Detección de Zonas de Alta Prevalencia Potencial de Chagas en Argentina

Antonio Vazquez Brust[1], Tomás Olego[1], Germán Rosati[1,2], Carolina Lang[1], Guillermo Bozzoli[1], Diego Weinberg[3], Roberto Chuit[3], Martin A. Minnoni[4], and Carlos Sarraute[*4]

[1]Fundación Bunge y Born
[2]University of San Martín
[3]Fundación Mundo Sano
[4]Grandata Labs

22 de noviembre, 2018

Resumen

Se presenta en este documento un Mapa de Prevalencia Potencial de la Enfermedad de Chagas (ECh) de alta desagregación espacial. El mismo tiene como objetivo la detección de zonas fuera de la ecoregión del Gran Chaco (hiperendémica para la ECh), caracterizadas por una alta afinidad con ésta última y por una elevada vulnerabilidad sanitaria. Se desarrollan las fuentes de datos, las técnicas de procesamiento y análisis de la información utilizadas para la construcción del mapa y los subproductos parciales.

[*]Agradecemos a todos los colegas que nos han acompañado a lo largo de este trabajo. A todo el equipo de la Fundación Bunge y Born, especialmente a Brenda Walter quien desempeñó un rol fundamental como coordinadora del proyecto y como miembro del equipo de investigación, así como también a Alejandra Candia y Gerardo della Paolera por sus valiosos aportes. También recibimos observaciones y comentarios muy útiles por parte de Marcelo Abril de Fundación Mundo Sano. El equipo de Grandata Labs aportó los Call Detail Records (CDRs) así como su conocimiento y experiencia en el procesamiento de datos. Agradecemos el generoso apoyo financiero de la Fundación Bunge y Born. Todos los hallazgos, interpretaciones y opiniones expresadas en esta publicación son de los autores y no representan necesariamente el punto de vista de sus respectivas instituciones.

1. Introducción

Este documento expone los criterios de procesamiento y las técnicas de análisis utilizadas para desarrollar un mapa ampliado con alta resolución espacial, que permita identificar zonas de la República Argentina con mayor probabilidad de estar habitadas por población potencialmente afectada por la Enfermedad de Chagas (ECh). En ausencia de información desagregada respecto a la prevalencia de la enfermedad en cada localidad, y en beneficio de las políticas públicas destinadas a tratarla, éste instrumento permitirá aproximar áreas geográficas que necesitan un seguimiento focalizado.

En ese sentido, la dimensión de análisis fundamental de este trabajo será la *prevalencia potencial de Chagas*, la cual será operacionalizada mediante un *Índice de Prevalencia Potencial de Chagas (IPPECh)*. El mismo queda definido como el indicador *proxy* para determinar una población en particular, en un espacio o área definido caracterizados por una probabilidad elevada de estar afectada por el Trypanosoma cruzi. Se trata de un indicador espacial que cuantifica el grado de contacto entre la población local y la residente en la ecoregión del Gran Chaco y el tipo de acceso a servicios de salud pública.

El índice combina un conjunto de características de la población, tales como su actividad en comunicaciones móviles, y una serie de variables e indicadores vinculados a la cobertura de salud en las zonas de residencia. Como será explicado luego, existen fundamentos para considerar un índice de esta clase como alternativa promisoria en países que no poseen registros epidemiológicos completos.

La primer fuente de datos utilizada fueron los Call Detail Records -CDRs-, provistos por una compañía de telecomunicaciones, que contienen información acerca de los llamados entrantes y salientes de los usuarios[1]. Este registro permitió construir un Índice de Afinidad, que, mediante el análisis de la actividad de comunicaciones móviles, determina el nivel de vinculación de la población residente en el territorio argentino, particularmente en áreas endémicas[2] (es decir, donde se observa una alta prevalencia de la ECh por cercanía y contacto con el vector natural, *Triatoma infestans*) con aquella que vive por fuera de estas áreas.

Por otro lado, la cercanía a efectores de salud y el nivel socioeconómico de la población como determinantes de la salud fueron utilizados para la construcción de un Índice de Vulnerabilidad Sanitaria, que complementó el análisis del Índice de Afinidad. Estos determinantes se vinculan en gran medida con el estado de salud de una persona, entendida en sentido amplio: salud biológica, psicológica y social. Si se hallan ausentes o en magnitud insuficiente, se produce un estado de vulnerabilidad [8].

Al combinar ambos índices, se generó un insumo que busca identificar zonas caracterizadas por un alto contacto con áreas donde la ECh es endémica y donde se registra un alto nivel de vulnerabilidad sanitaria, es decir, áreas que no alcanzan un umbral mínimo en el acceso a servicios de salud pública. Complementar la información de afinidad con la información de vulnerabilidad sanitaria permite interpretar la relación entre ambas y entender en forma global el fenómeno de la infección y la situación socioeconómica local.

[1] Cabe destacar que los datos de la identidad de las personas fueron preservados dado que no fueron accesibles en el desarrollo del trabajo.

[2] La definición de área endémica fue determinada a partir de la información histórica del Programa Nacional de Control de Chagas, publicaciones científicas y reportes de los Ministerios de Salud Provinciales.

La información presentada posee un nivel de desagregación elevado. La unidad mínima de análisis es el radio censal, la unidad estadística más pequeña de la que se dispone información sociodemográfica pública. En la Argentina, cada unidad político-administrativa se desagrega en fracciones y cada una de ellas se desagrega a su vez en radios. El tamaño de los radios en áreas urbanas se determina según la cantidad de viviendas: el radio tiene un promedio de 300 viviendas. Para bordes de localidades el radio urbano puede bajar a 200 viviendas, aproximadamente, y en localidades aisladas a 100 viviendas. En zonas rurales, los radios se determinan por la conjunción de distintos factores: características del terreno, accesibilidad y distancia entre las viviendas [23]. En zonas urbanas un radio censal puede equivaler a una cuadra. Sólo en zonas rurales (en las que la densidad poblacional es menor) los radios censales pueden adquirir dimensiones considerables, llegando a varios km^2. En Argentina, según datos del Censo 2010, existen algo más de 52.000 radios censales distribuídos en todo el país.

El mapa, al señalar estas zonas "calientes" de alta afinidad con la zona endémica y de alta vulnerabilidad sanitaria, fue conceptualizado y desarrollado como insumo para diferentes usuarios: investigadores, especialistas en salud pública, decisores de políticas públicas en salud y funcionarios públicos, entre otros actores. En particular para los decisores, este mapa facilitaría el desarrollo de estrategias costo-efectivas para mejorar el acceso al diagnóstico y tratamiento de la ECh.

El documento se estructura de la siguiente manera:

- En la Sección 2 se reseñan algunos antecedentes relevantes sobre la temática y se marcan los principales avances respecto a los trabajos de [10, 11, 31]
- En la Sección 3 se presenta la metodología utilizada para la construcción del Índice de Afinidad, junto con el procesamiento de la información vinculada a los CDRs .
- En la Sección 4 se presentan algunos fundamentos acerca del concepto de vulnerabilidad sanitaria y una explicación detallada de los diferentes procedimientos y técnicas aplicadas para la construcción del Índice de Vulnerabilidad Sanitaria.
- En la Sección 6, se desarrolla y justifica la forma funcional y el cálculo del *IPPCh* y en la Seccion 7 se presenta el mapa final y se realiza un primer análisis descriptivo de los resultados.
- Por último, se discuten las limitaciones y futuras líneas de investigación abiertas por el trabajo en la Sección 8.

2. Antecedentes

La continua aparición de grandes bases de datos, sumada a la creciente capacidad de procesamiento informático –en conjunto, el fenómeno llamado Big Data– ha producido grandes avances en distintas disciplinas, entre ellas la epidemiología. Por medio de miles de millones de registros de llamadas celulares y bajo ciertas condiciones, hoy es posible modelar el patrón de difusión de ciertas endemias ([14]).

Las bases de datos de registros de llamados han emergido como una fuente de particular interés para el estudio de migraciones humanas ([6]). Una revisión de literatura científica reciente evidencia un creciente corpus de trabajos empleando análisis de registros de llamadas para el estudio de patrones de movilidad a distintas escalas. El uso de esta nueva fuente en

estudios de movilidad ofrece notables ventajas respecto de opciones tradicionales como las encuestas origen-destino.

Los registros de llamadas ofrecen una cobertura geográfica y poblacional óptimas ([35]), y con un costo de adquisición en potencia mucho menor a los significativos recursos necesarios para llevar a cabo encuestas de movilidad a gran escala. Como se mostró en [26], el análisis de millones de registros de llamados móviles permite inferir patrones de movilidad agregados a distintas escalas geográficas, desagregando ciertas características por grupo demográfico. Por otro lado, la naturaleza aún novedosa de esta fuente implica que su aprovechamiento como recurso de investigación se encuentra en su período inicial y exploratorio. Siendo datos recolectados con fines distintos al uso para investigación científica, es esperable la presencia de sesgos, ruido y omisiones no documentadas. Particularmente, los primeros (sesgos) serán función del nivel de cobertura de telefonía celular en la zona analizada y la penetración de mercado que cada una las diferentes compañías de telefonía móvil tenga en dichas zonas. Además -en contraste con las encuestas de movilidad tradicionales- los registros de llamadas son en extremo detallados en sus atributos espaciales y temporales, pero mucho más superficiales en los sociodemográficos ([35]).

Para el presente trabajo se dispuso de registros de 50 millones de llamadas diarias entre celulares, a lo largo de varios meses. Distintos patrones de llamadas pueden ser inferidos, estableciéndose su relación con fenómenos sociales como las migraciones estacionales o de largo plazo. Más aún, de existir estadísticas a nivel detallado (esto es, al máximo nivel de desagregación posible), se podría inferir, bajo ciertos supuestos, los patrones migratorios de individuos infectados por parásitos, virus o bacterias.

En línea con esto, en 2015, la Fundación Mundo Sano y Grandata llevaron a cabo el análisis de 18.000 millones de comunicaciones para detectar posibles zonas de riesgo de ECh en dos países de América Latina (Argentina y México). El trabajo fue presentado en el conference paper "Uncovering the Spread of an Infectious Disease with Mobile Phone Data", expuesto en el 2° Simposio Argentino de Grandes Datos, AGRANDA 2016, y en "Analyzing the Spread of Chagas Disease with Mobile Phone Data", en ocasión de IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016. Este trabajo mostró que los registros de llamadas geolocalizados son ricos en información social e individual y que se pueden utilizar para inferir si un individuo ha vivido en un área endémica en algún momento de su vida.

En el presente estudio se avanzó en dos direcciones fundamentales respecto a los trabajos mencionados:

- Se complejizó el modelo de afinidad entre zonas endémicas y no endémicas utilizado previamente. En lugar de asumir la uniformidad de las probabilidades de transmisión de Chagas en la zona endémica, se utilizó información complementaria acerca de las viviendas radicadas en la zona endémica para poder identificar diferenciales en dichas probabilidades de transmisión. Este punto se aborda en la sección 3.
- Se incorporó una dimensión más general que afecta las probabilidades de contagio de la enfermedad: la vulnerabilidad sanitaria, desarrollada en la sección 4.

La construcción de un Mapa de Prevalencia Potencial del Chagas, del cual pueda desprenderse una interpretación en términos de la prevalencia (probabilidad de contagio) para cada coordenada del territorio argentino, es una tarea difícil de sortear: entre otras cosas, la escasa

disponibilidad de información acerca de la cantidad de casos positivos de la enfermedad -o al menos de indicadores aproximados-, dificulta la construcción de una función de densidad para las probabilidades de infección por *Trypanosoma cruzi*.

En esta oportunidad se refinarán las correlaciones encontradas entre comunicaciones, movilidad, acceso al sistema de salud, características demográficas y la distribución de la ECh; además de escalar el proyecto a nivel nacional.

3. Índice de Afinidad

Como se ha mencionado previamente, el Índice de Afinidad cuantifica el grado de vinculación entre áreas endémicas de la ECh en Argentina (también conocida como el Gran Chaco Argentino) y el resto del país. Este índice integra dos dimensiones diferenciadas que, como explicaremos en las siguientes secciones, utilizan diferentes fuentes de datos: registros telefónicos, condiciones de la vivienda, y localización de efectores de salud entre otros.

La primera está vinculada a la afinidad entre áreas endémicas y no endémicas, es decir, en qué medida las diferentes áreas del país se encuentran vinculadas a las áreas endémicas. Dado que fuera de las áreas endémicas no existen condiciones propicias para la reproducción del principal vector de la enfermedad (la especie *Triatoma infestans* de vinchuca), uno de los principales vehículos de propagación de la enfermedad se vincula a las corrientes migratorias. Este hecho es aún más relevante teniendo en cuenta que buena parte de las provincias que forman el Gran Chaco han sido productoras históricas de fuerza de trabajo para la economía argentina, cuanto menos entre principios y fines del siglo XX ([5]). Por ello, la posibilidad de detectar zonas con alta prevalencia potencial de ECh requiere una correcta detección y mapeo de los flujos migratorios entre el Gran Chaco y el resto del país.

Un primer obstáculo en este sentido reside en la falta de información demográfica con un nivel de desagregación adecuado para los objetivos planteados. En efecto, la fuente clásica (y virtualmente la única del Sistema Estadístico Nacional con cobertura nacional rural y urbana[3]) para el estudio de la migración interna son los Censos Nacionales de Población. Sin embargo, los mismos, en general, se limitan a cuantificar los flujos migratorios internos a nivel provincial. Si bien se indaga por los lugares de nacimiento, residencia habitual y residencia previa (cinco años atrás), los datos se publican a nivel provincial[4].

Existen otros métodos que pueden ser utilizados para la estimación indirecta de flujos migratorios con un nivel de desagregación menor al provincial. El método de las estadísticas vitales parte de la ecuación compensadora de la población:

$$P_t = P_{t-1} + N_{t-1,t} - D_{t-1,t} + I_{t-1,t} - E_{t-1,t} \tag{1}$$

Donde se asume que la población en un período dado (P_t) es el resultado de sumar a la población en el período anterior (P_{t-1}) los nacimientos ($N_{t-1,t}$) y las inmigraciones ($I_{t-1,t}$)

[3]Los relevamientos muestrales tales como la Encuesta Permanente de Hogares pueden ser utilizado para estos fines pero tiene cobertura restringida a los grandes aglomerados urbanos.

[4]El Censo Nacional de Población y Vivienda (CNPyV) 2010 constituyó una excepción: se publicaron las respuestas a estas preguntas a nivel municipal. No obstante, dado que fueron formuladas en el *Cuestionario Ampliado* la información no se encuentra disponible a nivel de radio censal

producidos entre ambos momentos y restar las defunciones ($D_{t-1,t}$) y emigraciones ($E_{t-1,t}$). En este esquema, el término $N_{t-1,t} - D_{t-1,t}$ constituye el crecimiento vegetativo de la población y el término $I_{t-1,t} - E_{t-1,t}$ el saldo migratorio. De esta forma, es posible obtener por "residuos"dicho saldo despejando

$$I_{t-1,t} - E_{t-1,t} = (P_t - P_{t-1}) - N_{t-1,t} + D_{t-1,t} \tag{2}$$

Este método tiene como ventaja su simplicidad: solamente requiere conocer la población al inicio y al final del período y la cantidad de defunciones y nacimientos del período. Pero el método es función de la calidad de los datos censales y de las estadísticas vitales[5]. A su vez, en el caso argentino, solamente es posible calcular dichos saldos a nivel departamental, dado que ese es el máximo nivel de desagregación disponible que poseen los registros de Estadísticas Vitales.

Estas limitaciones traen aparejada la necesidad de explorar otras fuentes de información para cuantificar los flujos migratorios. Como se ha mencionado, la cuantificación de la afinidad entre ambos tipos de zonas se basa en el supuesto de que una alta conexión de comunicaciones móviles (debidamente filtrados y procesados, como se verá a continuación) constituye un indicador aproximado válido de la existencia de flujos migratorios entre ambas áreas. Un alto grado de afinidad entre un área endémica y un área externa a ella podría ser considerado como un indicador indirecto de la presencia de población migrante de área endémica que vive al momento del análisis en un área no endémica.

