Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Maestría en Tecnologías de la Información

Trabajo Final de Maestría en Tecnologías de la Información

«Minería de datos espacial como técnica de enfoque epidemiológico. Caracterización del estado de salud general de adultos mayores. Ciudad de Corrientes periodo 2019»

Autor: Lic. Leonardo Alcides Vallejos

Directora: Dra. Sonia Itati Mariño

Co- Directora: Dra. Paola V. Britos

Año 2024

Minería de datos espaciales

como técnica de enfoque epidemiológico

CARACTERIZACIÓN DEL ESTADO
DE SALUD DE ADULTOS MAYORES

Dedicatoria

"A mi esposa Stefani y a mi hijo Leonardo quienes fueron y son mi sostén durante toda esta trayectoria, y en mi vida"

Resumen

Las instituciones públicas deben sensibilizar y capacitar a sus servidores públicos sobre la importancia de la innovación. En ese sentido el presente Trabajo Final de Maestría se enfoca en la implementación de técnicas supervisadas y no supervisadas de minería de datos espaciales (SDM) del repositorio de PAMI. La minería de datos es una tecnología que busca descubrir conocimientos significativos en diferentes dominios, como la salud. Se utilizan estas técnicas para analizar las características de los afiliados, la distribución de algunas de sus patologías, localizar posibles relaciones con factores sociales, económicos, como también la gestión eficiente de recursos. El objetivo es generar conocimiento para apoyar la toma de decisiones a nivel gerencial. La aplicación de la minería de datos espaciales puede ayudar a extraer patrones útiles que carecen de visibilidad a simple vista. Los resultados actuales de la implementación han demostrado ser altamente efectivos en el análisis de las características de los afiliados y la distribución de sus patologías. Por otro lado, han permitido identificar zonas donde se requiere mayor presencia del organismo. En definitiva, han facilitado la generación de conocimiento para respaldar la toma de decisiones técnicas en el ámbito de la salud y la seguridad social.

***Palabras claves**: Minería de datos espaciales, Algoritmos supervisados y no supervisados, Sistemas de información geográficos, Epidemiologia, Salud Publica, Adultos mayores, Asignación de recursos.*

Abstract

Public institutions must raise awareness and train their public servants about the importance of innovation. In that sense, this Master's Final Project focuses on the implementation of supervised and unsupervised spatial data mining (SDM) techniques from the PAMI repository. Data mining is a technology that seeks to discover significant knowledge in different domains, such as health. These techniques are used to analyze the characteristics of the members, the distribution of some of their pathologies, to locate possible relationships with social and economic factors, as well as the efficient management of resources. The objective is to generate knowledge to support technical decision making at the managerial level. The application of spatial data mining can help extract useful patterns that lack visibility with the naked eye. The current results of the implementation have proven to be highly effective in the analysis of the characteristics of the members and the distribution of their pathologies. These techniques have made it possible to identify areas where a greater presence of the organism is required. In short, they have facilitated the generation of knowledge to support technical decision-making in the field of health and social security.

«Minería de datos espacial como técnica de enfoque epidemiológico. Caracterización del estado de salud general de adultos mayores. Ciudad de Corrientes periodo 2019.»

Keywords: *Spatial data mining, Supervised and unsupervised algorithms, Geographic information systems, Epidemiology, Public Health, Older adults, Resource allocation.*

Reconocimientos

"A Dios, fuente de sabiduría y fortaleza, por iluminar mi camino y darme la fuerza para alcanzar otro logro académico. A la educación pública, pilar fundamental en mi formación, por brindarme las herramientas y oportunidades necesarias para alcanzar mis metas académicas. A mis estimadas tutoras, la Dra. Sonia Mariño y la Dra. Paola Britos, por su invaluable orientación, apoyo y dedicación a lo largo de este proceso de investigación. Sin sus acompañamientos y sabiduría, este logro no habría sido posible. A mis compañeros de cursado, especialmente Federico, Fernando, Walter y Griselda, con quienes compartí las largas horas de cursado, estudios e hicieron mucho más amena esta experiencia. Finalmente, a todos mis compañeros y compañeras de trabajo que me han brindado su apoyo, tiempo y conocimientos de forma total y desinteresada. A todos ellos, mi más profundo agradecimiento."

Índice de contenido

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1 Introducción

"Quien quiera estudiar la medicina adecuadamente, debe proceder de la siguiente manera: en primer lugar, considerar las estaciones del año y sus efectos. Luego, los vientos, el frío y el calor, en sus características que son comunes a todos los países y en las que son propias de cada localidad. Deberíamos también considerar las cualidades de las aguas…"

Ya en la antigua Grecia Hipócrates relacionaba el concepto de salud y enfermedad con el componente geográfico, [1].

El presente Trabajo Final de Maestría se refiere a la implementación de técnicas supervisadas y no supervisadas de minería de datos espaciales (SDM) [2], sobre datos contenidos en un repositorio de la obra social para jubilados y pensionados INSSJP (Instituto Nacional de Servicios Sociales para Jubilados y Pensionado) en adelante PAMI.

La minería de datos, como tecnología orientada a descubrir conocimientos significativos se aplica en diferentes dominios. Por citar algunos estudios relacionados al campo de la salud, son los expuestos en [3]y [4]modelos aplicados en salud pública y epidemiologia.

Por medio de estas técnicas se pretende responder y aportar conocimiento en relación a las características de los afiliados con sus patologías, la distribución de las mismas, la búsqueda de posibles relaciones con otros factores como ser sociales, económicos, geográficos. Este banco de datos almacena una gran cantidad de información, que con el transcurrir del tiempo se va acrecentando. Sin embargo, estos datos permanecen como "estancados", sin sacar ningún beneficio, ni siquiera para la elaboración de datos estadísticos.

La aplicación de las técnicas de minería de datos espaciales puede permitir la extracción de patrones potencialmente útiles, que a simple vista no pueden visualizarse. En una primera instancia se pretende convertir estos datos semi-estructurados contenidos en planillas de cálculos, archivos de textos, etc. producto de los informes de auditorías médicas y sociales en una base de datos espacial agregando el dato georreferenciado.

Esto permitirá una mejor compresión de las diferentes patologías que afectan a nuestros afiliados, mapeando su distribución, estableciendo posibles causas, respondiendo a preguntas como ¿la distribución de las patologías son producto del azar? o ¿hay otras causas que lo expliquen?

Con el fin de estructurar la investigación, se detallan los capítulos que componen esta investigación. El *capítulo I "Introducción*" ofrece una descripción general de la temática abordada, estableciendo ámbito del problema, los objetivos perseguidos. En el capítulo II *"Marco teórico"* con el objetivo

de lograr una comprensión más profunda se realizó un repaso de todos los conceptos considerados necesario para el presente trabajo y su vinculación con el mismo. El capítulo III *"Metodología"* se detalla la manera de encarar el trabajo. En el capítulo IV "Resultados" se exponen toda la información obtenida fruto de la aplicación de la metodología propuesta. El capítulo V *"Discusión"* se examinan los resultados obtenidos en base a los objetivos planteados al inicio, estableciendo las conclusiones de la investigación.

