones infanto-juveniles de Argentina: el impacto de los factores ambientales y socio-económicos en su distribución geográfica. Available from: https://ri.conicet.gov.ar/handle/11336/83720. [Thesis in Spanish].

Cociancic P, Torrusio SE, Garraza M, Zonta ML, Navone GT, 2021. Intestinal parasites in child and youth populations of Argentina: Environmental factors determining geographic distribution. Rev Argent Microbiol 53:225-32.

Cociancic P, Torrusio SE, Zonta ML, Navone GT, 2019. Sistemas de información geográfica (SIG) y sensores remotos aplicados a la epidemiología de las parasitosis intestinales en Argentina. Rev Argent Parasitol 44. [Material in Spanish].

Collins GS, Reitsma JB, Altman DG, Moons KGM, 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350:g7594.

Cuenca-León K, Sarmiento-Ordóñez J, Blandín-Lituma P, Pacheco-Quito EM, 2021. Prevalencia de parasitosis intestinal en la población infantil de una zona rural del Ecuador. Bol Malariol Salud Ambient 61:596-602. [Article in Spanish].

Del Popolo F, Jaspers D, Cepal N, 2014. Los pueblos indígenas en América Latina. Avances en el último decenio y retos pendientes para la garantía de sus derechos. CEPAL. Available from: https://www.sidalc.net/search/Record/dig-cepal-11362-37050/Description

De Bourmont S, Olmedo S, Rodríguez P, Valeggia C, 2020. Therapeutic itineraries of Qom mothers in a peri-urban community of Formosa. Arch Argent Pediatr 118:240-4. [Article in English, Spanish].

Diez Roux AV, 2015. Health in cities: is a systems approach needed?. Cad Saude Publica 31:9-13.

Dueñas AS, Gobel ND, Mota IFM, 2021. Aspectos relevantes de las enfermedades infecciosas desatendidas. Panor Cuba Salud 16:127-34. [Aricle in Spanish].

Echagüe G, Sosa L, Díaz V, Ruiz I, Rivas L, Granado D, Funes P, Zenteno J, Pistilli N, Ramírez M, 2015. Enteroparasitosis en niños bajo 5 años de edad, indígenas y no indígenas, de comunidades rurales del Paraguay. Rev Chil Infectol 32:649-657. [Article in Spanish].

Engels D, Zhou XN, 2020. Neglected tropical diseases: an effective global response to local poverty-related disease priorities. Infect Dis Poverty 9:10.

Esri, 2023a. Spatial autocorrelation (global Moran's I) (spatial statistics). Available from: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/spatial-autocorrelation.htm.

Esri, 2023b. Optimized outlier analysis (spatial statistics). Available from: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/optimizedoutlieranalysis.htm.

Esri, 2023c. Forest-based classification and regression. Available from: https://pro.arcgis.com/search/?q=regression&p=1&language=en&product=arcgis-pro&version=pro3.2&n=15&collection=help.

Esteban Mendoza MV, Arcila Quiceno VH, Morchón García R, 2020. Determinación de la seroprevalencia de Dirofilaria immitis en humanos del Área Metropolitana de Bucaramanga. Available from: http://hdl.handle.net/20.500.12494/18004. [Material in Spanish].

Gabrie JA, Rueda MM, Canales M, Sánchez A, 2012. Utilidad del método Kato-Katz para diagnóstico de Uncinariasis: experiencia en una zona rural de Honduras, 2011. Rev Med Hondur 80:3. [Article in Spanish].

Gamboa MI, Giambelluca LA, Navone GT, 2014. Distribución espacial de las parasitosis intestinales en la ciudad de La Plata, Argentina. Medicina (B Aires) 74:363-70. [Article in Spanish].

Gebreyes WA, Dupouy-Camet J, Newport MJ, Oliveira CJ, Schlesinger LS, Saif YM, King LJ, 2014. The global one health paradigm: challenges and opportunities for tackling infectious diseases at the human, animal, and environment interface in low-resource settings. PLoS Negl Trop Dis 8:e3257.

Han BA, Schmidt JP, Bowden SE, Drake JM, 2015. Rodent reservoirs of future zoonotic diseases. Proc Natl Acad Sci U S A 112:7039-44.

Hoyos CL, Cajal SP, Juárez M, Acosta NR, Krolewiecki AJ, Torrejón I, Gil JF, 2011. Clustering temporal de incidencia de la Leishmaniasis Tegumentaria Americana en el año 2009 y potencial exposición a Leishmania spp. en personas sin manifestaciones clínicas en la Localidad de Hipólito Yrigoyen. Universidad Nacional de Salta. Available from: http://eprints.natura.unsa.edu.ar/id/eprint/54. [Material in Spanish].

Iomini PA, Parodi JB, Farina JM, Saldarriaga C, Liblik K, Mendoza I, Sosa Liprandi A, Martínez-Sellés M, Burgos LM, Baranchuk A, 2021. Enfermedades tropicales desatendidas y su impacto sobre la salud cardiovascular (The NET-heart project). Medicina (B. Aires) 81:808-16. [Article in Spanish].