De este modo, cuantificar con alta precisión espacial el grado de afinidad de cada región permitiría identificar aquellas zonas donde reside una población con un mayores chances de haber contraído Chagas, por la dinámica de transmisión en Argentina, y de adquirir la enfermedad[7]. Así, la infección se habría producido en un período anterior, cuando dicha población habitaba en área endémica, o por vía transplacentaria, en cuanto aquellas zonas con población que proviene de zona endémica tendrán también mayor probabilidad de tener mujeres en edad fértil infectadas, que pueden transmitirle la enfermedad a su progenie.

La segunda dimensión constituye una ampliación respecto al trabajo original [10, 11, 31]. Aún cuando dichos trabajos consideraban a la zona endémica como un área homogénea, es razonable pensar que dentro de la misma las probabilidades de transmisión de la enfermedad no se encuentran distribuidas de manera uniforme. En efecto, el Gran Chaco no es una zona de transmisión vectorial homogénea por su condición geográfica, historia de acciones de control y condiciones sociodemográficas y edilicias de sus pobladores. Si bien no existen registros completos y confiables, la transmisión (o la incidencia anual) se ve facilitada en parte por una pobre infraestructura de las viviendas y el peridomicilio, que facilitan la transmisión en ciertas zonas dentro del área endémica[9].

Por este motivo, se buscó identificar la afinidad entre el hábitat humano y el vector de la enfermedad (vinchuca): es decir, si una vivienda presenta ciertas características (materiales precarios en techo y paredes), presentará una mayor probabilidad de alojar al vector, y por ende, una mayor probabilidad de transmisión de la enfermedad. En términos generales, si

[5] "La eficacia de este "método residual"depende de la calidad de las estadísticas de nacimientos y de defunciones, sin olvidar que las cifras censales también suelen estar afectadas por errores y que éstos pueden ser importantes. Tal exigencia constituye un serio obstáculo en la mayoría de los países latinoamericanos, cuyas estadísticas de nacimientos y defunciones, al nivel regional, ofrecen evidencias de errores de integridad difíciles de evaluar"[13, p. 151]

una vivienda presenta ciertas características - como materiales de baja calidad en techo y paredes- presentará una mayor probabilidad de alojar al vector y, por ende, una mayor probabilidad de transmisión de la enfermedad. A tal efecto se construyó una métrica para cuantificar en cada radio censal la cantidad de viviendas que son "afines.ª la vinchuca, es decir, que presentan condiciones habitacionales y de materiales que son propicias para que la vinchuca se asiente y forme colonia.

3.1. Condiciones de la vivienda

Dado que es posible flexibilizar el supuesto acerca de las chances de transmisión uniformes en toda la ecoregión del Gran Chaco, y considerando el componente social de la enfermedad, las condiciones de habitabilidad de las viviendas y su relación directa con las condiciones propicias para el alojamiento del vector, se propuso como objetivo desagregar la región en múltiples sub-áreas.

Figura 1: Límites de la ecoregión del Gran Chaco

(a) Fuente: elaboración propia sobre datos proporcionados por la Fundación Mundo Sano

Para ello, se utilizó información proveniente del CNPyV del año 2010[22] y se construyó un índice que logra cuantificar para cada vivienda sus condiciones de viabilidad para el alojamiento de la vinchuca.

Las viviendas tipo rancho cuyos materiales de construcción incluyen paredes de adobe sin revoque y techos de caña, barro o paja favorecen la domiciliación del vector:

"La caña, tabla o paja con barro, paja sola son elementos que suelen ser buenos aislantes térmicos, pero no son aislantes hidrófugos. Se usan en lugares secos y se construyen con mucha pendiente a fin de acelerar el escurrimiento del agua. Asimismo, tienen peligro de incendio, no son higiénicos, e inclusive favorecen a la propagación del Mal de Chagas [sic] a través de la vinchuca". [21]

Se trabajó con las siguientes variables:

- Material predominante de los pisos
  - Cerámica,baldosa,mosaico,mármol,madera,alfombrado
  - Cemento o ladrillo fijo
  - Tierra o ladrillo suelto
  - Otro
- Material predominante de la cubierta exterior del techo
  - Cubierta asfáltica o membrana
  - Baldosa o losa (sin cubierta)
  - Pizarra o teja
  - Chapa de metal (sin cubierta)
  - Chapa fibrocemento o plástico
  - Chapa de cartón
  - Caña, palma, tabla o paja con o sin barro
  - Otro
- Revestimiento interno o cielorraso del techo
  - Si
  - No

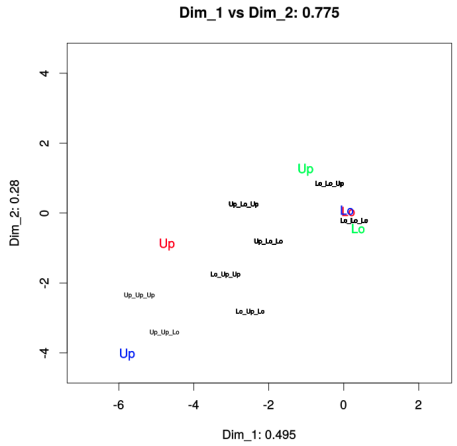
Con el propósito de analizar las condiciones de habitabilidad favorables para la vinchuca, se contruyeron los siguientes factores-variables que indican la presencia y ausencia de tales condiciones, es decir, se dicomotizaron las variables anteriores de la siguiente forma:

	Positivo (1)	Negativo (0)
Material predominante pisos	Tierra, ladrillo, suelto	Resto
Material predominante techos	Caña, palma, tabla, paja con o sin barro	Resto
Revestimiento interno o cielo raso	No	Si

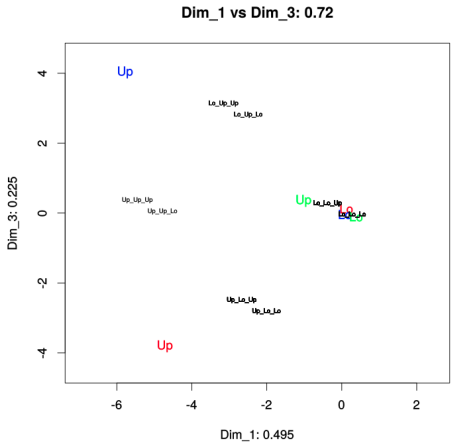
Cuadro 1: Variables utilizadas para analizar las condiciones de habitabilidad que favorecen la domiciliación de *Triatoma infestans*

En tanto se trata de variables de nivel de medición discretas y categóricas, se utilizó el Análisis de Correspondencias Múltiples como técnica de reducción de dimensionalidad para recuperar las variables latentes; luego se obtuvo la dimensión de mayor variabilidad y se recuperaron las coordenadas para dicha variable latente.

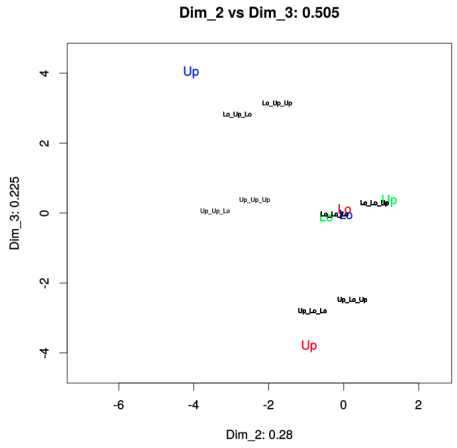
A continuación se grafican los resultados para las variables:



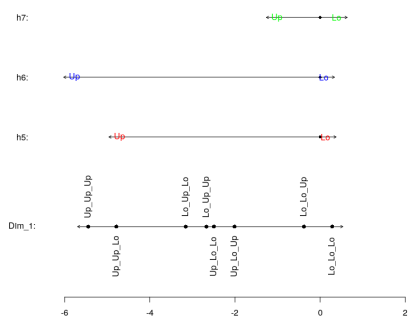
(a) MDS: Dim_1 vs Dim_2



(b) MDS: Dim_1 vs Dim_3



(c) MDS: Dim_2 vs Dim_3



(d) MDS: Dim_1 0.495

Las referencias para interpretar los gráficos es la siguiente:

- Rojo: H5 - Material predominante pisos-,
- Azul: H6 - Material predominante techos-,
- Verde: H7 - Revestimiento interno o cielo raso,

Se codificaron los 1 como 'Up' y los 0 como 'Lo', de manera que Up_Up_Up corresponde a H5=1, H6=1, H7=1.

A partir del gráfico D, la primera dimensión explica el crecimiento de los factores (presencia), de modo que a medida que disminuyen los valores de la dimensión, aumenta el nivel de precariedad de la vivienda y viceversa. Se observa que los valores máximos y mínimos corresponden con los valores mínimos y máximos de precariedad. Cabe destacar que dicha dimensión corresponde al 50 % de la variabilidad de los datos si se analiza el peso de cada

autovalor.

3.2. Índices a nivel antena

La información del polígono del área endémica posee un nivel de resolución asociada al polígono del radio censal. Ahora bien, la información para generar el grafo de contactos (ver sección 3.3) está asociada a las antenas, es decir, puntos fijos en el espacio. Por ello, el primer paso para la generación de los índices a nivel de antena consistió en estimar la superficie de cobertura de las antenas de la Argentina. Para ello se recurrió a un diagrama de Voronoi. El diagrama resulta de la subdivisión de un plano euclidiano de acuerdo a la ubicación de un conjunto de sitios, donde a cada sitio le corresponde una y sólo una región o celda. Uno de los atributos del diagrama de Voronoi es cada celda abarca los puntos del plano que se encuentran más cercanos a su sitio que a cualquier otro ([1]). La metodología es utilizada para el análisis de comunicaciones móviles cuando no se dispone de datos oficiales acerca de las zonas de servicio; al calcular la Celda de Voronoi correspondiente a cada antena se obtiene una aproximación a su área de cobertura ([33]).

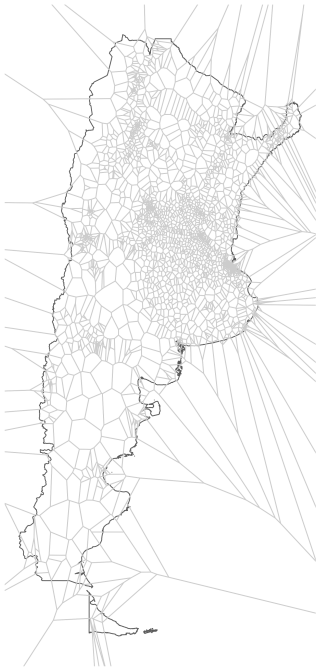


Figura 4: Celdas de voronoi generadas en base a la posición de las antenas analizadas

Los resultados del índice de calidad de la vivienda fueron agregados por cada celda de Voronoi, asignando una cantidad de hogares en función del porcentaje del área que intersecaba cada radio censal. En otras palabras, cada radio aportó un porcentaje de sus valores en función de la proporción de superficie compartía con el o lós polígonos de Voronoi en intersección.

De tal manera, para cada antena j se obtuvo un índice conformado por la suma

$$vinchucalmat_j = \sum_{i=1}^{R} w_{i,j} \sum_{k=1}^{8} Nh_{k,i}\ v_{k,i}$$

donde v_k es el puntaje obtenido en el análisis de correspondencias, con $k = 1:8$, $Nh_{k,i}$ es la cantidad de hogares en el radio i que obtuvieron puntaje v_k y $w_{i,j}$ la proporción de superficie compartía definida como

$$w_{i,j} = \frac{Area(radio_i \cap antena_j)}{Area(radio_i)}$$

Finalmente, se graduaron los valores obtenidos por las antenas en función de los cuantiles. De esta forma, se generaron cuatro subdivisiones que representan un cuartil respecto a un indicador que mide el grado en que los materiales de las viviendas locales favorecen la presencia del *Triatoma infestans*.

3.3. Datos telefónicos

El principal insumo utilizado para la construcción del Índice de Afinidad es la información de actividad de telefonía celular. Esta base de datos consta de una tabla de antenas geolocalizadas y una tabla de llamados (*Call Detail Records* o CDRs), donde cada llamado está asociado a una antena particular, que es la que se utilizó para efectuar la conexión.

La base de antenas geolocalizadas tiene aproximadamente 3.500 antenas con una ubicación válida[6] en los cuales se registra actividad en el periodo estudiado.

3.3.1. Call Detail Records (CDRs)

La base cuenta con aproximadamente 70.000.000 de registros por día, es decir 2.000.000.0000 por mes y 10.000.000.000 en el período disponible. Se dispone de 5 meses de datos, desde noviembre de 2011 a marzo de 2012.

Para contabilizar interacciones entre usuarios, estos fueron identificados unívocamente utilizando un "hash", es decir un código anonimizado. El "hashing", aplicado como paso previo a cualquier análisis de los datos, consiste en transformar los identificadores de usuario utilizando una función resumen que toma la secuencia de caracteres del identificador original y devuelve una nueva secuencia, denominada hash. La operación es irreversible: no es posible determinar la información de entrada -es decir, el código que identifica a un usuario- en base al hash resultante ([29]). De tal modo se resguarda la anonimidad de los usuarios registrados en la base, manteniendo a la vez la capacidad de discernir entre usuarios distintos.

Los CDRs están organizados en registros (filas), y cada uno aporta los siguientes datos (columnas).

[6] En los datos originales se encontraron antenas con datos inconsistentes, es decir, que no tenían información de localización (latitud/longitud) o bien que se geolocalizaban fuera del territorio argentino. Estos registros fueron descartados.

Indice	Tipo	Descripción
0	STRING	Hash del origen (destino si la dirección es I)
1	STRING	Hash del destino (origen si la dirección es I)
2	CHAR	O para llamados salientes, I para entrantes
5	YYYYMMDD	Día en el que ocurrió el llamado
6	HHMMSS	Hora en el que ocurrió el llamado
7	INT	Duración del llamado en segundos
8	STRING	Código de la antena usada (incluye sección de la antena)

Cuadro 2: Datos de CDRs

3.3.2. Metodología

El proceso de análisis de los CDRs y extracción del score de afinidad con áreas de Chagas endémico de cada antena puede resumirse en tres etapas secuenciales.

Detección de *antena_hogar* de cada usuario

El primer paso consistió en la detección, para cada usuario u, la *antena_hogar* H_u, con el objetivo de geolocalizar la información de las comunicaciones. Así, H_u se define como la antena en la que se registran más llamados del usuario correspondiente en el período estudiado, sobre el conjunto de llamados que se realizan los días de semana por la noche. Es decir, de lunes a jueves, entre las 8 de la noche y las 6 de la mañana del día siguiente. Se optó por ese rango asumiendo que coincide con la franja horaria durante la cual la mayor parte de las personas permanece en su residencia. De existir más de una antena con una cantidad máxima de llamados, se seleccionó en forma aleatoria una de ellas.

La cantidad de usuarios para los cuales se definió H_u ronda los 15 millones.

Cálculo de la afinidad semilla para cada antena

En este paso se determinó, para cada antena a, la *afinidad semilla* s_a. La afinidad depende de características de las viviendas en la zona donde se emplaza la antena. Por ello, depende sólo de atributos geográficos y demográficos, y no de aquellos relacionados con la comunicación telefónica.

Como entrada del proceso se utilizó una partición por cuantiles del indicador de condiciones de materiales de vivienda, cuya estimación se desarrolló en la sección 3.1. A cada antena contenida dentro de la ecoregión del Gran Chaco le corresponde un indicador de afinidad igual al cuartil de condiciones de vivienda al que pertenece (un número entero entre 1 y 4 inclusive), mientras que fuera del polígono del Gran Chaco el indicador vale 0. De esta manera, la s_a de cada antena puede tomar un número entero entre 0 y 4 como valor.

Cálculo de indicadores de afinidad para las antenas utilizando CDRs

Para cada habitante u de alguna antena del país, el objetivo fie asignarle un nivel de afinidad basado en sus comunicaciones telefónicas en el grafo social G. En particular se calculó el conjunto $V_u(G)$ de usuarios que conforma la *vecindad* de u en el grafo social. Cada nodo vecino $v \in V_u(G)$ modifica el score de u de acuerdo a la intensidad de la arista (u, v), y los indicadores demográficos semilla en la región de la antena donde vive v (es decir s_{H_v}).