1.2 Elección del tema

La investigación de esta problemática versó por el interés de aportar mayor conocimiento al momento de la planificación de los operativos en terrenos. Mejorar la asignación de recursos humanos, materiales, programas, en pos de detectar las poblaciones más vulnerables en la provincia. En definitiva, hacer uso inteligente de los recursos disponibles.

Por otra parte, profundizar conocimientos y habilidades en el campo de las tecnologías relacionadas al análisis de datos, ciencias de datos, etc. fue determinante en la elección.

1.3 Descripción del problema.

Desde una perspectiva macro, uno de los históricos inconvenientes en América Latina en general es la gestión ineficiente de los recursos públicos. Problemática que según [5], está más relacionada con la falta de planeación que con la escasez de recursos. Existía y se sigue observando un estilo de formular e implementar políticas públicas, caracterizado por la predisposición y la urgencia por actuar, con poco conocimiento y en forma inconsulta. Estilo que no respeta el denominado ciclo de políticas públicas. Donde en primera instancia surge y se identifica un problema, luego se analizan las posibles soluciones a aplicar, en tercer lugar, se opta por uno o más cursos de acción tendiente a resolver el problema (planificación) para terminar con las últimas fases de implementación y evaluación.

Tanto el pasado como el futuro - la dimensión temporal – son importantes en la gestión. El futuro, visto como un estado de situación deseable/ideal. Y el pasado, como referencia obligada para comparar si lo planificado y programado se ha logrado.

El sector de servicios sanitario no escapa a esta realidad. En el caso de las obras sociales, que es el tema que nos concierne en esta investigación, [6]presenta indicadores como apoyo a los procesos de planificación y evaluación. La búsqueda de medidas objetivas que reflejen el estado de salud de la población y midan la calidad y eficiencia del sistema es una antigua tradición en salud pública,

particularmente en epidemiología. Estos indicadores constituyen la materia prima para los análisis en el sector. Para una comprensión más en detalle el sistema de salud argentino - en cuanto a niveles, alcances y actores - dentro del cual opera el INSSJP se puede consultar el *Anexo N° 1*.

En razón de lo antes expuesto el lector puede deducir la importancia de los datos al momento de la planificación. El valor de los mismos radica en que sean relevantes para el cumplimiento de los objetivos del proyecto.

Sobre este punto es donde la implementación de técnicas de minería de datos puede generar un valor agregado en las instancias de análisis de los problemas y la asignación de recursos. En [7]encontramos otras áreas también que pueden ser beneficiadas, podemos citar las implementaciones en el diagnóstico y tratamiento de enfermedades, identificando los tratamientos más eficaces. Desde un enfoque comercial de la salud como empresa, la relación con los clientes (socios) y herramientas para manejarlas (como el CRM). Este es un tema importante, ya que uno de los criterios prioritarios a la hora de tomar una decisión es precisamente la relación (preferencias, reclamos, etc.) con los clientes y, hacerlo bien requiere poder adquirir un profundo conocimiento de los mismos.

Detección de fraudes e irregularidades. Para las empresas de seguros sanitarios contar con estas herramientas puede ser la diferencia al momento de prevenir delitos, fraudes.

El forecasting que ayuda a pronosticar la estadía de los pacientes hospitalizados, [8].

En cualquier caso, de los descriptos, se puede apreciar la utilidad de las técnicas en un campo tan sensible como es el de la salud. Y en especial cuando hablamos de la salud de los adultos mayores.

Exploración de los datos

Otra cuestión a tener en cuenta se relaciona en la manera en que se presentan los datos. Que, si bien no podemos de hablar de problema, si se puede mejorar su presentación. Esto ayudaría a una mejor interpretación de la información con la que se trabaja.

¿Qué importancia tiene la visualización de los datos?

La visualización de datos ayuda a contar historias, seleccionando los datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos. Una buena visualización, eliminando el ruido de los datos y resaltando la información útil es un producto final de consideración.

1.4 Objetivos generales

Desarrollar un proceso de minería de datos espaciales para apoyar toma de decisiones de la salud de adultos mayores. Caso de estudio de las tres enfermedades crónicas no transmisibles (ENT)1 más frecuentes detectadas en la ciudad de Corrientes, periodo 2019.

1.5 Objetivos específicos

Para la consecución de este objetivo se detallan los objetivos específicos del mismo

- Aplicar los conocimientos adquiridos en torno a las técnicas y las herramientas comprendidas en la minería de datos para el diseño de un proceso de datos espaciales.
- Proponer una solución innovadora a través del desarrollo de un proceso de minería de datos espaciales en el dominio de la salud publica
- Validar la propuesta utilizando un conjunto de datos representativos de las tres enfermedades más frecuentes detectadas en adultos mayores de la ciudad de Corrientes en 2019.

1.6 Preguntas de investigación

¿Cuáles son las áreas geográficas con mayores vulnerabilidades?

¿Existe un patrón o tendencia en la distribución espacial de las patologías?

¿Existen agrupaciones de interés?

¿Se pueden encontrar relaciones entre las variables estudiadas?

¿La elaboración de mapas con datos de salud tiene alguna utilidad?

1.7 Justificación y viabilidad

La presente investigación tiene los siguientes criterios de utilidad.

- Por un lado, una perspectiva teoría - práctica, que, a partir de la profundización de conocimientos referentes a ciencia de datos, minería de datos, algoritmos supervisados y no supervisadas demuestren su utilidad en el análisis de datos permitiendo:

- Determinar los niveles de morbilidad generales y por causas definidas.
- Identificar posibles agrupaciones espaciales de territorios con valores de los indicadores de morbilidad altos.
- Identificar posibles patrones espaciales morbilidad.

[1] Grupos de enfermedades que no son causadas por un agente infeccioso, estas incluyen cáncer, enfermedades cardiovasculares, diabetes y enfermedades pulmonares crónicas. También se incluyen lesiones y trastornos mentales.

- Identificar los territorios con una privación significativa de recursos y servicios en salud.

- Desde un enfoque metodológico se prevé una innovación en la fase de diseño y planificación de campañas que se desarrollen en territorio. Ya que estas técnicas de análisis de datos aportarían nuevos conocimientos sobre la realidad sobre la cual se interactúa. La posibilidad de presentar los datos de forma gráfica representa un potencial al momento de su interpretación, en contrapartida a los datos tabulados (datos en el repositorio).
- Viabilidad: el proyecto de investigación se considera factible y pertinente ya que se tiene acceso y autorización a los datos necesarios para el análisis seleccionado como caso de estudio.

Capítulo 2

Marco Teórico

2.1. Marco Teórico

En consideración a las diversas temáticas que incluye esta investigación se procederá a una breve introducción y así comprender la integración de los mismos.

En primera instancia introduciendo los conceptos sobre gestión de proyectos, luego se presenta información relacionada a la estadística, herramienta empleada en el análisis de nuestros datos. Dada la naturaleza espacial de los mismos, los sistemas de información geográfica (GIS) y el lenguaje de programación R ofrecen algoritmos para su tratamiento. Todo esto en el marco referencial de un proceso de minerías de datos espaciales.

Concluyendo esta sección con aportes de la salud pública, epidemiologia y el adulto mayor, dado el enfoque epidemiológico del trabajo

2.2 Gestión de proyectos

La definición de la asignación de recursos hace referencia al proceso de asignación y programación de activos que respaldan de manera efectiva los objetivos de un equipo, organización, etc., eligiendo los mejores recursos disponibles para sus proyectos, [9].