Juárez MM, Rajala VB, 2013. Parasitosis intestinales en Argentina: principales agentes causales encontrados en la población y en el ambiente. Rev Argent Microbiol 45:191-204. [Article in Spanish].

Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory Methods 26:1481-96.

Lary DJ, Alavi AH, Gandomi AH, Walker AL, 2016. Machine learning in geosciences and remote sensing. Geosci Front 7:3-10.

Longhi F, Cordero ML, Cesani MF, 2022. Malnutrición infantil en Río Chico (Tucumán, Argentina). Evolución y manifestaciones actuales en el contexto de la transición nutricional. Rev Univ Geogr 31:1-3. [Article in Spanish].

Lundberg SM, Lee S, 2017. A unified approach to interpreting model predictions. Available from: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Menghi CI, Iuvaro FR, Dellacasa MA, Gatta CL, 2007. Investigación de parásitos intestinales en una comunidad aborigen de la provincia de Salta. Medicina (B. Aires) 67:705-8. [Article in Spanish].

Ministerio de Salud y Ambiente de la Nación, 2004. Atención Primaria de la Salud. Boletín PROAPS-Available from: https://bancos.salud.gob.ar/sites/default/files/2020-06/boletin-remediar-14.pdf

Müller I, Gall S, Beyleveld L, Gerber M, Pühse U, Du Randt R, Utzinger J, 2017. Shrinking risk profiles after deworming of children in Port Elizabeth, South Africa, with special reference to Ascaris lumbricoides and Trichuris trichiura. Geospat Health 12:601.

Rivero MR, De Angelo C, Feliziani C, Liang S, Tiranti K, Salas MM, Salomon OD, 2022. Enterobiasis and its risk factors in urban, rural and indigenous children of subtropical Argentina. Parasitology 149:396-406.

Romero-Ramírez SC, 2022. Caracterización epidemiológica de la parasitosis intestinal. Rev Arbitr Interdiscip Cienc Salud Salud Vida 6:35-43. [Article in Spanish].

Roski J, Bo-Linn GW, Andrews TA, 2014. Creating value in health care through big data: opportunities and policy implications. Health Aff 33:1115-22.

Scavuzzo CM, Scavuzzo JM, Campero MN, Anegagrie M, Aramendia AA, Benito A, Periago V, 2022. Feature importance: opening a soil-transmitted helminth machine learning model via SHAP. Infect Dis Model 7:262-76.

Stelling J, Yih WK, Galas M, Kulldorff M, Pichel M, Terragno R, Platt R, 2010. Automated use of WHONET and SaTScan to detect outbreaks of Shigella spp. using antimicrobial resistance phenotypes. Epidemiol Infect 138:873-83.

Tapia-Veloz E, Gozalbo M, Guillén M, Dashti A, Bailo B, Köster PC, Santín M, Carmena D, Trelis M, 2023. Prevalence and associated risk factors of intestinal parasites among schoolchildren in Ecuador, with emphasis on the molecular diversity of Giardia duodenalis, Blastocystis sp. and Enterocytozoon bieneusi. PLoS Negl Trop Dis 17:e0011339.

Taranto NJ, Cajal SP, De Marzi MC, Fernandez MM, Frank FM, Bru AM, Minvielle MC, Basualdo JA, Malchiodi EL, 2003. Clinical status and parasitic infection in a Wichi Aboriginal community in Salta, Argentina. Trans R Soc Trop Med Hyg 97:554-8.

Weatherhead EC, Reinsel GC, Tiao GC, Meng X-L, Choi D, Cheang W-K, Keller T, DeLuisi J, Wuebbles DJ, Kerr JB, Miller AJ, Oltmans SJ, Frederick JE, 1998. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. J Geophys Res Atmos 103:17149-61.

Wetchayont P, Waiyasusri K, 2021. Using Moran's I for detection and monitoring of the Covid-19 spreading stage in Thailand during the third wave of the pandemic. Geogr Environ Sustain 14:155-67.

WHO, 2010. Working to overcome the global impact of neglected tropical diseases: first WHO report on neglected tropical diseases. Available from: https://apps.who.int/iris/handle/10665/44440. Accessed on: 20/11/2022.

WHO, 2021. Poner fin a la desatención para alcanzar los objetivos de desarrollo sostenible: hoja de ruta sobre enfermedades tropicales desatendidas 2021-2030. World Health Organization. Available from: https://www.who.int/es/publications/i/item/9789240010352. Accessed on: 20/11/2022. [Material in Spanish].

WHO, 2023. Soil-transmitted helminth infections. Available from: https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections. Accessed on: 20/11/2022.

Wiens J, Shenoy ES, 2018. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 66:149-53.

Zonta ML, Navone GT, Oyhenart EE, 2007. Parasitosis intestinales en niños de edad preescolar y escolar: situación actual en poblaciones urbanas, periurbanas y rurales en Brandsen, Buenos Aires, Argentina. Parasitol Latinoam 62:54-60. [Article in Spanish].

Zonta ML, Cociancic P, Oyhenart EE, Navone GT, 2020. Intestinal parasitosis, undernutrition and socio-environmental factors in schoolchildren from Clorinda Formosa, Argentina. Rev Salud Pública 21:224-31.