Dados los indicadores modificados para cada usuario, se pueden agregar los resultados agrupando por antena hogar, para hallar una distribución del indicador de afinidad para cada una de ellas.

Primero, para cada usuario u se definió su afinidad semilla como:

$$s_u = s_{H_u}$$

es decir, se asoció a cada usuario con la afinidad semilla de la zona en la que vive.

Luego, dados u y su vecindad $V_u(G)$, se definió el indicador de afinidad s'_u de cada usuario como:

$$s'_u = max_{v \in V_u(G)}(s_v)$$

Esto quiere decir que el indicador de afinidad modificado por el grafo social se obtiene calculando cuál es la afinidad semilla para cada $v \in V_u(G)$, y luego tomando el máximo de las afinidades que se hallaron. Notemos que este número también es un entero entre 0 y 4.

Se definió en este sentido una serie de indicadores para cada antena a, que es la información de cuántos usuarios entre los habitantes de a tienen cada uno de los valores s'_u. Llamamos $H_{a,k}$ al subconjunto de los habitantes H_a de una antena a, tales que su s' vale k. Es decir:

$$H_{a,k} = \cup_{(h \in H_a / s'_h = k)}\{h\}$$

Si se reporta la cantidad de usuarios en $H_{a,k}$ para cada uno de los valores $k \in 0, \cdots, 4$, se tiene un resumen de la distribución de niveles de afinidad en cada antena a en particular. Es decir que, para cada antena, la distribución de los usuarios está dada por la tupla:

$$< a, |H_{a,0}|, |H_{a,1}|, |H_{a,2}|, |H_{a,3}|, |H_{a,4}| >$$

4. Índice de Vulnerabilidad Sanitaria

Es posible plantear que las chances de contraer la ECh, o de estar infectado por Trypanosoma cruzi, no se encuentran asociadas tan sólo a las probabilidades de contacto con el vector o con los movimientos o contactos con las zonas endémicas. Existen también determinantes de carácter más general que podrían ser englobados bajo el concepto de “Vulnerabilidad Sanitaria". Dichos determinantes se encuentran asociados al acceso a los servicios y cobertura de salud de diferentes segmentos de la población. Al desarrollar un índice que cuantifica ésta dimensión se complementa al de afinidad con áreas de Chagas endémico, ya que estimar la Vulnerabilidad Sanitaria permite priorizar "zonas calientes"fuera del área endémica, aquellas que tienen bajo acceso a servicios sanitarios.

4.1. Definición

La existencia de disparidades en el acceso a servicios de salud es un fenómeno que ha sido estudiado y documentado en múltiples ocasiones, hallando evidencia de disparidades de acceso y de cobertura médica constante para distintos estratos de población. En líneas generales, los segmentos de población más pauperizados y/o residentes en zonas aisladas presentan menores niveles de acceso a dichas prestaciones de salud ([4], [34]).

En la literatura sobre la problemática de salud suele diferenciarse esta noción de "acceso desigual" de la llamada "vulnerabilidad" que remite, en última instancia, al riesgo de desarrollar ciertas enfermedades o al hecho de estar expuesto a ciertos factores de riesgo ambientales. En este sentido, el estudio de *poblaciones vulnerables* o de *factores de vulnerabilidad* resultan de interés.

Un obstáculo para la cuantificación de "vulnerabilidad" emerge de la necesidad de considerar múltiples factores que podrían explicar las desigualdades en el acceso al sistema de salud. En [17] se relevan las principales dimensiones e indicadores vinculados a estas nociones de *vulnerabilidad* y *desigualdades en el acceso a la salud* utilizados en un conjunto de trabajos seleccionados:

Aspectos de la Vulnerabilidad	Papers	Porcentaje
Pobreza	21	91.3 %
Minoría Racial/Étnica	18	78.3 %
Enfermedades crónicas (físicas o mentales)	12	52.2 %
Falta de cobertura de salud	8	34.8 %
Edad	6	26.0 %
Encarcelamiento	3	13.0 %
Condición de migrante	3	13.0 %
Bajo nivel educativo	3	13.0 %
Residencia en áreas de baja cobertura	2	8.7 %
Desempleo	1	4.3 %
Condición de viudez	1	4.3 %
Carencia de vivienda	1	4.3 %

Cuadro 3: Indicadores vinculados al concepto de Vulnerabilidad Sanitaria. Tomados de C. Grabovschi, C. Loignon, and M. Fortin: *Mapping the concept of vulnerability related to health care disparities: a scoping review*

Puede verse en la Tabla 3 que los principales indicadores utilizados por los estudios analizados en [17] se vinculan a la condición de pobreza, a la pertenencia a minorías étnicas o raciales, a la presencia de enfermedades mentales o físicas de carácter crónico y a la falta de cobertura médica. Cabe destacar que estos indicadores están definidos a nivel de individuos. Existen también otros factores que determinan el nivel de vulnerabilidad sanitaria que están más vinculados a la dimensión ambiental o del entorno [30].

En la aproximación realizada se consideraron varios de los indicadores mencionados. La noción de vulnerabilidad sanitaria considerada se compone de los siguientes factores asociados:

- *Acceso a prestaciones y servicios de salud por parte del estado:* para esta dimensión se usó como indicador principal la cercanía a efectores de salud. Se construyó un dataset que contiene la ubicación (latitud y longitud) de la gran mayoría de los efectores de salud estatales de todo el país. Para ello, se integraron fuentes provenientes del Estado Nacional y de los Estados Provinciales. Se calculó el tiempo de caminata desde diversos puntos hasta el Centro de Salud más cercano.
- *Nivel Socioeconómico de la Población:* para construir el Índice de Nivel Socioeconómico (INSE), se utilizó información censal. Si bien este punto se detalla más adelante, puede mencionarse que el cálculo del indicador implicó el procesamiento de información censal,

correspondiente al CNPyV del año 2010, a nivel individuo. Para ello, se seleccionó una serie de variables relevantes -nivel educativo, indicadores de Necesidades Básicas Insatisfechas, etc.- y se combinaron utilizando *variational autoencoders*, un método para la reducción de dimensionalidad basado en redes neuronales (ver sección 4.4).

4.2. Metodología de construcción

Las dimensiones anteriormente mencionadas se combinaron para construir un *Mapa de Vulnerabilidad Sanitaria*. El objetivo del mapa de vulnerabilidad es identificar zonas con un déficit potencial en la cobertura sanitaria de la población, es decir, que no logran superar un umbral mínimo en el acceso a servicios de salud. Teniendo en cuenta este objetivo, se construyó una métrica que permite ordenar y clasificar a las diferentes zonas en función de este déficit potencial.

4.2.1. Fuentes de información utilizadas

Para la construcción del mapa se recabaron y analizaron las siguientes fuentes:

- Datos censales (Censo Nacional de Población, Hogares y Vivienda 2010 [22])
- Polígonos de radios censales
- Ubicación de Efectores de Salud Pública: hospitales públicos, centros de salud y postas sanitarias (en total, 16.564)
- Ejes de calles (rutas nacionales, provinciales, caminos y trazas urbanas) utilizadas para calcular por simulación distancias entre hogares y efectores de salud.

4.2.2. Criterios y técnicas utilizadas para el procesamiento de la información

Como se ha mencionado, si bien se procesaron datos a nivel de desagregación menor (como por ejemplo, datos individuales del CNPyV o datos de los efectores de salud), para la construcción del mapa final se procedió a agregar esta información a nivel de radio censal.

En los siguientes apartados se detallan los diferentes procedimientos empleados para la preparación de la información a explotar estratos de población y las diferentes técnicas de procesamiento y análisis empleadas.

4.3. Cercanía a Hospitales y Centros de Salud

4.3.1. Construcción y limpieza del dataset de Efectores de Salud

Para la construcción del indicador *Cercanía a Efectores de Salud* el primer paso fue la construcción de un dataset con registros de ubicación de la mayor cantidad posible de efectores de salud en todo el país, localizados con coordenadas de latitud y longitud.

Este dataset fue construido a partir de la integración de diferentes fuentes de datos oficiales:

- *Base nacional de Hospitales y Centros de Atención Primaria*: la misma fue compilada por el Sistema de Información Sanitaria Argentina (SISA - https://sisa.msal.gov.ar/), y publicada por SEDRONAR en el sitio de IDERA (http://catalogo.idera.gob.ar). Este dataset fue utilizado como punto de inicio y *base maestra*. El mismo fue enriquecido y corregido en base a la información obtenida de fuentes adicionales.

Otras fuentes utilizadas:

- *Efectores de salud del programa SUMAR*: La información publicada en el sitio http://programasumar.com.ar/efectores/ fue recogida via *scraping*[7] permitiendo reunir en una base consolidada los múltiples listados de efectores con sus direcciones.
- *Listados de hospitales y centros de atención de salud del Programa Nacional de Salud Sexual y Procreación Responsable (Ministerio de Salud)*: Se dispone de datos por provincia. Se descargaron los datos del sitio y se georreferenciaron.[8]

Fuentes provinciales

Se expone a continuación el listado de fuentes a nivel provincia[9] exploradas. En general, se trata de Ministerios de Salud provinciales.

	Fuente (con hipervínculo)	Tarea	Resultado
Buenos Aires	Portal datos abiertos	Descarga	Agrega casos
CABA	Portal datos abiertos	Descarga	Agrega casos
Córdoba	Portal datos abiertos	Descarga	Agrega casos
Entre Ríos	Ministerio de Salud	Revisión	No agrega casos
Formosa	Plan Nacer	Revisión	No agrega casos
Mendoza	Ministerio de Salud	Revisión	No agrega casos
Misiones	Ministerio de Salud		
Neuquén	Ministerio de Salud	Revisión	Alto esfuerzo[10]
San Luis	Ministerio de Salud	Revisión	No agrega casos
Santa Cruz	Ministerio de Salud	Revisión	Pendiente
Santa Fe	Ministerio de Salud	Scrap	Agrega casos
Santiago del Estero	Ministerio de Salud	Revisión	Pendiente
Tierra del Fuego	Ministerio de Salud - Río Grande	Revisión	No agrega casos
	Ministerio de Salud - Tolhuin	Revisión	No agrega casos
	Ministerio de Salud - Ushuaia	Revisión	No agrega casos

Cuadro 4: Listado de fuentes provinciales de efectores de salud

En todos los casos, se aplicó el siguiente procedimiento:

1. Se descargaron archivos consolidados (si los hubiera) o se recolectó el contenido del

[7] El término "scraping", o "web scraping", refiere al conjunto de técnicas y tecnologías utilzadas para extraer información de sitios web que no se encuentra debidamente formateada para su extracción

[8] El archivo de la provincia de San Juan fue publicado vacío, por lo cual no pudo ser considerado.

[9] Para el resto de las provincias no se encontraron fuentes alternativas.

[10] Datos en formato gráfico. No era trivial el esfuerzo requerido para transcribir y georeferenciar los datos.

sitio correspondiente
2. Se georreferenciaron las direcciones de los efectores de salud utilizando la API de GoogleMaps.
3. Los registros sin coordenadas fueron descartados.
4. Se "superpusieron" los datos resultantes en el mapa junto con el dataset inicial.
5. Se retuvieron los puntos que no se superponen -más allá de un *buffer* de 100 metros-, permitiendo así descartar sitios duplicados por aparecer en múltiples fuentes.

A título de ejemplo, se presentan los diferentes puntos provenientes de cada una de las fuentes para la Ciudad Autónoma de Buenos Aires (CABA).

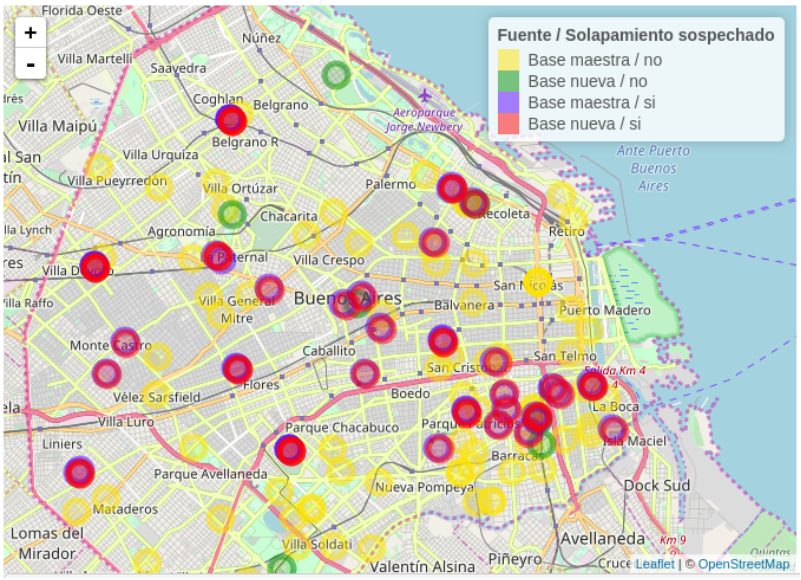


Figura 5: Comparación entre la base maestra (SISA) y las diferentes fuentes consultadas para la CABA.

Puede notarse que el dataset de CABA fue notablemente enriquecido: si bien existen bastantes puntos de superposición (rojos y azules), existe una considerable cantidad de puntos que fueron agregados a la *base maestra* (amarillos).

Al completarse el procedimiento con cada una de las fuentes de datos adicionales adquiridas, el número de efectores de salud pública georeferenciados aumentó de los 4.419 a 16.564.

4.3.2. Clasificación de los efectores de salud según nivel de complejidad

Con ello se completó la compilación del dataset de efectores de salud en todo el país. Posteriormente, y como última etapa del proceso de limpieza del dataset, los efectores de salud fueron clasificados según su nivel de complejidad. Se buscó el modo de reflejar que la cercanía a un hospital de alta complejidad implica un acceso a prestaciones de salud a priori mayor que la cercanía a una posta sanitaria o una "salita". El tipo de problemas, las emergencias atendidas y la atención que pueden brindar estos establecimientos difieren notablemente. Si bien existían en los diferentes datasets utilizados clasificaciones en función

de la noción de complejidad, los criterios variaban entre fuentes: las clasificaciones no eran homogéneas en los diferentes listados de efectores de salud consultados y consolidados en la base final.

Se realizó un trabajo de revisión con expertos de la Fundación Mundo Sano, que permitió arribar a una clasificación que unificara las diferentes denominaciones con los que se describe a los establecimientos de salud, produciendo una clasificación simple según nivel de resolución. El resultado final fue una clasificación de los efectores de salud en tres categorías, en orden decreciente de resolución de complejidad en la atención::

1. Hospital
2. Centro de Salud
3. Posta Sanitaria

Tras descartar los efectores de salud pública que no pertenecen a ninguna de las categorías definidas (por ejemplo, geriátricos u oficinas administrativas), se conservaron 15.903 registros de los 16.654 del total recopilado.

4.3.3. Cálculo del tiempo al centro de salud más cercano

El paso siguiente en el proceso fue el cálculo del tiempo necesario para llegar al efector de salud más cercano. A partir de este punto, se comenzó a agregar la información a nivel radio censal.

Se planteó hallar el efector público más cercano para cada radio censal. Este punto representó un desafío dado que los radios son polígonos y los efectores, puntos. Una primera opción fué utilizar el centroide del radio y calcular la distancia y el tiempo desde esas coordenadas[11].

Sin embargo, dado que la forma, los límites y la superficie de los radios censales son muy disímiles a lo largo del país (especialmente, en zonas rurales o poco pobladas) se decidió calcular las distancias y tiempos de la siguiente forma:

1. Dentro de cada radio se seleccionaron 5 puntos (pares de coordenadas) al azar
2. Se identificó para cada punto el efector de salud más cercano
3. Se calculó para cada punto la distancia/tiempo al efector de salud más cercano
4. Se promediaron las 5 distancias/tiempos para obtener el valor final

Este procedimiento se realizó para cada una de las categorías de efectores de salud -Hospital, Centro de Salud y Posta Sanitaria. Ahora bien, para el paso 2. se utilizó un algoritmo kNN -k vecinos cercanos-. Las distancias en metros representadas por diferencias iguales en latitud y longitud difieren incrementalmente a medida que uno se acerca a los polos y se aleja de la línea de Greenwich. No obstante, para distancias cortas y suficientemente alejadas de los extremos de la grilla de coordenadas Mercator (tal es el caso en cuestión) el error es bajo. Por esta razón y para aprovechar la eficiencia de kNN, se aceptó dicha limitación.