Pasos para la asignación de recursos:

- Crear una escala de tiempo del proyecto.
- Identificar los recursos significativos y el presupuesto.
- Designar los recursos disponibles.
- Asignar y distribuir los recursos.
- Seguimiento de la programación del proyecto.
- Elegir una técnica para abordar los obstáculos.

2.3 Análisis estadístico.

Seguidamente se proporcionan conceptos básicos mencionados en [10] relacionados a la estadística que nos familiarizan con las técnicas consideradas en el proyecto de investigación. Los siguientes párrafos no intentan ser exhaustivos, más bien entender de manera afable términos que no se relacionan directamente con las ciencias de la computación. No obstante, son indispensables para el entendimiento de las técnicas de minería de datos aplicadas.

Estadística espacial

En [11] indica que son muchas las áreas del conocimiento que requieren de herramientas aportadas por la estadística para el análisis de sus datos. Que si bien todas hacen uso de las características comunes que ella tiene, esto ha motivado la aparición de metodologías específicas. Tal es el caso de la geoestadística, una rama de la estadística que se centra en los conjuntos de datos de variables en el espacio, conocidas como variables regionalizadas, en las que cada valor está asociado a una posición particular en el espacio. Dicho de otra manera, la geoestadística se encarga del estudio de fenómenos con correlación espacial.

Su interés primordial es la estimación, predicción y simulación de dichos fenómenos.

Haciendo una breve reseña la geoestadística tiene sus orígenes en la década del 1970, con el propósito de predecir valores de las variables en sitios no muestreados. La geoestadística es solo una de las áreas del análisis de datos espaciales. Esta última se subdivide en tres grandes áreas:

- Geoestadística: Las ubicaciones s provienen de un conjunto D continuo y son seleccionadas a juicio del investigador (D fijo). Por ejemplo, valores de precipitación en un país determinado, medido en las diferentes estaciones meteorológicas en un mes dado. Es importante resaltar que en geoestadística el propósito esencial es la interpolación y si no hay continuidad espacial pueden hacerse predicciones carentes de sentido.
- Lattices: Las ubicaciones s pertenecen a un conjunto D discreto y son seleccionadas por el investigador (D fijo). Algunos ejemplos podrían ser: tasa de accidentes de tránsitos en diferentes zonas de una ciudad o tasa de morbilidad de Covid - 19 en Argentina medida por provincias. En estos ejemplos se observa que el conjunto de ubicaciones de interés es discreto, es decir, corresponden a agregaciones espaciales más que a un conjunto de puntos del espacio. Para estos casos no tienen sentido la interpolación espacial debido al tipo de datos.
- Patrones espaciales: las ubicaciones pertenecen a un conjunto D que puede ser discreto o continuo y su selección no depende del investigador (D aleatorio). Ejemplos de datos de este tipo son: localización de nidos de pájaros en una región dada; puntos de imperfectos dentro de una placa metálica. No es difícil observar la naturaleza aleatoria de los datos. No dependen del investigador la localización de los mismos. En [12] hallamos un trabajo de investigación sobre la leucemia infantil en la ciudad de North Humberside (Inglaterra).

Análisis de datos exploratorios (EDA)

Las técnicas detalladas a continuación son aplicadas en el proceso denominado análisis exploratorio de datos sus siglas en ingles EDA (Exploratory Data Analysis). Este análisis fue propuesto por John Tukey[2] por cuatro objetivos:

- Sugerir hipótesis acerca de las causas de los fenómenos observados.
- Comprobar las suposiciones sobre las cuales se basarán las inferencias estadísticas
- Dar soporte a la selección de las herramientas estadísticas para el análisis posterior de esos datos
- Sentar las bases para expandir la base de datos mediante relevamientos y experimentación.

Las técnicas EDA han sido incorporadas en la minería de datos. El EDA apela a recursos gráficos y no gráficos, y a técnicas estadísticas uni y multivariada; también a procedimientos matemáticos sofisticados como el análisis por componentes principales.

El EDA se apoya fuertemente en los métodos gráficos, ya que el ser humano es bueno en identificar patrones visualmente; histogramas, diagramas de caja, gráficos de cuantiles y otros, son recursos muy utilizados.

- Regresión lineal: ampliamente utilizado para el análisis predictivo.
- Correlación: La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante). Es una herramienta común para describir relaciones simples sin hacer afirmaciones sobre causa y efecto. Es una tarea descriptiva que analiza el porcentaje de similitud entre los valores de dos variables numéricas.
- Análisis de Componentes principales (PCA): El análisis de componentes principales (PCA) es una técnica matemática para reducir la dimensión de un conjunto de datos
- Análisis de agrupamiento: implica agrupar un conjunto de objetos de modo tal que los objetos contenidos en un grupo, o cluster, muestren una mayor semejanza entre sí que si se los compara con los objetos de los otros clusters. Es una actividad propia de la minería de datos, que considera al agrupamiento una técnica de aprendizaje no supervisado. El análisis de agrupamiento incluye dos grandes divisiones, el Análisis de Agrupamiento Jerárquico, también conocido como

[2]Estadístico estadounidense nacido en New Bedford, Massachusetts, conocido entre otras cosas por el desarrollo de la FFT (Fast Fourier Transform, que es un algoritmo para el cómputo de la DFT o Discrete Fourier Transform), así como el Diagrama de la caja y bigotes, o Box plot.

agrupamiento basado en conectividad, y el Análisis de Agrupamiento por Promedios k (K-means), también conocido como agrupamiento basado en centroides.

2.4 Sistemas de información geográficos (GIS)

Esta sección trata sobre los sistemas de información geográficos. Su elección para este proyecto se justifica en base a la capacidad de esta tecnología para soportar y procesar información proveniente de diferentes fuentes y formatos, [13]. Esto hace obligatorio adentrarse a términos relacionados con la misma.

Antes de abordar cuestiones técnicas de un GIS, consideramos adicionar el concepto de Sistemas de Vigilancia en Salud Publica y su relación con un sistema GIS.

[14] define la Vigilancia en Salud Publica como un proceso permanente y continuo de observación, recolección, análisis, interpretación y divulgación sobre los eventos de salud y sus factores condicionantes.

La posibilidad de contar con agentes (trabajadores sociales, médicos, administrativos) que despliegan sus trabajos en terreno hace posible el levantamiento de información que alimente al sistema de información geográfica a desarrollar. A través de informes de auditorías socio-ambientales en domicilios particulares (afiliados), registros de internaciones en centros de salud (hospitales, clínicas, geriátricos, etc.), además de los registros censales de población obtenidos por organismos estatales, conforman los insumos del sistema.

Son muchas las obras que mencionan la importancia de los datos con referencia espacial, hace tiempo que cerca del 70% de los datos manejados en la mayoría de las disciplinas contienen un componente espacial/geográfico. Esto va en aumento en la actualidad gracias al desarrollo de herramientas que proporcionan este tipo de dato, GPS, sensores, y un largo etc. Cuando hablamos de dato o información georreferenciada hacemos alusión a información que puede ser ubicada geográficamente, por ejemplo, pozos de aguas, eventos de salud, fábricas en una determinada región, etc. Esto ha permitido a la geografía ser un elemento de suma importancia en relación a muchas otras áreas del conocimiento.