Para la determinación final del tiempo se utilizó Open Source Routing Machine (OSRM), un sistema de ruteo de alta performance que indica las ruta más corta a través de vías

[11] La noción de accesibilidad ha sido recientemente utilizada como un indicador *proxy* de la distribución desigual de bienes y recursos en las sociedad. Un paper reciente[36] ha mostrado que existe una correlación positiva y relevante entre los tiempos agregados de acceso a aglomerados urbanos de las diferentes regiones del mundo y muchos de los indicadores socioeconómicos habitualmente utilizados -niveles de ingreso medios, nivel educativo de la población, tasas de mortalidad infantil, etc.-

públicas entre cualquier par de coordenadas origen-destino ([20]). Para determinar las rutas OSRM utiliza grillas de calles descargadas de OpenStreetMap ([28]), un repositorio público de información geográfica cuya calidad de datos lo ha establecido como fuente frecuente para estudios de movilidad ([18], [25])

4.3.4. Tiempos a pie

El indicador calculado para la medición del acceso a cobertura de salud fue la distancia a pie hasta el efector de salud más cercano. Este indicador resulta relevante dado que existe evidencia de que, al menos para cierto tipo de tratamientos médicos, la distancia a pie a un establecimiento de salud es un buen predictor de las probabilidades de culminación de dicho tratamiento:

> Results from this study of Baltimore City clients attending outpatient drug addiction treatment programs suggest that having to travel more than 1 mile from the treatment center reduces clients' chances of completing treatment by almost 50 %, after controlling for the effects of demographic variables and type of drug problem. Moreover, living more than 4 miles away from treatment decreases the expected length of treatment by almost 13 days in comparison to clients traveling less than 1 mile. [4] [12]

En efecto, en el estudio sugiere que una distancia mayor a una milla (aproximadamente 1,6 km.) redunda en un considerable incremento de la probabilidad de no culminación de un tratamiento de rehabilitación. A su vez, distancias mayores a 6.4 km redundan en una baja de la duración media del tratamiento en casi dos semanas.

Para un número muy reducido de radios no fue posible determinar un trayecto a pie hasta un efector de salud cercano (10 en total sobre 52.406 radios analizados). Para cada uno de ellos, se realizó inspección visual de imágenes aéreas. En todos ellos se comprobó que coinciden con áreas no urbanizadas que abarcan zonas de esteros y bañados, o dedicadas a la agricultura, donde no hay vías peatonales. En esos casos la distancia a pie fue imputada, usando los valores máximos encontrados en el resto de los radios en la misma región administrativa (departamento).

A continuación se presentan mapas a nivel radio censal de los tiempos de llegada a Hospitales, Centros de Salud y Postas Sanitarias para el total del país.

[12] Resultados de este estudio de pacientes de la ciudad de Baltimore que asisten de forma ambulatoria a los programas de tratamiento para la adicción a las drogas sugieren que tener que viajar más de 1 milla desde el centro de tratamiento reduce las posibilidades de los clientes de completar tratamiento en casi el 50 %, controlando los efectos de variables demográficas y tipo de problema de drogas. Además, el hecho de vivir a más de 4 millas de distancia del tratamiento disminuye la longitud esperada de tratamiento por casi 13 días en comparación con los pacientes que viajan menos de 1 milla

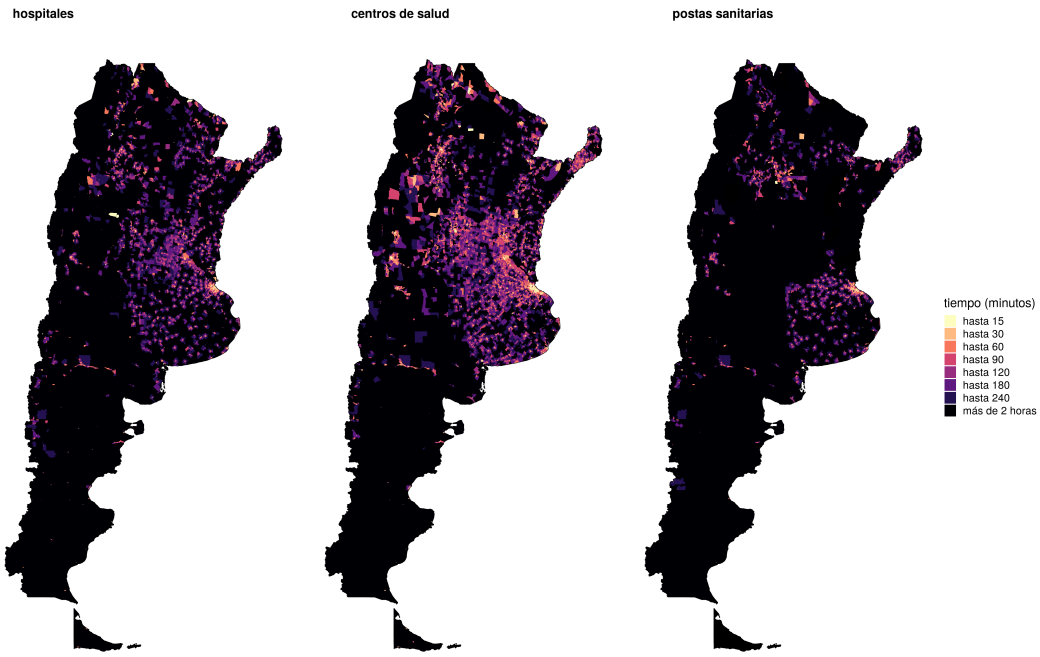


Figura 6: Tiempos de viaje a pie hasta efectores de salud agregados en promedio por radio censal.

Como puede verse, los mapas de tiempo de los dos primeros tipos de efectores de salud -Hospitales y Centros de Salud-son notablemente similares. El de Postas Sanitarias resulta más contrastante. En efecto, puede notarse una estructura mucho más segmentada: existe una gran cantidad de regiones con tiempos de viaje notablemente elevado hasta postas sanitarias. De hecho, en una primera inspección visual, pareciera que muchas de estas zonas con altos tiempos de viaje hasta Postas a su vez se caracterizan por valores inversos en tiempos a Hospitales y Centros de Salud.

Esta característica puede deberse a dos factores (no excluyentes entre sí):

- *diferentes políticas provinciales en la asignación y asignación de postas sanitarias*,
- *errores de cobertura en los datos analizados*, es decir, que halla zonas en las que la omisión de postas sanitarias sea muy elevada.

Se difiere a una etapa posterior de investigación determinar el grado de confiabilidad de los datos recopilados en relación a postas sanitarias.

La distancia final para cada radio Δ_r es la distancia mediana entre todos los puntos muestreados en cada radio, según ilustra el siguiente mapa:

$$\Delta_r = MED(\Delta_hosp_1, ..., \Delta_hosp_5, \Delta_centro_{i1}, ..., \Delta_centro_{i5}, \Delta_posta_{i1}, ..., \Delta_posta_{i5}) \quad (3)$$

Es decir incluye todos los efectores-distancias en un radio censal r.

Aún con las reservas mencionadas respecto a la fiabilidad de los datos, se incluyó la capa de Postas Sanitarias para el cálculo del indicador, dado que la toma de la distancia mediana minimiza el efecto de potenciales *outliers*.

4.4. Nivel Socioeconómico

4.4.1. Datos de entrada

Para el cálculo del *Índice de Nivel Socioeconómico* (INSE) se utilizaron datos a nivel individuo provenientes del CNPyV del año 2010. Dado que el INSE suele ser una variable medida a nivel del hogar, se optó por calcular un índice para cada jefe de hogar en el dataset del Censo.

Para la estimación de los valores del índice se utilizaron las siguientes variables (ver Anexo I), las cuales fueron ordinalizadas:

Variable	Unidad
Condición de propiedad de la vivienda	Vivienda
Calidad Materiales	Vivienda
Calidad de Conexión a Servicios Básicos	Vivienda
Calidad de Construcción	Vivienda
Hacinamiento	Hogar
Presencia de algún indicador NBI	Hogar
Nivel Educativo del total del Hogar	Hogar
Cantidad de Desocupados en el Hogar	Hogar
Existencia de servicios doméstico	Hogar
Condición de actividad	Individuo (jefe)
Nivel educativo	Individuo (jefe)

Cuadro 5: Indicadores utilizados para la construcción del INSE.

Para construir el Índice, se utilizó una codificación de termómetro para las variables ordinales. Sea N la cantidad de casos y $v_1, \ldots, v_I$ las variables. Para cada variable v_i, existen K_i categorías. Se crearon las siguientes variables codificadas $x_{k_i}^{(i)}$ para cada variable v_i y para cada categoría k_i donde $2 \leq k_i \leq K_i$. En cada caso j con $1 \leq j \leq N$ vale:

$$x_{k_i}^{(i)}(j) = \begin{cases} 0, & si\ v_i(j) < k_i \\ 1, & si\ v_i(j) \geq k_i \end{cases} \tag{4}$$

Supongamos que tenemos una variable como el INMAT[13], con cuatro categorías:

[13]Puede verse una descripción detallada de esta variable en el Anexo 1.

INMAT original	INMAT_2	INMAT_3	INMAT_4
1	0	0	0
2	1	0	0
3	1	1	0
4	1	1	1

Cuadro 6: Ejemplo de codificación termómetro

En este ejemplo, INMAT_1 sería el *baseline* o categoría de referencia, y tomaría un valor constante, por eso es redundante.

4.4.2. Construcción del Nivel Socioeconómico

Para la construcción del índice final se utilizó una técnica de reducción de dimensionalidad llamada *autoencoder* [16]. Los *autoencoders* son una arquitectura basada en redes neuronales. En líneas generales, un autoencoder tiene como objetivo encontrar una representación de los datos de input (*encoding*) generalmente con el objetivo de realizar una reducción de la dimensionalidad. En general, los autoencoders funcionan simplemente aprendiendo a replicar los inputs en los outputs. Si bien esto parece un problema trivial introduciendo diversas restricciones a la red puede hacerse esta tarea muy compleja. Por ejemplo, puede limitarse el tamaño de la representación interna o bien agregarse ruido a los inputs. Los autoencoders intentan aprender la función identidad bajo ciertas restricciones [15].

Un autoencoder está compuesto de dos partes:

- un encoder (o *recognition network*) que convierte los inputs a una representación interna, seguida
- un decoder (o *generative network*) que reconvierte la representación interna a los outputs.

Suele tener una arquitectura análoga a un Multi-Layer Perceptron, con la única excepción de que el número de neuronas en la capa de output debe ser igual a la de input.

El modelo entrenado posee como capa final una función logística y realiza *dropout* (de 0.5) en sus capas intermedias para regularizar y lograr coeficientes con buena capacidad de generalización.

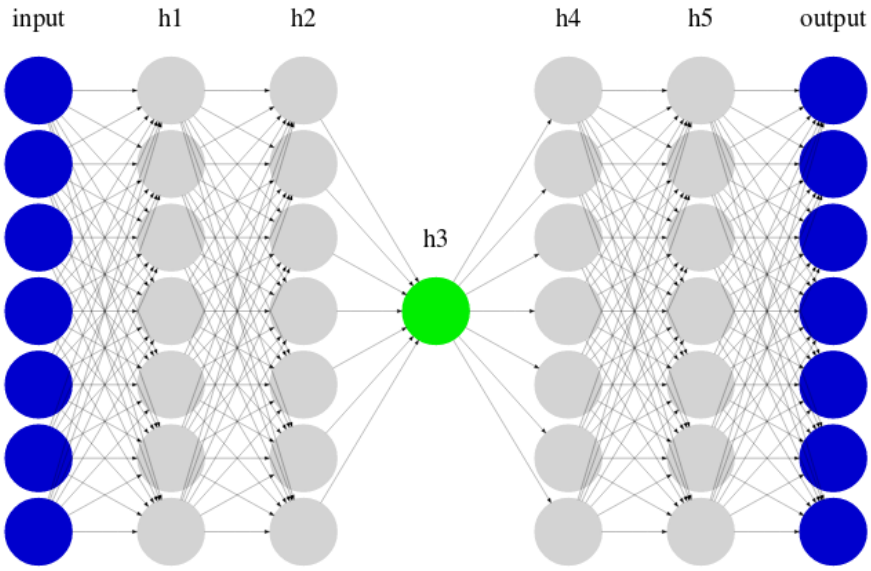


Figura 7: Esquema del autoencoder utilizado

El modelo que se estimó es el siguiente:

$$
\begin{aligned}
r_i &\sim \text{Bernoulli}(p = 0{,}5) \\
x &= input \\
h_1 &= \tanh(W_1(x * r_1) + b_1) \\
h_2 &= \tanh(W_2(h_1 * r_2) + b_2) \\
h_3 &= \tanh(W_3(h_2 * r_3) + b_3) \\
h_4 &= \tanh(W_4(h_3 * r_4) + b_4) \\
h_5 &= \tanh(W_5(h_4 * r_5) + b_5) \\
\hat{x} &= \sigma(W_6 h_5 + b_6)
\end{aligned}
$$

donde W_i es la matriz de coeficientes de la capa i, b_i es la media de la capa h.

De esta forma, la log-probabilidad para cada caso se puede escribir como:

$$
\Pi(\hat{x}, x) = \sum_{i=1}^{I} \sum_{k_i=2}^{K_i} w_{k_i}^{(i)} \left(x_{k_i}^{(i)} \log\left(\hat{x}_{k_i}^{(i)}\right) + \left(1 - x_{k_i}^{(i)}\right) \log\left(1 - \hat{x}_{k_i}^{(i)}\right) \right)
$$

donde $w_{k_i}^{(i)}$ se define para cada variable v_i con sus K_i categorías como:

$$
w_{k_i}^{(i)} = \frac{1}{K_i} \cdot \left(\frac{\sum_\ell K_\ell}{I} \right)
$$

La función de perdida (*loss function*)se definió mediante una verosimilitud pesada:

$$L(\hat{x}, x) = \text{argmin}_{W,b} \sum_{j=1}^{N} \Pi(\hat{x}(j), x(j))$$

El modelo fue entrenado con ADAM [15] definiendo como bachas de datos remuestreos con repetición sobre la distribución empírica de los casos para favorecer la convergencia.

Como se trabajó con la población completa (ya que el conjunto de datos es el mismo censo), el objetivo del modelo era la generación de una medida descriptiva. Es por ello que se tomó como criterio de evaluación del modelo la capacidad de explicar a la misma población utilizando como métrica el promedio ponderado de la probabilidad de cada variable-categoría:

$$Error(\hat{x}, x) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{I} \sum_{k_i=2}^{K_i} \frac{1}{\sum_\ell K_\ell} e^{x_{k_i}^{(i)} \log\left(\hat{x}_{k_i}^{(i)}\right) + \left(1 - x_{k_i}^{(i)}\right) \log\left(1 - \hat{x}_{k_i}^{(i)}\right)}$$

El modelo final conserva un 85 % del total de la información de input.

De esta forma, a partir de h_3 cada jefe de hogar, y por ende, cada hogar, queda clasificado con un valor resultado del autoencoder el cual llamaremos s_i, INSE.

A partír de INSE se generó una medida agregada para cada radio tomando como base el nivel socioeconómico que posee cada jefe de hogar. Siendo así, se define para cada jefe de hogar i con un índice de nivel socioecnómico s_i que habita en el radio censal r con una población de n_r jefes de hogar la variable *eta* como

$$\eta_r = \frac{1}{4} Q_{,25}(\mathbf{s}_{r_{1,\dots,n_r}}) + \frac{1}{2} Q_{,5}(\mathbf{s}_{r_{1,..,n_r}}) + \frac{1}{4} Q_{,75}(\mathbf{s}_{r_{1,..,n_r}}) \tag{5}$$

donde Q_p es el correspondiente p cuantil y η_r el resultado de aplicar la medida resumen conocida como "Tukey Trimean"la cual logra un compromiso entre robustez y eficiencia comparada a la mediana.

Como puede verse en los density plots previos, el índice construido parece captar en buena medida las disparidades provinciales:

- la CABA presenta una distribución claramente sesgada hacia la derecha (valores más altos del INSE)
- provincias como Chaco, Formosa, Jujuy, Salta presentan distribuciones sesgadas hacia la izquierda.

4.5. Cálculo del Índice de Vulnerabilidad Sanitaria

4.5.1. Dimensiones consideradas

La construcción del Índice de Vulnerabilidad Sanitaria contempló las siguientes dimensiones:

- Nivel Socioeconómico

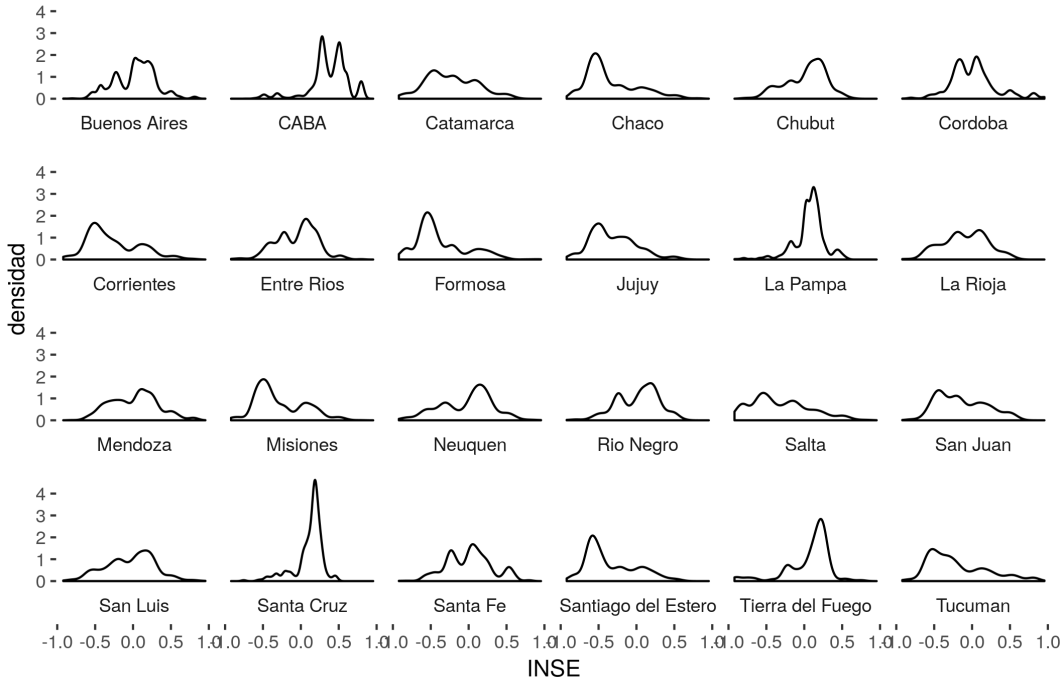


Figura 8: Density plot del INSE por provincia

- Acceso a Centros de Salud

La primera de ellas, como hemos mencionado, puede ser considerada un determinante de carácter general de la situación sanitaria de la población y se refiere al acceso a bienes sociales como la educación, la salud y a las condiciones de habitabilidad. Si bien estos bienes se encuentran asociados a trayectorias individuales y sociales, también mantienen una relación con el lugar de residencia.

Para construir el Índice de Vulnerabilidad Sanitaria final fue necesario combinar ambas variables a nivel radio censal. La estrategia adoptada para generar un índice compuesto fue la de componentes principales, es decir buscar la combinación lineal de las variables que pueda explicar la máxima varianza.

Ahora bien, cuando la distribución de las variables presentan distribuciones atípicas, es decir multimodales, asimétricas y/o con colas pesadas, la interpretación de los componentes principales resulta dificultosa ya que el método es sensible a la escala de las variables [24]. Una posible solución es transformar las variables utilizando los rangos de los datos [3] [2]

Por ello, siguiendo a [32], se buscó estandarizar cada una de las X_j variables mediante la transformación rankit:

$$rankit(X_{i,j}) = \frac{r_j(x_{i,j}) - 0{,}5}{n} \tag{6}$$

Para cada observación i de cada variable X_j se calcula el rango $r_j(X_{i,j})$ -que varía entre 1 y n-, se le resta 0.5 y se divide por el total de registros n.

De esta manera, se eliminaron tanto las diferencias de unidades que pudieran existir entre variables, y se logró invarianza frente a cambios de escala, desplazamientos y transformaciones monótonas.

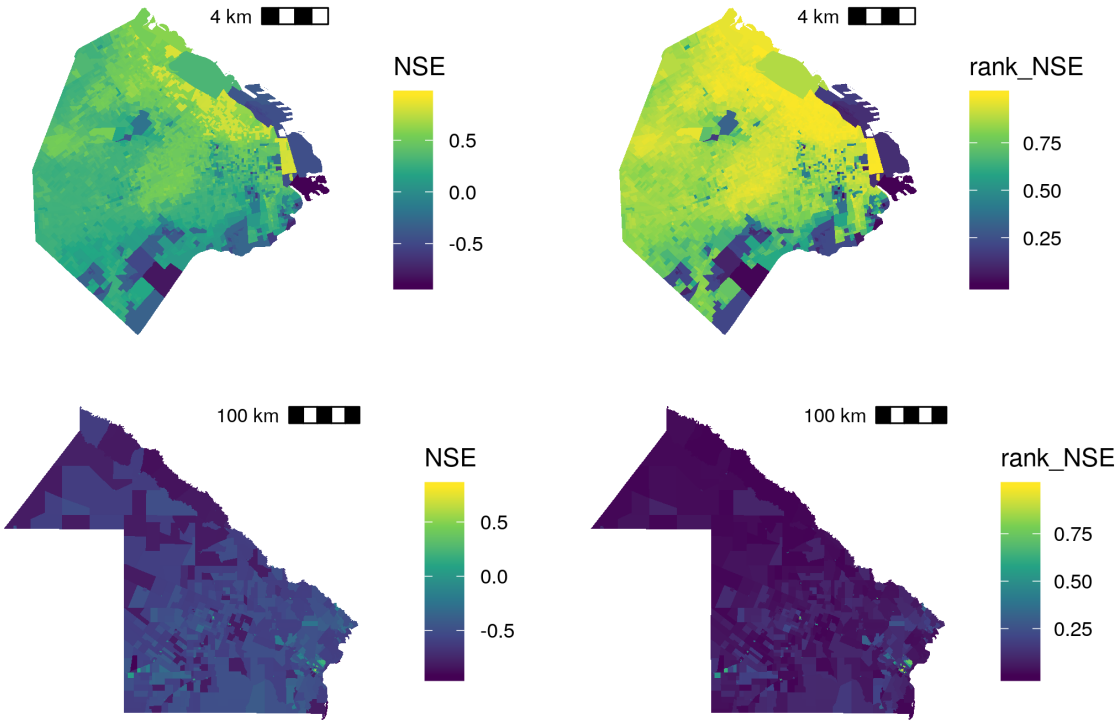


Figura 9: INSE en escala natural y en escala rankit. Arriba CABA, debajo Chaco.

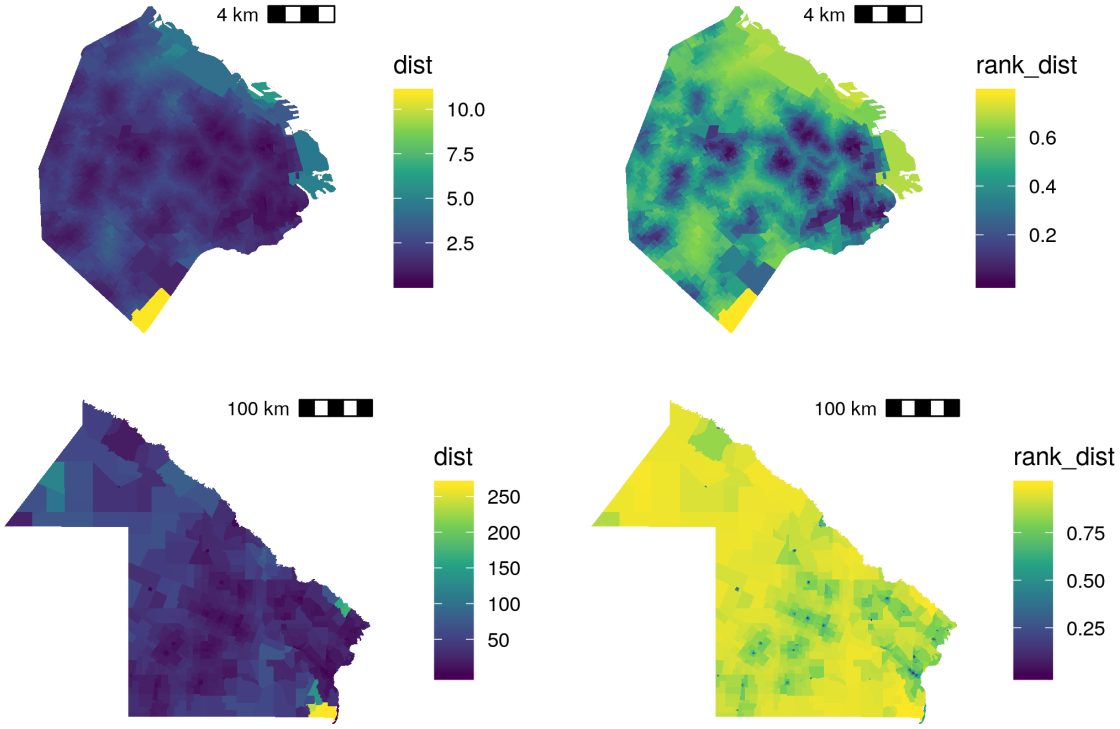


Figura 10: Distancias a Centros de Salud en escala natural y en escala rankit. Arriba CABA, debajo Chaco.

Puede notarse en los plots anteriores el efecto que la transformación rankit tiene en las dos variables utilizadas: Nivel Socioeconómico -η_r- y Acceso a Centros de Salud -Δ_r-.

Es fácil notar que la transformación rankit no es otra cosa que un estimador no paramétrico de la función distribución acumulada. Luego, y siguiendo a Egger et all. [12] y a Han y Liu [19], se procedió a realizar un análisis de componentes principales semiparamétrico:

Para realizar la combinación de ambas variables se calculó la correlación de Spearman sobre los rankits [19] y se descompuso la matriz de correlación no paramétrica. Se encontró que el principal autovector absorbía el 72 % de la variabilidad de los rankits y que estaba orientado a la dirección inversa de crecimiento mutuo. Este primer autovector fue considerado, entonces, la combinación óptima entre ambas variables. De esta manera, y mediante la aplicación sucesiva de transformaciones no paramétricas, se generó un índice tolerante a contaminación y sobre todo independiente a las características de medición de cada variable de entrada.

Como se verá mas adelante, el interés radica en tener una índice entre 0 y 1 cuya distribución sea homogenea para todas las unidades de análisis. Se aplicó nuevamente una transformación del tipo de función de distribución acumulada estimada mediante logsplines [27] usando AIC como criterio de regularización sobre la principal dirección, quedando conformado de esta

Algorithm 1 Componentes Principales Semiparamétrico

procedure S-PCA(X)
    for `j en variables` **do**
        for `i en casos` **do**
            $z_{i,j} = \Phi^{-1}\left(\frac{r_j(x_{i,j})-0{,}5}{n}\right)$ ▷ Φ^{-1} es la inversa de la f.d.a. gaussiana
        end for
    end for
    calcular Σ ▷ Matriz de covarianza
    encontrar U, S tal que $\Sigma = \frac{1}{n} U S^2 U^t$ ▷ SVD
    retornar ZU^t, S ▷ Coordenadas y autovalores
end procedure

forma el índice VS_r para cada uno de los radios:

A diferencia de rankit, al modelar la distribución con logspline se respetan las distancias relativas.

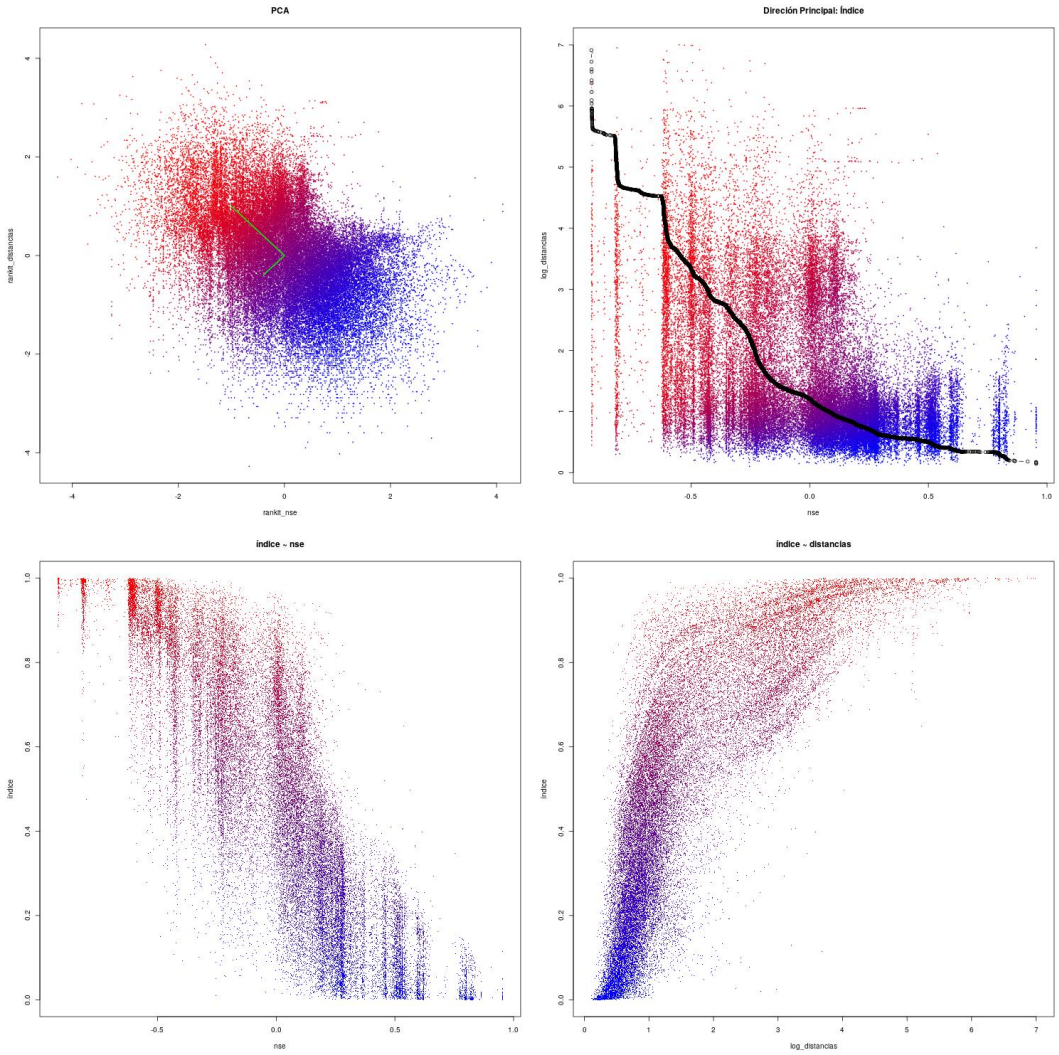


Figura 11: Índice de Vulnerabilidad Sanitaria

En el primer gráfico de la figura nueve se observa el resultado del S-PCA en el espacio de las Z, y los colores representan el valor del índice resultante. En el segundo gráfico se observa el resultado en el espacio de las X. Finalmente en los dos gráficos inferiores se observa el índice en función de INSE y distancias respectivamente.