En cuanto a una definición precisa de qué es un GIS no es posible, hay variadas definiciones de acuerdo a las posibilidades o funcionalidades que más se valore. Aun cuando existen diversas definiciones podemos resaltar algunos conceptos sobre los GIS, a saber:

- Lectura, edición, almacenamiento y, en términos generales, gestión de datos espaciales.

- Análisis de dichos datos. Esto puede incluir desde consultas sencillas a la elaboración de complejos modelos, y puede llevarse a cabo tanto sobre la componente espacial de los datos (la localización de cada valor o elemento) como sobre la componente temática (el valor o el elemento en sí).
- Generación de resultados tales como mapas, informes, gráficos, etc.

De lo anterior se desprende que un GIS funciona como un elemento integrador de tecnologías y personas de distintas disciplinas. Si hablamos de disciplinas relacionadas a la tecnología y manejo de información podemos citar a ciencias de la información, la informática, el diseño de bases de datos o el tratamiento digital de imágenes, la estadística y las matemáticas, [15].

Otras dedicadas al estudio de la Tierra desde un punto de vista físico ambiental (geología, oceanografía, la ecología), y desde un plano social tenemos a la antropología, la geografía o la sociología, entre otras.

Componentes de un SIG [14].

Es habitual citar tres subsistemas fundamentales:

- Datos. Los datos son la materia prima necesaria para el trabajo en un SIG, y los que contienen la información geográfica vital para la propia existencia de los SIG.
- Métodos. Un conjunto de formulaciones y metodologías a aplicar sobre los datos.
- Software. Es necesaria una aplicación informática que pueda trabajar con los datos e implemente los métodos anteriores.
- Hardware. El equipo necesario para ejecutar el software.
- Personas. Las personas son las encargadas de diseñar y utilizar el software, siendo el motor del sistema SIG.

Para poder considerar a un GIS herramienta útil, debe poder integrar estos subsistemas.

Los datos son un componente fundamental en un sistema de información geográfico, se podría decir que es pilar al cual se conectan los demás (visualización y análisis).

El dato geográfico tiene dos componentes: un componente espacial que hace referencia a la posición dentro de un sistema de referencia establecido, y la componente temática responde a la pregunta ¿qué? Esta pregunta se relaciona con cierto proceso o en ella aparece algún fenómeno dado, como los ejemplos citados con anterioridad (pozos de agua, eventos de salud, etc.)

Un concepto a tener en cuenta en relación con las componentes de la información geográfica es la dimensión, la cual se aprecia en la Fig. 1.

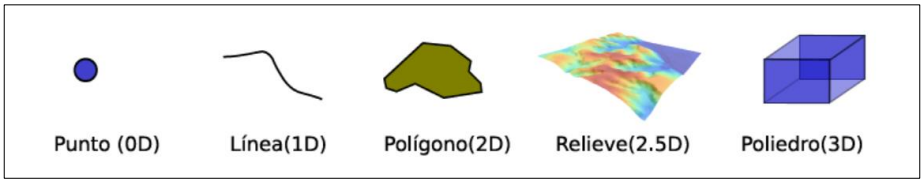


Fig. 1. Dimensión de los datos geográficos. Fuente [14]

Capas. División vertical de la información.

Uno de los grandes éxitos de los SIG es su estructura de manejo de información geográfica, que facilita todas las operaciones que se llevan a cabo con esta. Esto se logra a través de las capas. Una capa es un archivo, o parte de un archivo, que contiene información espacial de una sola variable, pudiendo contener información espacial y temática. Para ejemplificar se presenta la Fig. 2.

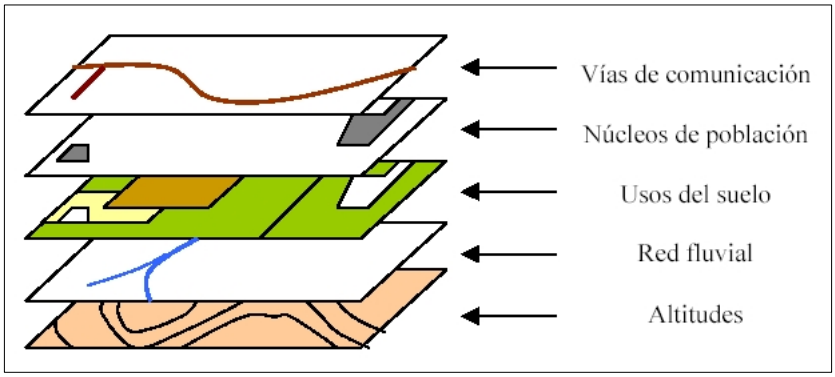


Fig. 2. Estructura de la información en un GIS. Fuente: [18]

A pesar de la heterogeneidad de la información manipulada por un GIS, existen dos modelos para simplificar y modelar la realidad. Este proceso de modelización es complejo porque depende de muchos factores, la cuestión consiste en elegir el o los elementos (entidad del sistema) mediante los cuales se quiere representar de la forma más fiel posible la realidad territorial, ver Fig. 3. Estos modelos que se emplean en la representación son el modelo vectorial y el ráster.

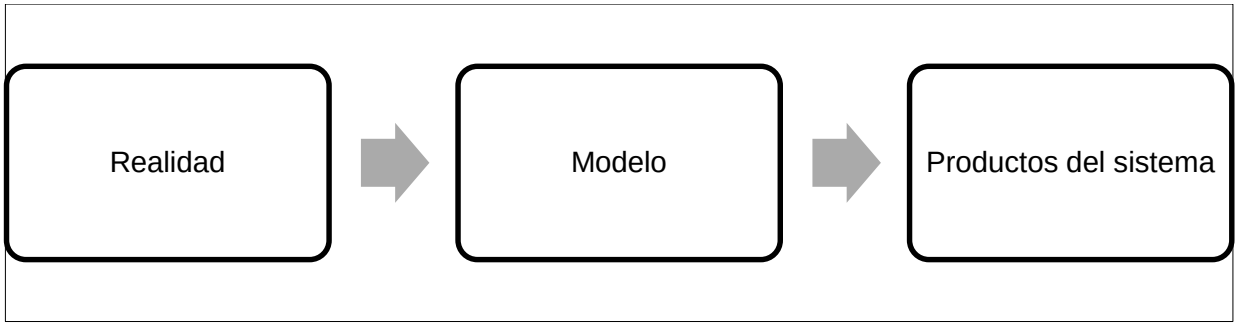


Fig. 3. Proceso de modelización en un GIS. Fuente: elaboración propia (2023)

Modelo vectorial: por medio de criterios se elige una entidad gráfica, puntos, líneas o polígonos para representar cada uno de los objetos de la realidad a tratar. Por lo general se agrupan en capas objetos del mismo "tipo grafico", (ver Fig. 4). Este modelo representa cada objeto por medio de dos entidades.

- Carácter gráfico: punto, línea o polígono.
- Carácter alfanumérico: Tabla cuyo contenido son campos que almacenan propiedades descriptivas de cada objeto de estudio.