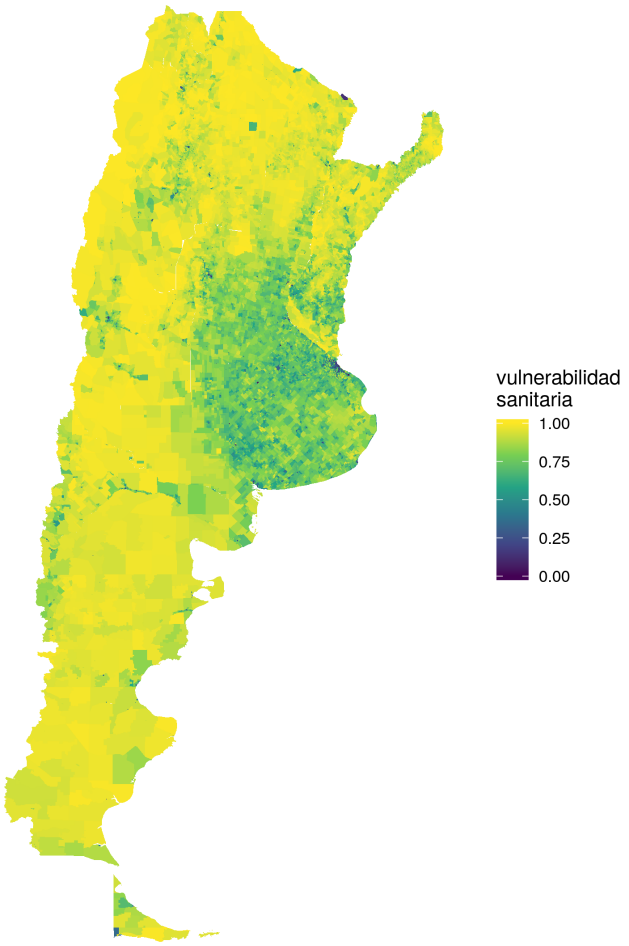


Figura 12: Índice de Vulnerabilidad Sanitaria a nivel radio censal (Argentina).

Ahora bien, esta visualización general debe ser tomada con cuidado debido a la escala (total del país) con la que se está trabajando. En efecto, el mapa anterior parece presentar una situación bastante esperable, dado que es posible detectar dos grandes zonas:

- Región “central": en esencia Buenos Aires, CABA y los grandes aglomerados urbanos de cada provincia, caracterizada por mayores valores de vulnerabilidad sanitaria,
- el resto del país, principalmente las zonas menos densamente pobladas, con valores críticos.

Ahora bien, al agregar la información a un nivel superior (fracción censal[14]), el panorama es ligeramente distinto:

[14]Una fracción censal es la subidivisión inmediatamente anterior al radio censal. Es decir, una fracción

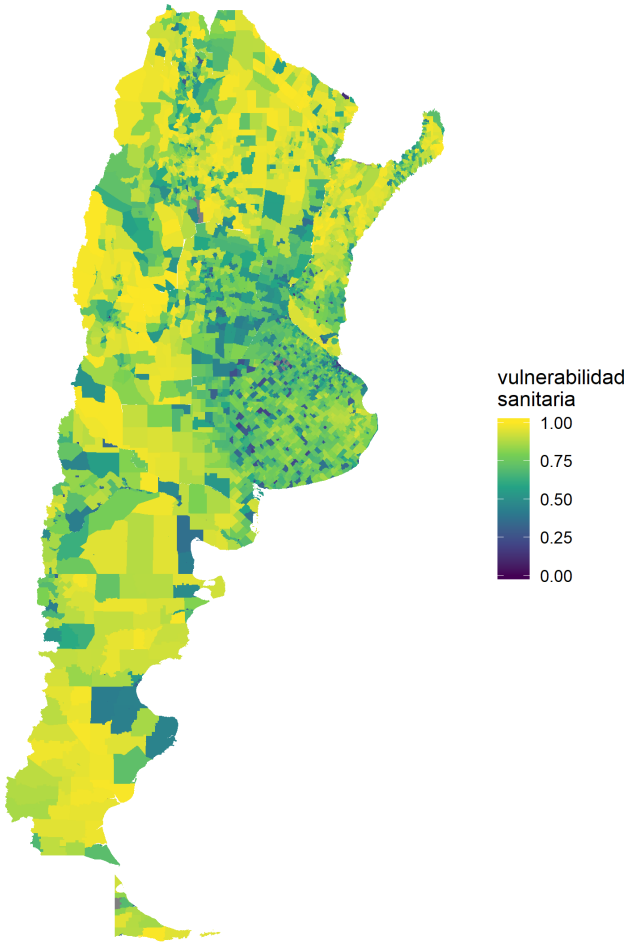


Figura 13: Índice de Vulnerabilidad Sanitaria a nivel fracción censal (Argentina).

Este mapa arroja una visión más matizada: las regiones detectadas previamente no son tan marcadas. Se observan en la zona central regiones con valores críticos. A su vez, en la zona no central, aparecen ahora sectores mejor rankeados en relación a este indicador.

A su vez, al focalizar la mirada en ciertas áreas (pero mantentiendo el nivel de desagregación por radio), también se logran destacar patrones invisibles en el mapa agregado. Si se observa la distribución final del Índice de Vulnerabilidad Sanitaria construido para el total del país y para las provincias de Chaco y la CABA se obtiene el siguiente resultado:

está compuesta por un conjunto de radios censales cercanos[23]

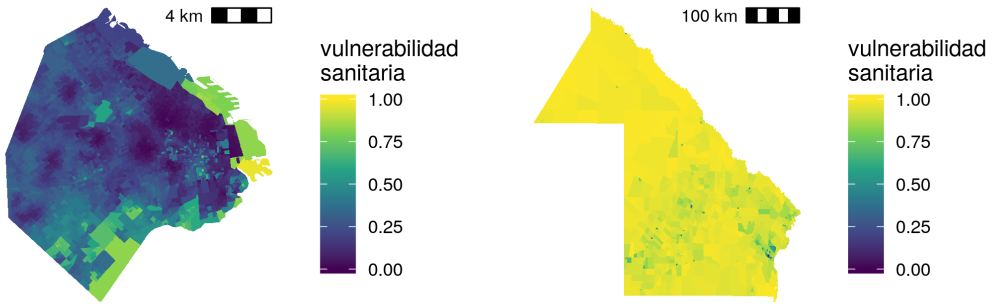


Figura 14: Índice de Vulnerabilidad Sanitaria. CABA (izq.), Chaco (der.)

Así, en la CABA es posible detectar nichos de alta vulnerabilidad santiaria y en Chaco existen zonas que presentan valores satisfactorios de este indicador.

5. Densidad poblacional

5.1. Densidad poblacional como factor de escala

Como se ha mencionara antes, el Mapa de Prevalencia Potencial de Chagas tiene como uno de sus objetivos principales la detección de áreas con altas chances de hallar población potencialmente portadora de la enfermedad fuera del área endémica ampliamente conocida. El fin último del mapa es lograr orientar acciones e intervenciones por parte del Estado (en sus diferentes niveles). Esto implica que la concentración de la población en cada uno de los radios es un factor de escala a considerar en el índice final.

Para el diseño de estrategias de intervención costo-efectivas en esta área es necesario cuantificar el impacto potencial en términos de cantidad de personas afectadas. En ese sentido, puede ser preferible (en términos del impacto final de la intervención en cuestión) realizar acciones en zonas con un nivel de prevalencia potencial ligeramente menor, pero con una concentración mayor de población en el territorio: el costo por persona tratada puede ser menor en estas zonas de alta concentración de población, minimizándose costos de transporte.

Es por ello que la densidad poblacional representa el componente final del índice. En efecto, para cada radio r se calculó la densidad como el cociente entre la cantidad de habitantes en el radio h_r y la superficie total en km^2 del radio censal Sup_r. Para poder combinarlo en el índice final, nuevamente se usó la estrategia de estandarización mediante la evaluación de la distribución de probabilidad acumulada F sobre la variable, lo cual remueve tanto las diferencia de unidades y logra invarianza frente a cambios de escala, desplazamientos y transformaciones monótonas. La función de densidad acumulada fue estimada mediante logsplines.

$$ds_r = \hat{F}_d\left(\frac{h_r}{Sup_r}\right) \tag{7}$$

6. Generación del Mapa de Alta Precisión de zonas de riesgo de Chagas en Argentina

6.1. Índice de Prevalencia Potencial de Chagas (IPPECh)

Para la construccion del *IPPECh* se tuvieron en cuenta las siguientes variables:

- Afinidad local con zona endémica de Chagas
- Vulnerabilidad Sanitaria
- Densidad de población

6.2. Forma funcional del *IPPECh*

Desde el punto de vista de las políticas públicas, con fondos a priori limitados, tendría sentido priorizar zonas con alta densidad de población y/o alta "afinidad" . Este índice compuesto, adecuadamente calibrado, permitiría establecer un orden en la expansión cual expandir políticas sanitarias destinadas a la atención de la ECh que tomen en cuenta distintos factores.

El índice final quedó compuesto de la siguiente manera:

$$IPPCh_r = \frac{{vs_r}^{\alpha}\ {d_r}^{\beta}\ Af_r}{\frac{1}{R}\sum_{r=1}^{R} {vs_r}^{\alpha}\ {d_r}^{\beta}\ Af_r} \tag{8}$$

donde:

* ${vs_r}^{\alpha}$ modera el efecto de la componente de Afinidad siendo α el parámetro que determina el impacto,

* ${d_r}^{\beta}$ penaliza a las areas despobladas y β funciona como el regulador,

* $0 \leq \alpha, \beta$, de manera tal que cuando $\alpha, \beta \longrightarrow 0$, se anula el efecto de la densidad y/o la vulnerabilidad sanitaria respectivamente. En cambio cuando $\alpha, \beta \longrightarrow \infty$, el índice se ve dominado la densidad y/o la vulnerabilidad sanitaria.

Respecto a $IPPCh_r$, ha de interpretarse como un índice relativo ya que sus valores indican cuánto mayor es la prevalencia potencial de Chagas en un radio censal respecto al valor mediano poblacional.

7. Resultados

El análisis de los resultados que se presenta en esta sección se propone detectar áreas de interés para futuro trabajo de campo y validación del modelo de estimación de prevalencia potencial[15]. Es decir, procura encontrar aquellas zonas en el páis donde el IPPECh es alto.

[15]Debe tenerse en cuenta que el listado es provisorio y deberá ser validado posteriormente considerando complejidades propias de la logística del trabajo de campo.

Esto quiere decir que se observa una alta afinidad -actividad de llamados celulares- con las zonas endémicas y valores elevados en el Índice de Vulnerabilidad Sanitaria como factor adicional para priorizar una subsiguiente intervención in-situ. Ambas dimensiones se encuentran combinadas en el índice final presentado en la sección anterior.

Para tal fin, se consideraron sólo áreas ubicadas fuera de la eco-región del Gran Chaco -aquellas donde un alto nivel de comunicación con el Gran Chaco no sea atribuible al volumen esperable de llamados entre personas físicamente cercanas.

Ahora bien, debe tenerse en cuenta que la unidad original del mapa (radios censales) no resulta viable para la organización del trabajo de campo: seleccionar radios censales relevantes (con alta afinidad) y sus correspondientes contrastes conduciría a una gran dispersión territorial de las áreas a visitar. Por esta razón se agregó la información a nivel localidad.[16]

Así, luego de filtrar aquellos radios con menos de 350 habitanes, se construyeron dos métricas para cada localidad:

1. el promedio de afinidad a lo largo de todos los radios que componen la localidad (ponderado por la población en cada radio),
2. el promedio de afinidad de los radios de la localidad que estuvieran entre los valores más altos de la distribución nacional de afinidad.[17]

Estas métricas dan lugar a tres tipos de situaciones posibles:

1. localidades que presentan alta afinidad pero "dispersa", exhibiendo una concentración similar en todos los radios que la componen; las llamaremos de *alta afinidad media*,
2. localidades que presentan valores de afinidad altos y altamente concentrados: es decir, pocos radios de la localidad presentan valores altos de afinidad; nominadas como localidades de *radios extremos*,
3. localidades que cumplen ambos criterios.

Lo anterior fue considerado para la construcción del listado que se presenta a continuación. Además, dentro de cada localidad se descartaron aquellos radios censales donde la densidad de población no alcanza los 350 habitantes por km^2, en tanto es necesario alcanzar una concentración de población local mínima para garantizar la viabilidad operativa de una eventual intervención en terreno.

Para favorecer constrastes y permitir variabilidad, fueron seleccionadas (selección estratificada) al interior de cada provincia localidades de los tres tipos mencionados.

[16] Para el INDEC una localidad "...es una porción de la superficie de la tierra caracterizada por la forma, cantidad, tamaño y proximidad entre sí de ciertos objetos físicos artificiales fijos (edificios) y por ciertas modificaciones artificiales del suelo (calles), necesarias para conectar aquellos entre sí." Es decir, "la localidad se concibe como concentración espacial de edificios conectados entre sí por una red de calles." Este es el criterio implícito en todos los censos argentinos, y explicitado a partir del censo de 1991.". Operativamente, cada radio censal pertenece a una y solo una localidad, por lo cual es posible agregar la información de cada radio a nivel localidad.

[17] Este umbral fue definido luego de comprobar que la distribución nacional de la afinidad es bimodal, y más específicamente, una mixtura de distribuciones. De esta forma, el umbral es el valor que mejor logra separar ambas distribuciones. En este caso, se usó un método clásico basado en el *boxplot*: el umbral fue calculado como $Q_{75} + 1{,}5 \times IQR$, es decir, el cuarto cuartil más 1,5 veces el rango intercuartílico. Todas las localidades cuyo valor de afinidad estuviera por encima de dicho umbral fueron consideradas en el cálculo de esta segunda métrica.

1. **Localidades de alta afinidad media:** fueron seleccionadas las tres con mayor afinidad media ponderada por la población de cada radio en cada provincia,
2. **Localidades de radios extremos y localidades con ambos tipos:** se siguió el mismo criterio, la diferencia es que en este caso, el promedio se calcula sobre los radios que sobreviven al filtro anterior; esto puede hacer que en algunas provincias no presenten localidades de estos tipos.

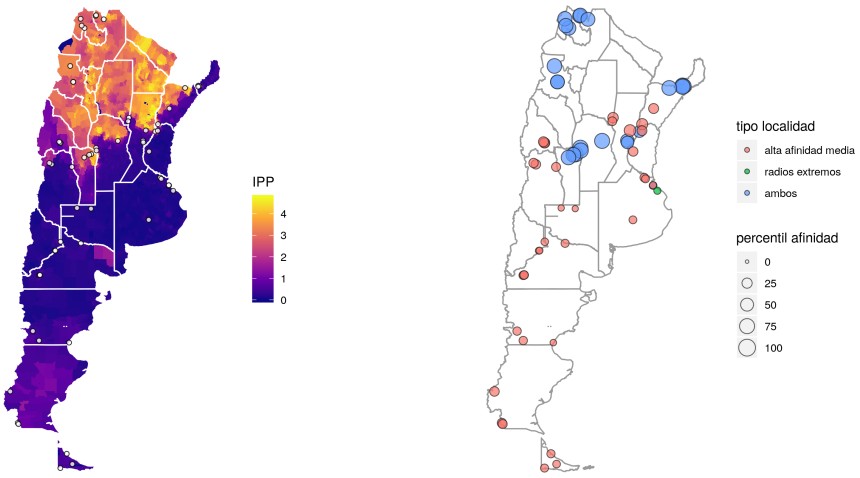


Figura 15: Localidades (homogéneas y concentradas) seleccionadas como áreas potenciales de intervención según niveles del IPPECh y según percentil de afinidad.

La figura 15 muestra las localidades seleccionadas, proyectadas sobre el mapa del IPPECh. Tal como se estableciera en [11], las áreas con mayores índices de Afinidad se concentran en la Región Metropolitana de Buenos Aires (incluyendo a la ciudad de La Plata) y en las provincias de la Patagonia Argentina -Neuquén, Río Negro, Chubut, Santa Cruz y Tierra del Fuego-. También aparecen como relevantes algunas localidades de áreas fronterizas al Gran Chaco, en provincias como Misiones, Entre Ríos y Mendoza.

Un primer punto a notar es que existe solamente una localidad (La Plata, Buenos Aires) que se caracteriza por ser presentar radio extremo pero no alta afinidad media. A su vez, puede notarse que entre las zonas fronterizas al Gran Chaco, predominan dentro de las localidades seleccionadas aquellas que combinan radios extremos y localidades de alta afinidad media. Estas localidades se caracterizan (como era esperable) por mayores valores en los percentiles de afinidad.

En cambio, en las zonas centro y Patagonia, la situación parece ser diferente: las localidades seleccionadas se caracterizan por ser de afinidad media alta y por menores valores en los percentiles de afinidad.

Provincia	Depto	Localidad	Tipo	Población	Afinidad	Vul_Sanit	Pct_usuarios
Buenos Aires	Zárate	COUNTRY CLUB EL CASCO	Alta Afinidad Media	532	0.1108	0.7566	0.1421
Buenos Aires	Zárate	LIMA	Alta Afinidad Media	10219	0.1472	0.7557	0.1909
Buenos Aires	Olavarría	VILLA ALFREDO FORTABAT	Alta Afinidad Media	3673	0.1072	0.6011	0.0373
Buenos Aires	La Plata	LA PLATA	Radio Extremo	645569	0.0773	0.3303	0.0489
Catamarca	Antofagasta	ANTOFAGASTA	Ambos	730	0.7067	0.7591	0.9139
Catamarca	Belén	EL DURAZNO	Ambos	785	0.6072	0.9664	0.9856
Catamarca	Belén	JACIPUNCO	Ambos	785	0.6072	0.9664	0.9856
Chubut	Río Senguer	ALTO RÍO SENGUER	Alta Afinidad Media	1570	0.1355	0.8920	0.1822
Chubut	Escalante	COMODORO RIVADAVIA	Alta Afinidad Media	175196	0.0525	0.4894	0.0811
Chubut	Río Senguer	RÍO MAYO	Alta Afinidad Media	2791	0.1365	0.8350	0.1835
CABA	Comuna 01	COMUNA 1	Alta Afinidad Media	205886	0.0859	0.2593	0.0922
CABA	Comuna 02	COMUNA 2	Alta Afinidad Media	157932	0.0747	0.0643	0.0616
CABA	Comuna 03	COMUNA 3	Ambos	187537	0.0845	0.1579	0.0984
Córdoba	San Justo	ARROYITO	Ambos	22147	0.7581	0.4437	0.9658
Córdoba	San Javier	LA TRAVESIA	Ambos	431	0.7713	0.9073	0.9090
Córdoba	San Javier	QUEBRADA DE LOS POZOS	Ambos	398	0.7889	0.7667	0.9377
Corrientes	Mercedes	FELIPE YOFRE	Alta Afinidad Media	1435	0.2365	0.8571	0.3119
Corrientes	Esquina	PUEBLO LIBERTADOR	Alta Afinidad Media	932	0.3271	0.9704	0.3601
Corrientes	Ituzaingó	VILLA OLIVARI	Ambos	1276	0.7223	0.8828	0.9421
Entre Ríos	Diamante	GENERAL RAMÍREZ	Alta Afinidad Media	8868	0.1974	0.3248	0.2038
Entre Ríos	La Paz	SAN GUSTAVO	Alta Afinidad Media	1325	0.2228	0.8876	0.2303
Entre Ríos	La Paz	LA PAZ	Ambos	24307	0.3851	0.6066	0.4222
Jujuy	Cochinoca	ABDÓN CASTRO TOLAY	Ambos	376	0.6617	0.8659	0.8802
Jujuy	Susques	CORANZULI	Ambos	520	0.6617	0.9638	0.8802
Jujuy	Rinconada	NUEVO PIRQUITAS	Ambos	672	0.6617	0.8632	0.8802
La Pampa	Curacó	GOBERNADOR DUVAL	Alta Afinidad Media	413	0.1216	0.9505	0.1723
La Pampa	Chalileo	SANTA ISABEL	Alta Afinidad Media	2697	0.0465	0.7905	0.0576
La Pampa	Loventué	TELÉN	Alta Afinidad Media	1240	0.0430	0.6464	0.0565
Mendoza	Luján de Cuyo	CACHEUTA	Alta Afinidad Media	701	0.2005	0.9349	0.2560
Mendoza	La Paz	DESAGUADERO	Alta Afinidad Media	641	0.1730	0.9688	0.1952
Mendoza	Luján de Cuyo	POTRERILLOS	Alta Afinidad Media	653	0.1944	0.8292	0.2367
Misiones	Candelaria	CANDELARIA	Ambos	13871	0.7816	0.6593	0.7849
Misiones	Candelaria	CERRO CORÁ	Ambos	684	0.7816	0.5185	0.7849
Misiones	Candelaria	PROFUNDIDAD	Ambos	417	0.7816	0.8254	0.7849
Neuquén	Confluencia	EL CHOCON	Alta Afinidad Media	467	0.0665	0.7586	0.0838
Neuquén	Añelo	SAN PATRICIO DEL CHAÑA	Alta Afinidad Media	6474	0.0936	0.3498	0.1437
Neuquén	Confluencia	VILLA EL CHOCÓN	Alta Afinidad Media	707	0.0665	0.6971	0.0838
Río Negro	Pilnaniyeu	CAÑADÓN CHILENO	Alta Afinidad Media	400	0.1427	0.9285	0.1759
Río Negro	Pilnaniyeu	COMALLO	Alta Afinidad Media	1497	0.1429	0.4506	0.1761
Río Negro	Pilnaniyeu	LAGUNA BLANCA	Alta Afinidad Media	400	0.1427	0.9285	0.1759
Salta	Orán	AGUAS BLANCAS	Ambos	2395	0.6161	0.6855	0.9681
Salta	Santa Victoria	NAZARENO	Ambos	916	0.6388	0.9380	0.9783
Salta	Santa Victoria	POSCAYA	Ambos	470	0.6388	0.9870	0.9783
San Juan	San Martín	DOS ACEQUIAS	Alta Afinidad Media	972	0.2378	0.7891	0.8782
San Juan	San Martín	VILLA DOMINGUITO	Alta Afinidad Media	619	0.2435	0.8815	0.8900
San Juan	San Martín	VILLA DON BOSCO	Alta Afinidad Media	2728	0.2434	0.7288	0.8896
San Luis	Lib. San Martín	LAS AGUADAS	Ambos	487	0.8143	0.9663	0.9492
San Luis	Ayacucho	SAN FRANCISCO DEL MONTE DE	Ambos	3732	0.6952	0.4792	0.8393
San Luis	Lib. San Martín	SAN MARTÍN	Ambos	697	0.7074	0.3581	0.7465
Santa Cruz	Güer Aike	28 DE NOVIEMBRE	Alta Afinidad Media	6145	0.1892	0.5642	0.2548
Santa Cruz	Lago Argentino	EL CHALTÉN	Alta Afinidad Media	1627	0.2058	0.8903	0.2566
Santa Cruz	Güer Aike	ROSPENTEK	Alta Afinidad Media	677	0.1745	0.5921	0.2426
Santa Fe	San Justo	RAMAYÓN	Alta Afinidad Media	550	0.3676	0.9028	0.3853
Santa Fe	La Capital	CANDIOTI	Ambos	888	0.5573	0.7285	0.7415
Santa Fe	La Capital	RECREO	Ambos	13336	0.4668	0.6728	0.2823
Santiago del Estero	Rivadavia	COLONIA ALPINA	Alta Afinidad Media	545	0.1573	0.8451	0.2027
Santiago del Estero	Rivadavia	SELVA	Alta Afinidad Media	2878	0.2747	0.7501	0.3351
Tierra del Fuego	Río Grande	RÍO GRANDE	Alta Afinidad Media	66475	0.1251	0.3677	0.1652
Tierra del Fuego	Río Grande	TOLHUIN	Alta Afinidad Media	2949	0.0986	0.8531	0.1273
Tierra del Fuego	Ushuaia	USHUAIA	Alta Afinidad Media	56593	0.1171	0.3341	0.1496

Cuadro 7: Localidades seleccionadas por provincia, según Afinidad, Vulnerabilidad Sanitaria, Población y Tipo

La Tabla 7 resume los atributos de las áreas más salientes, representando las 60 localidades seleccionadas en función de los criterios antes mencionados.

Si bien comparten altos niveles de afinidad, las zonas no exhiben patrones similares en cuanto a su vulnerabilidad sanitaria. En esa dimensión la variabilidad es alta, abarcando áreas en todo el rango desde la baja vulnerabilidad en lo relativo al resto del país (CABA, Vicente Lopez, La Plata), hasta altos niveles: El Durazno y Jacipunco (Catamarca), Pueblo Liberatdor (Corrientes), Coranzulis (Jujuy), Poscaya (Salta).

Entre las localidades seleccionadas, en la figura 16 se observan 6 con población mayor a 5000 personas, cada una en una provincia distinta y en orden decreciente de afinidad. De forma esperable por su cercanía geográfica a la zona endémica, las localidades en Misiones, Córdoba y Santa Fe exhiben una mayor afinidad respecto a sus contrapartes en el Centro y Sur del país. Sin embargo, cabe notar que la afinidad no disminuye en forma homogénea, es decir en gradiente continuo a medida que las localidades se alejan del área endémica. Al contrario, se detectaron localidades en la Provincia de Buenos Aires y en la Patagonia cuyo grado de afinidad es mucho más alta que el de centros poblados en provincias más cercanas a la EC -como La Pampa-. Esto sugiere la existencia de migración considerable desde regiones endémicas hacia las localidades resaltadas. Un indicio adicional al respecto es el alto porcentaje de usuarios locales de telefonía celular que llamó, o recibió llamadas desde la EC: Mas de una cuarta parte en 28 de Noviembre, cerca de un 20 por ciento en Lima y un 13 por ciento en Tolhuin.

Figura 16: De izquierda a derecha, y de arriba a abajo: Candelaria (Misiones), Arroyito (Córdoba), Recreo (Santa Fe), 28 de Noviembre (Santa Cruz), Lima (Buenos Aires), Tolhuin (Tierra del Fuego)

8. Discusión y Conclusión

Un primer resultado de la generación de este mapa supone la apertura de algunas líneas de trabajo que pueden resultar de interés. La más evidente de ellas supone un proceso de revisión, mejora y actualización. En efecto, varios de los indicadores y fuentes de información usadas como insumo son susceptibles de un proceso de mejora relevante.

Por un lado, podrían ampliarse las fuentes de información utilizadas incorporando, por ejemplo, registros obtenidos a lo largo de varios años para identificar dinámicas de cambio relacionadas con patrones migratorios. Así, la incorporación de datos censales que puedan cuantificar dichos patrones migratorios es una tarea pendiente.

Vinculado a este último punto, en el presente trabajo se utilizó el volumen de llamados de residentes de zonas no endémicas con el área endémica como un indicador proxy de la existencia de movimientos y contactos de población entre ambos tipos de zonas. Esto implica un salto inferencial: se asume que la la comunicación es un buen indicador de la existencia de flujos migratorios desde áreas endémicas. A su vez, un segundo supuesto es que parte de la población migrante está o estuvo infectada, al menos, con una probabilidad comparable a la de la población no migrante.

Si bien se trata de un supuesto razonable, también es esperable que los mismos CDRs permitan -a partir de nuevos procesamientos- estimar patrones y volúmenes migratorios efectivos. Es decir, a partir de la sucesión de conexiones de los usuarios a diferentes antenas podrían mapearse movimientos de población reales entre zonas endémicas y no endémicas.

Por otro lado, los datos provenientes de los CDRs presentan limitaciones que es necesario señalar. Quizás la más importante se vincule con el grado de cobertura de las antenas de telefonía. En efecto, esta variable altera la granularidad espacial del análisis. En zonas de baja densidad, la mayor distancia entre antenas hace que el área asignada a cada una sea más extensa, y por lo tanto mayor la dificultad para identificar el radio censal subyacente al que atribuir la actividad detectada. Por ello, que se recomienda como trabajo futuro realizar un análisis espacial de conglomerados con el fin de identificar áreas homogéneas de radios para mitigar dicho problema.

En cuanto a utilización de fuentes adicionales integradas en el Mapa (especialmente, el de Vulnerabilidad Sanitaria), una segunda iteración en la limpieza y consolidación de los datos georreferenciados de efectores de salud emerge como una tarea necesaria. Esta tarea implicaría la incorporación de fuentes que no fueron utilizadas en esta primera aproximación. A su vez, la clasificación según nivel de complejidad de la atención de los efectores de salud merece una revisión. Dichos efectores fueron clasificados en tres niveles en base a su nombre, o tipología declarada en la fuente original, pero sin precisiones acerca de la capacidad real de cada uno. Es decir, no se determinó para cada instancia si el efector cuenta con los recursos necesarios para diagnosticar o atender enfermos de Chagas.

También la incorporación de nuevas fuentes de datos y dimensiones vinculadas al riesgo y vulnerabilidad ambiental resultan de potencial interés para el trabajo futuro.

El trabajo resultó en la obtención de resultado valioso de carácter general, que tiene que ver con la generación de una metodología de trabajo replicable y aplicable a otros ámbitos. La cuantificación de interacciones entre áreas geográficas dispersas, así como los indicadores de Vulnerabilidad Sanitaria, pueden ser considerados como dimensiones transversales que

afectan la evolución y la transmisión de otras patologías además de la ECh. Por ello, una segunda línea de trabajo a futuro se vincula con la posibilidad de considerar mapas como el aquí descripto (y sus posteriores actualizaciones) como un insumo para el estudio de otras enfermedades infecciosas.

Referencias

[1] F. Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[2] J. Bacon-Shone. Ranking methods for compositional data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(3):533–537, 1992.

[3] M. J. Baxter. Standardization and transformation in principal component analysis, with applications to archaeometry. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):513–527, 1995.

[4] K. Beardsley, E. Wish, D. Fitzelle, K. O'Grady, and A. Arria. Distance traveled to outpatient drug treatment and client retention. *Journal of Substance Abuse Treatment, 25*, 2003. https://linkinghub.elsevier.com/retrieve/pii/S0740547203001880.

[5] G. Busso. Migración interna y desarrollo territorial en argentina a inicios del siglo xxi. brechas e impactos sociodemográficos de la migración interna interprovincial. In *IX Jornadas Argentinas de Estudios de Población. AEPA*, Oct 2007.

[6] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41(22):224015, 2008.

[7] R. Chuit, R. E. Gurtler, L. Mac Dougall, E. L. Segura, and B. Singer. Chagas disease-risk assessment by an environmental approach in northern argentina. *Revista de Patologia Tropical*, 30(2):193–208, 2001.

[8] R. Chuit, L. Mac Dougall, M. C. Evequoz, A. M. Bressan, and M. Frias. Travel and infectious disease. *Revista de Patologia Tropical*, 32(1), 2008.

[9] R. Chuit, E. Subias, P. AC., I. Paulone, C. Wisnivesky-Colli, and E. Segura. Usefulness of serology for the evaluation of trypanosome cruzi transmission in endemic areas of chagas's disease. *Revista da Sociedade Brasileira de Medicina Tropical*, 22(3):119–129, 2003.

[10] J. de Monasterio, A. Salles, C. Lang, D. Weinberg, M. Minnoni, M. Travizano, and C. Sarraute. Analyzing the spread of Chagas disease with mobile phone data. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, aug 2016.

[11] J. de Monasterio, A. Salles, C. Lang, D. Weinberg, M. Minnoni, M. Travizano, and C. Sarraute. Uncovering the spread of an infectious disease with mobile phone data. In *Simposio Argentino de GRANdes DAtos (AGRANDA 2016)*, 2016.

[12] B. Egger, D. Kaufmann, S. Schönborn, V. Roth, and T. Vetter. Copula eigenfaces - semiparametric principal component analysis for facial appearance modeling. In N. Magnenat-Thalmann, P. Richard, L. Linsen, A. Telea, S. Battiato, F. H. Imai, and J. Braz, editors, *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 1: GRAPP, Rome, Italy, February 27-29, 2016.*, pages 50–58. SciTePress, 2016.

[13] J. C. Elizaga and J. Macisco. *Migraciones Internas. Teoría, método y factores sociológicos.* CELADE, 1975. https://repositorio.cepal.org/handle/11362/7611.