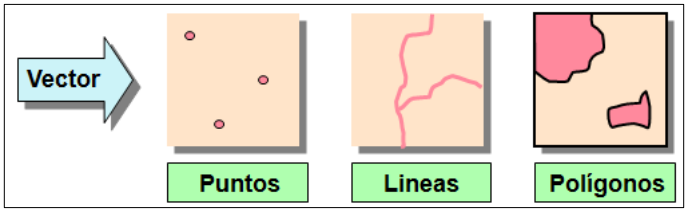


Fig. 4. Modelo de datos Vectorial. Fuente: [18].

Modelo ráster: en este caso solo existe un tipo de entidad denominada "celdilla" o "tesela". Este define unas dimensiones específicas, en función de las cuales tendremos mayor o menor nivel de detalle sobre la realidad a representar (Fig. 5).

1	1	1	1	3
1	1	1	3	3
1	1	1	3	2
1	3	2	2	2
3	3	3	2	2

Código	uso
1	Cereal
2	Pastos
3	Pinar

Fig. 5. Modelo de datos Ráster. Fuente elaboración propia (2023)

En la figura n° 5 se representa a modo de ejemplo un objeto ráster. Se divide en celdas regulares, en forma de filas y columnas, cada una de estas se recoge la información pertinente que la describe. En el caso de ejemplo se trata de "suelos", donde los distintos códigos indican suelos de pastoreo, plantación de pinos y cereales. Este modelo codifica el interior de los objetos geográficos, registrando de forma implícita la frontera de los mismos.

Seguidamente se expresan definiciones de términos que es necesario tenerlos en cuenta.

Mapa: representación gráfica, sobre una superficie plana, ya sea de una parte o el total de la superficie terrestre.

Planos: mapas realizados a una escala relativamente grande. Contienen mucho detalle y representan una pequeña parte de la superficie terrestre.

Escala: relación de semejanza que se establece entre las dimensiones reales de un objeto y su imagen sobre el mapa.

Proyección cartográfica: es el método usado para realizar una representación de una parte o el total de la superficie terrestre obtenida por transformación de una superficie básicamente esférica (como puede ser la Tierra) en una superficie plana (mapa) más fácil y cómoda de manejar.

Coordenadas geográficas: son los valores de latitud y longitud que permiten determinar la posición de un punto sobre una superficie básicamente esférica, como puede ser la Tierra.

Coordenadas cartesianas: valores de abscisas y ordenadas, X e Y respectivamente que permiten determinar la posición de un punto sobre una superficie plana.

Geodesia: es la ciencia que tiene por objeto el estudio de la forma, dimensiones y campo gravitatorio de la Tierra. Las desviaciones de la forma de la Tierra respecto de una esfera son relativamente pequeñas, sin embargo, para la elaboración de mapas resultan muy importantes afectando directamente las precisiones con la cual los datos geográficos se transfieren a los mapas.

Para el desarrollo de la geodesia se determinaron ciertos puntos denominados “vértices geodésicos” sobre la superficie terrestre.

Para situar estos puntos se han referido a una superficie irregular llamada geoide. Por tanto, estas irregularidades dificultarían la proyección de los vértices geodésicos. Por ello se acepta como superficie de referencia el “elipsoide de revolución”.

Hasta 1924 cada país utilizaba un elipsoide de referencia, así España empleaba el Struve, en Europa Central se usó Bessel, Gran Bretaña y Francia Clarke, etc.

En 1924 en la asamblea de la Unión Geofísica Internacional se acordó la utilización del elipsoide de Hayford. Posteriormente en 1964 se realizan pequeñas modificaciones.

Una vez conocidos datos u observaciones de puntos sobre la superficie terrestre, hay que trasladarlas al elipsoide escogido. Esta operación se conoce con el nombre de reducción. Y es el elipsoide la base que se utilizara en la proyección cartográfica para elaborar mapas.

Datum: en este punto coinciden las verticales al geoide y al elipsoide, ver Fig. N°6. En España se utiliza el "Datum Europeo o Datum Postman".

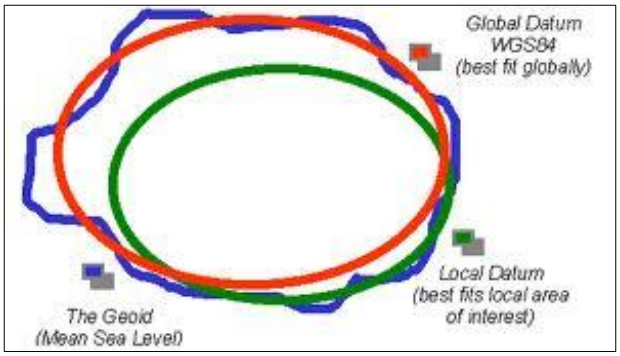


Fig. 6. Elipsoide, geoide y Datum. Fuente [18]

Sistemas de coordenadas

Se encuentra una gran complejidad al momento de situar un punto sobre la superficie terrestre. Si la concebimos como una superficie plana, es posible medir y representar los fenómenos mediante el sistema de coordenadas rectangulares planas o cartesianas.

Sin embargo, desde la antigüedad, en Grecia, se consideró la superficie del planeta como una esfera perfecta. Las coordenadas (longitud y latitud) que se miden sobre la esfera responden a una geometría esférica, las cuales se denominan coordenadas geográficas.

Con advenimiento de la Geodesia se comprobó que la superficie es un geoide lleno de irregularidades, por lo tanto, se realiza una aproximación a través de un elipsoide (elipsoide de referencia). Las coordenadas medidas sobre esta figura se llaman coordenadas geodésicas.

En los distintos procesos cartográficos y topográficos se necesitan realizar las transformaciones entre coordenadas geodésicas y planas. Y este problema de transformación lo resuelven los sistemas de Proyección. Existen una multitud de sistemas y es casi imposible clasificarlos a todos ellos. No obstante, ello, el sistema más importante utilizado en las cartas de navegación fue ideado por Gerardus Mercator en 1569.

2.5 Minería de datos espaciales

Se define a la minería de datos como el proceso de extraer conocimiento interesante, como patrones y relaciones, de un gran volumen de datos [2]. Es un campo pluridisciplinario en el cual intervienen la estadística, reconocimiento de patrones, manejo de tablas, y otros. La utilidad del conocimiento

extraído representado por los modelos o patrones hallados queda bajo el criterio subjetivo del usuario o investigador.

Podemos clasificar en dos grupos, supervisadas o predictivas y no supervisadas o descriptivas (ver Fig. 7). En caso de necesitar predecir valores, fenómenos, se puede hacer uso de técnicas predictivas. Más cuando se requiera la compresión de los datos se utilizan técnicas descriptivas.

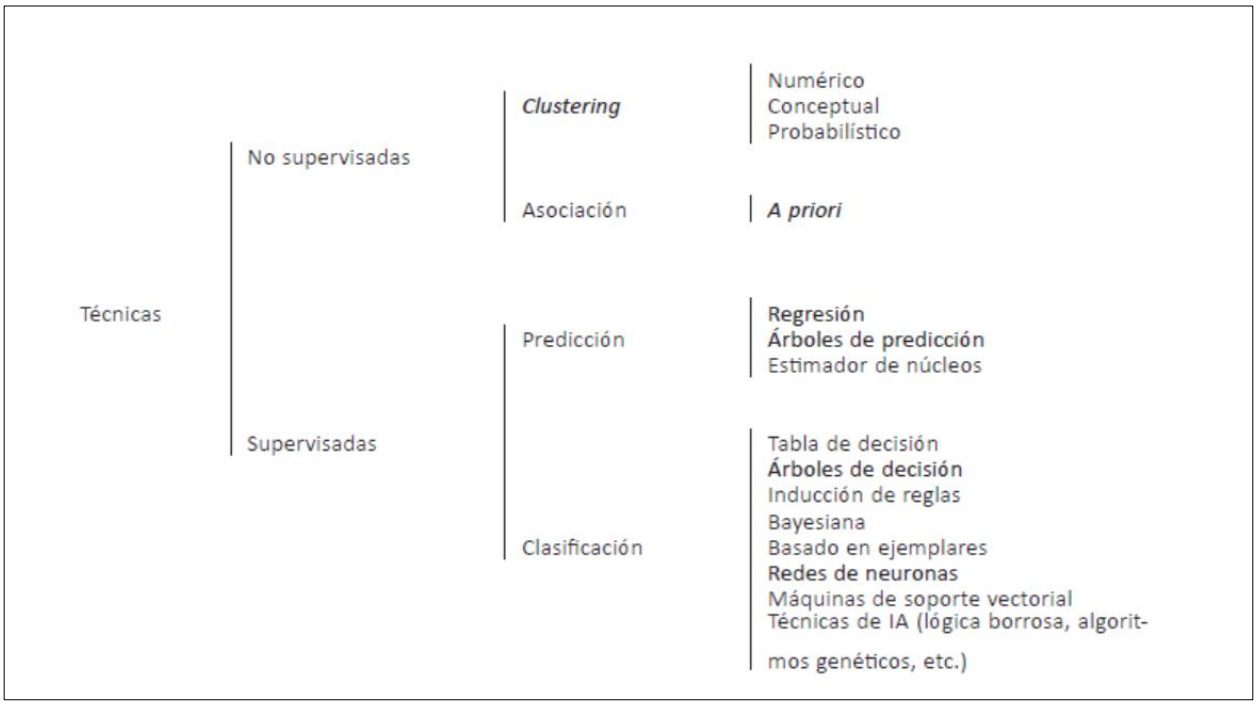


Fig. 7. Clasificación de técnicas de datamining. Fuente: elaboración propia (2023)

En el caso que esta investigación las técnicas y modelos proporcionadas por la minería tradicional no son suficiente o adecuadas debido que no tienen en cuentan el componente espacial de los datos. Los algoritmos de minería de datos espaciales que trabajan con información espacial deben cumplir con ciertas características:

- Debe poder operar en conjuntos de datos de tamaño considerable. Las bases de datos espaciales tienen la potencialidad de almacenar grandes cantidades de información [16], [17].
- Deben realizar su tarea de manera rápida.
- Deben tener en cuenta el razonamiento espacial y las técnicas existentes de optimización de búsquedas espaciales.

Taxonomía.

Presentamos en la Fig. 8, la clasificación de las técnicas SDM según el trabajo realizado por el profesor Zhang Xiang.

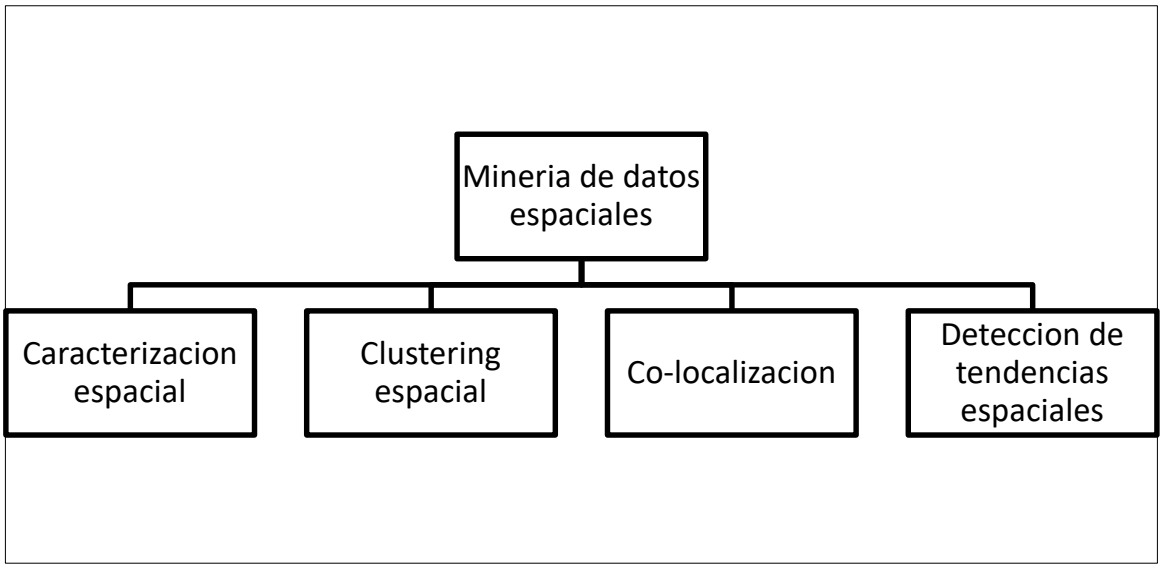


Fig. 8. Clasificación de técnicas de spatial datamining. Fuente [2]

- Co-localización: La co-localización espacial representa los conjuntos de características de tipo booleano que ocurren regularmente con gran proximidad geográfica. Ejemplos comunes en el mundo real podrían ser la presencia de estaciones de gasolina junto a vías.
- Clustering: se denomina al agrupamiento de objetos de una base de datos en subclases con significado, y es una de las técnicas más utilizadas en la minería de datos.
- Caracterización espacial de un conjunto de objetos espaciales con respecto a la base de datos que los contiene, se define como una descripción compacta de las propiedades espaciales y no espaciales que son típicas para los objetos examinados y no para el conjunto completo de objetos disponibles en la base de datos.

  Detección de tendencias espaciales: una tendencia espacial has sido definida como un cambio regular de uno o más atributos espaciales cuando se aleja de un objeto inicial dado.

Diferentes perspectivas de la minería de datos espaciales

Para comprender más en detalle este proceso de análisis de datos resulta oportuno traer aquí algunos conceptos vertidos en [2] sobre los distintos enfoques.

- Como Disciplina. SDM es una materia interdisciplinaria que coincide con la multidisciplinar filosofía del pensamiento humano y aborda adecuadamente la complejidad, incertidumbre y variedad presentes al informar datos y representar reglas. SDM es el resultado de la etapa en la que se desarrollan algunas tecnologías, como como tecnología de acceso a datos espaciales, tecnología de bases de datos espaciales, estadísticas espaciales, y sistemas de información espacial; por lo tanto, reúne los frutos de varios campos. Sus teorías y técnicas están vinculadas con la minería de datos, el conocimiento descubrimiento, sistemas de bases de datos, análisis de datos, aprendizaje automático, reconocimiento de patrones, ciencia cognitiva, inteligencia artificial, estadística matemática, red tecnología, ingeniería de software, etc.
- Como Análisis. SDM descubre reglas desconocidas y útiles a partir de enormes cantidades de datos a través de un conjunto de manipulaciones interactivas, repetitivas, asociativas y orientadas a datos. Utiliza principalmente ciertos métodos y técnicas para extraer varios patrones de conjuntos de datos espaciales. Los patrones descubiertos describen reglas existentes, o predicen una tendencia. Pueden ayudar a los usuarios a hacer pleno uso de los repositorios espaciales bajo el conjunto de diversas aplicaciones, haciendo hincapié en la implementación eficiente y oportuna respuesta a los comandos del usuario.
- Desde la Lógica. SDM es una técnica avanzada de razonamiento espacial deductivo. La inducción se utiliza para descubrir el conocimiento, mientras que La deducción se utiliza para evaluar el conocimiento descubierto. Los algoritmos de minería son una combinación de inducción y deducción.