[14] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 57–64. IEEE, 2011.

[15] A. Gerón. *Hands-On Machine Learning with Scikit-Learn and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly, 2017. http://shop.oreilly.com/product/0636920052289.do.

[16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016. http://www.deeplearningbook.org.

[17] C. Grabovschi, C. Loignon, and M. Fortin. Mapping the concept of vulnerability related to health care disparities: a scoping review. *BMC Health Service Research, 13, 94*, 2013. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626765/pdf/1472-6963-13-94.pdf.

[18] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design*, 37(4):682–703, 2010.

[19] F. Han and H. Liu. Scale-invariant sparse pca on high-dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287, 2014.

[20] S. Huber, C. Rust, et al. Calculate travel time and distance with openstreetmap data using the open source routing machine (osrm). *Stata J*, 16:416–423, 2016.

[21] INDEC. Habitat y vivienda por medio de datos censales. calidad de los materiales de la vivienda (calmat). Technical report, Instituto Nacional de Estadística y Censos, 12 2003. https://www.indec.gov.ar/micro_sitios/webcenso/censo2001s2_2/datos/metod_construccion.pdf.

[22] INDEC. Censo nacional de población, hogares y viviendas 2010. base de datos. definiciones de la base de datos. Technical report, Instituto Nacional de Estadística y Censos, 2011. http://200.51.91.245/redarg/CENSOS/CPV2010rad/Docs/base.pdf.

[23] INDEC. Unidades geoestadísticas. cartografía y códigos genográficos del sistema estadístico nacional. Technical report, Instituto Nacional de Estadística y Censos, 2011. https://geoservicios.indec.gov.ar/codgeo/index.php?pagina=definiciones.

[24] I. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer New York, 2006.

[25] S. Juran, P. N. Broer, S. J. Klug, R. C. Snow, E. A. Okiro, P. O. Ouma, R. W. Snow, A. J. Tatem, J. G. Meara, and V. A. Alegana. Geospatial mapping of access to timely essential surgery in sub-saharan africa. *BMJ global health*, 3(4):e000875, 2018.

[26] C. Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, and X. Ma. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Geoinformatics, 2010 18th International Conference on*, pages 1–7. IEEE, 2010.

[27] C. Kooperberg and C. J. Stone. A study of logspline density estimation. *Computational Statistics and Data Analysis*, 12(3):327–347, 1991.

[28] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[29] M. Peyravian and N. Zunic. Methods for protecting password transmission. *Computers & Security*, 19(5):466–469, 2000.

[30] A. Prüss-Üstün and M. Neira. *Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks*. World Health Organization, 2016.

[31] C. Sarraute, C. Lang, J. de Monasterio, and D. Weinberg. Descubriendo Chagas con big data. In *XVII Simposio Internacional sobre Enfermedades Desatendidas*, Aug 2015.

[32] S. Solomon and S. Sawilowsky. Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, 2009.

[33] J. Steenbruggen, E. Tranos, and P. Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346, 2015.

[34] J. Timyan, G. B. SJ, D. M. Measham, and B. Ogunleye. Access to care: more than a problem of distance. 1993.

[35] Z. Wang, S. Y. He, and Y. Leung. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141–155, 2018.

[36] D. Weiss, A. Nelson, and H. e. a. Gibson. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature, 25*, 2018.

9. Anexo 1. Descripción de las variables utilizadas en el Índice de Nivel Socioeconómico (INSE)

Se describe en este anexo las variables utilizadas para la construcción del índice. Para una referencia más exhaustiva, ver [22].

9.1. Relación de parentesco (P01):

Indica la relación de cada miembro del hogar con quien ellos hayan designado como jefe o jefa del mismo. Refiere tanto a las relaciones de parentesco (sean consanguíneas o no) como a las relaciones de amistad, de trabajo o de otro tipo. Las categorías son:

1. Jefe(a): persona reconocida como tal por los demás miembros del hogar.
2. Cónyuge: es la pareja de unión legal (unión civil/matrimonio) o la pareja de unión "de hecho" del jefe o jefa.
3. Hijo(a) / hijastro (a): incluye hijos e hijas biológicos y adoptivos legales o de hecho del jefe(a) o criados del jefe o jefa y/o también incluye a los hijos e hijas del o la cónyuge aunque no lo sean del jefe o jefa.
4. Yerno / nuera: cónyuge o pareja del hijo(a) o hijastro(a) del jefe o jefa del hogar.
5. Nieto(a): nietos del jefe o jefa del hogar. Se incluye también a los hijos de hijastros(as).
6. Padre / madre / suegro(a): estas relaciones son aplicables también a vínculos funcionalmente equivalentes, tales como los de padrastros y madrastras con yernos, nietos, etcétera, referidos tanto a los matrimonios como a las uniones de hecho.
7. Otros familiares: personas con algún otro tipo de parentesco con el jefe o la jefa del hogar (por ejemplo: tíos, etcétera).
8. Otros no familiares: personas que forman parte del hogar y que no tienen parentesco con el jefe(a) del hogar (amigos, etcétera) y que no están comprendidos en ninguna de las categorías precedentes.
9. Servicio doméstico y sus familiares: persona contratada (legalmente o de hecho) para desarrollar tareas relacionadas con el cuidado y mantenimiento de la vivienda y/o de los integrantes del hogar. La retribución por su trabajo puede ser en dinero o en especie. Incluye a sus familiares.

9.2. Sabe leer y escribir (condición de alfabetismo - P07):

Refiere a la capacidad de leer, escribir y comprender una frase sencilla sobre la vida cotidiana en cualquier idioma. Se requiere el conocimiento de ambas capacidades.

1. Si
2. No

9.3. Asiste o asistió a un establecimiento educativo (P08):

Concurrencia actual o pasada (si asiste/asistió) a un establecimiento reconocido del sistema de la enseñanza formal. Comprende a los establecimientos del sector estatal o privado.

Según el Ministerio de Educación es importante destacar que para el cálculo de indicadores educativos y asistencia escolar, las edades deberían ser recalculadas al 30 de junio. Por ese motivo, actualmente se está trabajando para incorporar una variable "edad calculada al 30 de junio", con la que se podrán obtener las tasas e indicadores correspondientes.

1. Asiste
2. Asistió
3. Nunca asistió

9.4. Nivel educativo que cursa o cursó (P09):

Nivel que el censado cursó o está cursando en Argentina o en el exterior en la fecha del Censo. Las categorías para la variable Nivel educativo son:

1. Nivel inicial (jardín/preescolar): nivel de la estructura implementada a través de la Ley Federal de Educación y la Ley de Educación Nacional que comprende a los/as niños/as desde los cuarenta y cinco (45) días hasta los cinco (5) años de edad inclusive, siendo obligatorio este último año.
2. Primario: comprende los niveles de escolaridad primaria, de carácter obligatorio, cuya duración puede ser de seis o siete años (1° a 7° grado o 1° a 6° grado).
3. EGB: nivel de la estructura implementada a través de la Ley Federal de Educación, actualmente vigente en la provincia de Buenos Aires y otras provincias del país. La duración es de nueve años y se encuentra compuesta por tres ciclos de tres años cada uno, EGB 1 (1° a 3° grado); EGB 2 (4° a 6° grado) y EGB 3 (7° a 9° grado).
4. Secundario: nivel de escolaridad media o secundaria aún vigente. La duración puede ser de cinco o seis años (1° a 5° año o 1° a 6° año). Las escuelas técnicas/industriales y las dependientes de la Universidad, también tienen una duración de seis años; este sistema aún está vigente en la Ciudad Autónoma de Buenos Aires y en algunas provincias de nuestro país.
5. Polimodal: nivel de la estructura implementada a través de la Ley Federal de Educación, actualmente vigente en la provincia de Buenos Aires y otras provincias del país. La duración es de tres años.
6. Superior no universitario: nivel de estudios que se realiza en instituciones de educación terciaria no universitaria, estatales o privados, con planes de estudios aprobados por el Ministerio de Educación (de la Nación o de las provincias). Incluye los profesorados de nivel inicial, de adultos, de enseñanza especial (sordomudos, ciegos, sordos, etcétera) y educación física, historia, letras, etcétera. Comprende también especialidades no docentes, por ejemplo: especialización técnica industrial, periodismo, turismo, computación, bellas artes y la formación de oficiales de las fuerzas armadas.
7. Universitario: nivel de estudios que se realiza en universidades nacionales, provinciales o privadas. Comprende exclusivamente las carreras que otorgan títulos profesionales (abogado, agrimensor, fonoaudiólogo, ingeniero, profesor, etcétera) y las licenciaturas (en ciencias de la educación, letras, matemática, sistemas, etcétera).
8. Post-universitario (especialización, maestría o doctorado): nivel de estudios que comprende a las carreras de especialización, maestrías y doctorados acreditadas por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU) o por entidades privadas debidamente reconocidas por el Ministerio de Educación. Para acceder a este tipo de estudio se requiere contar con título universitario de grado. Este tipo de

educación se desarrolla en instituciones universitarias y en centros de investigación e instituciones de formación profesional superior que suscribieron convenios con universidades a esos efectos. La realización de carreras de Postgrado conduce al otorgamiento del título académico de Especialista, Magíster o Doctor.

9. Educación especial (para personas con discapacidad): modalidad del Sistema Educativo Nacional destinada a asegurar el derecho a la educación de las personas con discapacidades, temporales o permanentes, en todos los niveles y modalidades del Sistema Educativo. La educación especial brinda atención educativa en todas aquellas problemáticas específicas que no puedan ser abordadas por la educación común. Son escuelas donde pueden concurrir niños con discapacidad mental (leve, moderada o severa), con discapacidad sensorial, con discapacidad motora sin compromiso intelectual. Es el conjunto de servicios, técnicas, estrategias, conocimientos y recursos pedagógicos - dentro del Sistema Educativo Nacional- orientados a la atención de las personas con necesidades educativas especiales a causa de una discapacidad.

9.5. Completó ese nivel (P10):

Refiere a la aprobación del último año de estudio y a la obtención del diploma o certificado correspondiente a un determinado nivel.

1. Sí
2. No
3. Ignorado

9.6. Nivel educativo:

A partir de estas tres variables (P08, P08, P10) se construye la variable final que tiene las siguientes categorías:

1. Primario incompleto
2. Primario completo
3. Secundario incompleto
4. Secundario completo
5. Universitario incompleto
6. Universitario completo o superior

9.7. Condición de actividad (CONDACT)

Comprende a la población de 14 o más años que, en el período de referencia adoptado por el Censo, estuvo:

1. Ocupada: población que durante por lo menos una hora en la semana anterior a la fecha de referencia del censo desarrolló cualquier actividad (paga o no) que genera bienes o servicios para el "mercado". Incluye a quienes realizaron tareas regulares de ayuda en la actividad de un familiar, reciban o no una remuneración por ello y a quienes se hallaron en uso de licencia por cualquier motivo. Se excluye de la actividad económica los trabajos voluntarios o comunitarios que no son retribuidos de ninguna manera.

2. Desocupada: es la población que no hallándose en ninguna de las situaciones descriptas, desarrolló, durante las cuatro semanas anteriores al día del censo, acciones tendientes a establecer una relación laboral o iniciar una actividad empresaria (tales como responder o publicar avisos en los diarios u otros medios solicitando empleo, registrarse en bolsas de trabajo, buscar recursos financieros o materiales para establecer una empresa, solicitar permisos y licencias para iniciar una actividad laboral, etcétera).
3. Económicamente inactiva: comprende a la población de 14 y más años no incluida en la población económicamente activa. Incluye a jubilados, estudiantes y otras situaciones.

9.8. Hacinamiento (INDHAC):

Representa el cociente entre la cantidad total de personas del hogar y la cantidad total de habitaciones o piezas de que dispone el mismo (sin contar baño/s y cocina/s). Las categorías son:

1. Hasta 0,50 personas por cuarto
2. De 0,51 a 1,00 personas por cuarto
3. De 1,01 a 1,50 personas por cuarto
4. De 1,51 a 2,00 personas por cuarto
5. De 2,01 a 3,00 personas por cuarto
6. Más de 3,00 personas por cuarto

9.9. Tenencia de la vivienda y propiedad del terreno (PROP):

Refiere al conjunto de normas jurídico-legales o de hecho en virtud de los cuales el hogar ocupa toda o parte de una vivienda. Las categorías son:

1. Propietario de la vivienda y del terreno: la vivienda y el terreno en el que está ubicada la misma, pertenecen a alguno/s de los integrantes del hogar. El hogar tiene capacidad (garantizada legalmente) para disponer de la vivienda y del terreno, aún cuando alguno de ellos esté pendiente de pago o tenga posesión de los mismos sin haber escriturado. El propietario de una vivienda tipo "departamento", también lo es del terreno.
2. Propietario sólo de la vivienda: la vivienda (pero no el terreno en el que está ubicada) pertenece a alguno/s de los integrantes del hogar. El hogar tiene capacidad (garantizada legalmente) para disponer de la vivienda aún cuando ésta esté pendiente de pago o tenga posesión de la misma sin haber escriturado.
3. Inquilino: el hogar paga, por la utilización de toda o parte de una vivienda, una cantidad en dinero o en especie (anual, mensual, quincenal, etcétera), independientemente de que medie un contrato legal.

4. Ocupante por préstamo: el hogar utiliza la vivienda que le es facilitada gratuitamente por el propietario. La vivienda no es propiedad de ninguno de los ocupantes, no está en régimen de alquiler y no existe contraprestación alguna por el uso de la misma.
5. Ocupante por trabajo: el hogar utiliza la vivienda que es facilitada gratuita o semi-gratuitamente por el patrón, organismo u empresa donde trabaja alguno de los miembros del hogar en virtud de su relación laboral. Un ejemplo son los porteros, serenos, caseros, trabajadores rurales, etcétera.

6. Otra situación: el hogar utiliza la vivienda con una modalidad que no se ajusta a ninguna de las anteriores

9.10. Al menos un indicador NBI (ALGUNBI)

0. No
1. Sí

9.11. Calidad de los materiales (INMAT):

Refiere a la calidad de los materiales con que están construidas las viviendas (material predominante de los pisos y techos), teniendo en cuenta la solidez, resistencia y capacidad de aislamiento, así como también su terminación. Se clasifica la calidad de los materiales en:

1. Calidad I: la vivienda presenta materiales resistentes y sólidos tanto en el piso como en techo; presenta cielorraso.
2. Calidad II: la vivienda presenta materiales resistentes y sólidos tanto en el piso como en el techo. Y techos sin cielorraso o bien materiales de menor calidad en pisos
3. Calidad III: la vivienda presenta materiales poco resistentes y sólidos en techo y en pisos.
4. Calidad IV: la vivienda presenta materiales de baja calidad en pisos y techos

9.12. Calidad de conexión a servicios básicos (INCALSERV):

Refiere al tipo de instalaciones con que cuentan las viviendas para su saneamiento. Para este indicador, se utilizan las variables procedencia del agua y tipo de desagüe. Las categorías clasificatorias son

1. Calidad satisfactoria: refiere a las viviendas que disponen de agua a red pública y desagüe cloacal.
2. Calidad básica: describe la situación de aquellas viviendas que disponen de agua de red pública y el desagüe a pozo con cámara séptica.
3. Calidad insuficiente: engloba a las viviendas que no cumplen ninguna de las 2 condiciones anteriores.

9.13. Calidad constructiva de la vivienda (INCALCONS):

Este indicador se construye a partir de la calidad de los materiales con los que está construida la vivienda y las instalaciones internas a servicios básicos (agua de red y desagüe) de las que dispone.

1. Calidad satisfactoria: refiere a las viviendas que disponen de materiales resistentes, sólidos y con la aislación adecuada. A su vez también disponen de cañerías dentro de la vivienda y de inodoro con descarga de agua.

2. Calidad básica: no cuentan con elementos adecuados de aislación o tienen techo de chapa o fibrocemento. Al igual que el anterior, cuentan con cañerías dentro de la vivienda y de inodoro con descarga de agua.
3. Calidad insuficiente: engloba a las viviendas que no cumplen ninguna de las 2 condiciones anteriores.