Operación del objeto real. El modelo de datos puede ser jerárquico, de red, relacional, orientado a objetos, relacionado con objetos, semiestructurado o no estructurado. El formato de datos puede ser datos espaciales vectoriales, ráster o vector-ráster. Los repositorios pueden ser sistemas de archivos, bases de datos, data markets, almacenes de datos, etc. El contenido de datos puede incluir ubicaciones, gráficos, imágenes, textos, secuencias de vídeo o cualquier otra recopilación de datos organizada en conjunto, como datos multimedia y datos de red.

2.6 Dominio de salud publica

El desarrollo del presente trabajo se enmarca dentro de la Salud Pública, más precisamente un trabajo de investigación epidemiológico. Se pretende demostrar como amplían las posibilidades el uso de

técnicas de minerías de datos en este campo. Por encima de la utilización de herramientas estadísticas cual el caso de una investigación tradicional.

En [18] se describen los objetivos de la salud pública. Proteger, promover y restaurar la salud de las personas mediante acciones colectivas. De éstos se desprenden un conjunto de funciones esenciales encaminadas a:

- Valorar las necesidades de salud de la población, lo que significa comprender y medir los determinantes de la salud de la población en su contexto social, político y ecológico.
- Desarrollar las políticas de salud, lo que implica contribuir a la construcción de respuestas sociales para mantener, proteger y promover la salud.
- Garantizar la prestación de servicios de salud, con garantía de eficiencia, sostenibilidad, subsidiariedad, equidad y paridad en las políticas, programas y servicios para la salud.

Entre las actividades necesarias para el desarrollo de las mencionadas funciones se encuentra la gestión y, más concretamente la planificación y la programación.

2.7 Epidemiologia

La epidemiología es una de las ramas de la salud pública. En [19] se indica que la función esencial de la epidemiología es mejorar la salud de las poblaciones. Se encarga del estudio de cómo se distribuyen las enfermedades en las poblaciones y los factores que determinan o influyen en esta distribución.

Los conceptos relacionados a esta disciplina son esenciales para la base de este proyecto. Ya que estos ayudan entender la dinámica de la salud en la población, sus elementos componentes y las fuerzas que la gobiernan. Estos conocimientos son de gran importancia para poder hacer este tipo de intervenciones.

Principales aplicaciones de la Epidemiología en Salud Pública:

- Estudiar la historia natural de la enfermedad y su pronóstico en términos cuantitativos. Es decir, por qué hay enfermedades que son más graves y rápidamente mortales, mientras que otras que van acompañados de periodos más largos de tiempo.
- Identificar la etiología o la/s causa/s de una enfermedad o de estados relacionados con la salud y sus factores de riesgo relevantes. Descubrir los factores que aumenta su probabilidad.
- Descripción de la distribución, frecuencia, tendencias y velocidad de contagio de la enfermedad en las poblaciones.

Dentro del universo de las enfermedades, las que nos ocupan en este trabajo son las denominadas enfermedades crónicas, no transmisibles (ENT). El término, enfermedades no transmisibles se refiere a un grupo de enfermedades que no son causadas por un agente infeccioso, y que dan como resultado consecuencias para la salud a largo plazo y con frecuencia crean una necesidad de tratamiento y cuidados a largo plazo.

En este conjunto se encuentran o encontramos a las siguientes patologías: cánceres, enfermedades cardiovasculares, diabetes y enfermedades pulmonares crónicas. También se incluyen lesiones y trastornos de salud mental.

Los mapas de Jhon Snow. Un caso paradigmático.

Los comienzos de la epidemiología científica se vinculan con las observaciones que realizara John Snow durante las epidemias de cólera que afectaron a Londres en 1849 y 1854.

Este médico demostró que la transmisión del mal era debida a la contaminación del agua por las materias fecales de los enfermos. 700 personas fallecieron en el barrio de Soho en menos de una semana, en un área de apenas medio kilómetro de diámetro. El Dr. venía utilizando desde hacía tiempo el uso de mapas en sus artículos y exposiciones como ayuda a la hora de argumentar sus hipótesis, por lo que aprovechó para comprar un mapa del barrio y, ayudado del párroco local Henry Whitehead, ir anotando en él las muertes que se habían producido por cólera en el mes de septiembre. Para ello recurrió al trabajo de campo, visitando uno por uno los edificios del área afectada, y ayudándose de los registros del hospital de Middlesex, a donde se trasladaban muchas de las víctimas. El mapa recogía las defunciones y el resultado fue clarificador: la mayor parte de las muertes se habían producido en las proximidades de Broad Street.

Este hecho conjuga varios aspectos de los cuales se tienen en cuenta en este trabajo de investigación, ellos son, epidemiologia, el uso del dato georreferencial y la importancia de la visualización de los datos, [20].

Las fases consideradas en una investigación epidemiológicas son:

- Planteamiento del problema.
- Marco teórico y Fijación de los objetivos.
- Formulación de la hipótesis y variables a estudiar.
- Definición de la unidad de observación y de la unidad de medida.
- Determinación de la población y de la muestra.

- La recolección.
- Depuración y procesamiento de los datos.
- Presentación de los datos
- Análisis de los resultados.
- Conclusiones y publicación.

La secuencia anterior y su orden cronológico se pueden ir modificando según el diseño de investigación. No se trata de un proceso lineal, sino que es posible avanzar en forma paralela.

2.8 El Adulto Mayor afiliado al INSSJP

A nivel mundial, producto del envejecimiento poblacional, la proporción de personas mayores de 60 años ha crecido a lo largo de las últimas décadas. Se espera que, en 2030, la proporción de adultos mayores respecto a la población total sea del 16,5% y que en 2050 esta cifra ascienda a 21,5%, según UNDESA[3] (United Nations, Department of Economic and Social Affairs).

En nuestro país la cuestión no es muy diferente, las proyecciones también prevén un incremento constante de la población de individuos adultos mayores (pertenecientes a la cuarta edad, compuesta por aquellos que tienen entre 64 y 79 años y la quinta edad, que la integran aquellos mayores de 80 años).

Este fenómeno es un gran desafío para todos los actores responsables de la elaboración de políticas en general, y a los sistemas previsionales y de salud en particular, pues exige evaluar los mecanismos que protejan a las personas en esta etapa del ciclo de vida. Vale recordar, etapa de la vida en donde los ingresos se reducen y los gastos en salud se incrementan. Por lo tanto, se hace necesario garantizar el acceso a todos los bienes y servicios que impactan en la salud de esta población.

En Argentina, el 82% de las personas mayores de 64 años y el 96% de las personas mayores de 79 años, cuentan con la cobertura de salud brindada por del Instituto Nacional de Servicios Sociales para Jubilados y Pensionados (INSSJP). De esta manera, tanto por cantidad de afiliados como por el flujo de fondos administrados, el INSSJP es el organismo asegurador más importante del sistema de salud argentino. Su financiamiento proviene de aportes que realizan los empleados activos, sus empleadores, jubilados, pensionados y el tesoro nacional.

[3] Es parte de la Secretaría General de Naciones Unidas. Ayuda a países de todo el mundo a establecer sus agendas y a tomar sus decisiones con el objetivo de hacer frente a sus problemas económicos, sociales y medioambientales.

Su principal reto es garantizar que los adultos mayores afiliados accedan a los servicios que necesitan para restaurar o mantener su salud, sin tener que pasar por dificultades financieras. Sin embargo, este organismo también ofrece actividades recreativas, físicas, culturales, de voluntariado, turismo y educativas, entre otras. Estas prestaciones destinadas a las personas mayores tienen el objetivo de promover el crecimiento personal, mejorar la calidad de vida y brindar un lugar de encuentro y capacitación de los afiliados. En este contexto, adquiere relevancia disponer de información de los afiliados al INSSJP considerando aspectos demográficos, socioeconómicos, de utilización de bienes y servicios relacionados con el cuidado de la salud, de las barreras que impiden a los adultos mayores acceder a mismos.

Se prevé que los datos recopilados y procesados durante esta investigación, al igual que las herramientas empleadas sirvan como un aporte a ser considerado por los responsables de la toma de decisiones en el instituto. Siendo un insumo para idear acciones o ajustes en la administración de los recursos, con el fin de beneficiar la calidad de vida de sus afiliados.

Capítulo 3

Metodología

3.1 Metodología

Actualmente la investigación y el manejo de la información médica, están absolutamente vinculados en una gran sinergia con la informática médica y las nuevas tecnologías. En este sentido, la epidemiología y la minería de datos tienen similitudes en su enfoque de recopilación, análisis y presentación de datos para identificar patrones y tendencias. Ambas disciplinas utilizan métodos estadísticos y herramientas de software para analizar grandes conjuntos de datos y extraer conclusiones significativas.

En referencia a la metodología en el TFM, el diseño de la investigación se corresponde con un ***estudio descriptivo***. Se busca describir la frecuencia y características (variables, eventos, etc.) más importantes del problema a investigar que se detallan en la fase I (comprensión del negocio) de CRISP-DM. La unidad de análisis es una población que representa los afiliados de la obra social para adultos mayores internados en los diferentes nosocomios de la provincia de Corrientes durante el año 2019.

En lo concerniente al tipo de asignación de la exposición o variable de estudio este trabajo se enmarca como ***estudio observacional***, dado que se busca reconstruir la ocurrencia natural de fenómenos sobre una población de estudio, sin influir sobre ellos.

Desde su dimensión espacio-temporal, es un ***estudio transversal*** (se trabaja sobre los datos correspondientes al periodo 2019). Manipulando datos ***cuantitativos*** especialmente. Para más información sobre tipos de investigaciones consultar Anexo N° 2.

Por lo expuesto, desde una perspectiva metodológica para este proyecto de minería de datos se optó por adaptar CRISP-DM (Cross Industry Standard Process for Data Mining) [21]. Es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema.

I. Compresión del negocio: desde un enfoque epidemiológico, se identifica el problema de salud pública a abordar. El objetivo perseguido fue obtener la mayor cantidad de información relevante para conocer el modelo de negocio sobre el cual aplicar nuestros modelos. Es una etapa importante, que llevada a cabo en buen término puede ayudar de reducir riesgos futuros. Se componen de sub-procesos:

a) Objetivos del negocio.

b) Evaluar situación actual.

c) Objetivos del datamining.

d) Plan del proyecto.

II. Compresión de los datos: se recopilan datos de vigilancia epidemiológica, datos demográficos y otros datos relevantes para comprender la situación actual. Se estudió más en profundidad los datos disponibles, los cuales procedían de diferentes fuentes. Por lo general es la fase más extensa de un proyecto. La comprensión de datos implicó acceder a los datos y explorarlos con la ayuda de tablas y gráficos. En gran medida nos ayudó a determinar la calidad de los mismos y describir los resultados. Sub-procesos:

a) Captura de los datos.

b) Descripción de los datos.

c) Exploración de los datos.

d) Verificaciones y gestión de la calidad.

III. Prepararlos datos: es uno de los aspectos más importantes y con frecuencia que más tiempo exigen en la minería de datos. Se considera que la preparación de datos suele llevar el 50-70 % del tiempo y esfuerzo de un proyecto. Incluyo tareas como fusión de registros de datos, selección de una muestra de un subconjunto de datos, agregación de registros, derivación de nuevos atributos, clasificación de los datos para el modelado, eliminación/sustitución de valores en blanco o ausentes y división en conjuntos de datos de prueba y entrenamiento. A continuación, los sub-procesos que componen esta etapa.

a. Selección de los datos.

b. Limpieza de datos.

c. Construcción del juego de datos.

d. Integración de los datos.

IV. Modelado: en esta fase los modelados fueron ejecutados de manera iterativa. Se fue probando con distintos valores y parámetros. En ocasiones se tuvo que volver sobre la fase anterior de preparación de los datos. Se determinó aplicar los modelos en función de la naturaleza del problema, utilizando herramientas de Sistemas de Información Geográfica (GIS) para el análisis espacial de los datos geoespaciales, [22].

V. Evaluación del modelo: este es el momento en que se corrobora si los modelos son técnicamente correctos y efectivos en función de los criterios de rendimiento de minería de datos que ha definido previamente.

VI. Implementación: en esta fase el conocimiento descubierto es presentado ante las autoridades correspondientes con el objetivo de mejorar la toma de decisiones sobre las intervenciones y políticas desarrolladas en terreno.

Después de construir y evaluar los modelos en la fase de modelado, se pasa a la fase de evaluación. En esta etapa, se analizan los resultados de los modelos para determinar su eficacia y precisión. Si los modelos no cumplen con los objetivos del proyecto o no son lo suficientemente precisos, se puede volver a la fase de modelado para ajustar los modelos o probar diferentes enfoques.

Por otra parte, una vez que se han evaluado y validado los modelos, se procede a la fase de implementación, donde se despliegan los modelos en el entorno operativo del negocio. Durante esta fase, es importante monitorear el rendimiento de los modelos en producción y recopilar datos sobre su funcionamiento en el mundo real.

Volvemos al inicio, nuevamente comprensión del negocio. Después de implementar los modelos y recopilar datos de su desempeño en el negocio, se regresa a la fase de comprensión del negocio. En esta etapa, se analizan los resultados obtenidos en la implementación y se comparan con los objetivos iniciales del proyecto. Esta retroalimentación ayuda a identificar nuevas oportunidades, ajustar los modelos existentes o desarrollar nuevos enfoques. Al repetir este ciclo se logra un proceso iterativo que permite refinar continuamente los modelos de minería de datos y garantizar que estén alineados con las necesidades y objetivos del negocio, (ver Fig. 9).

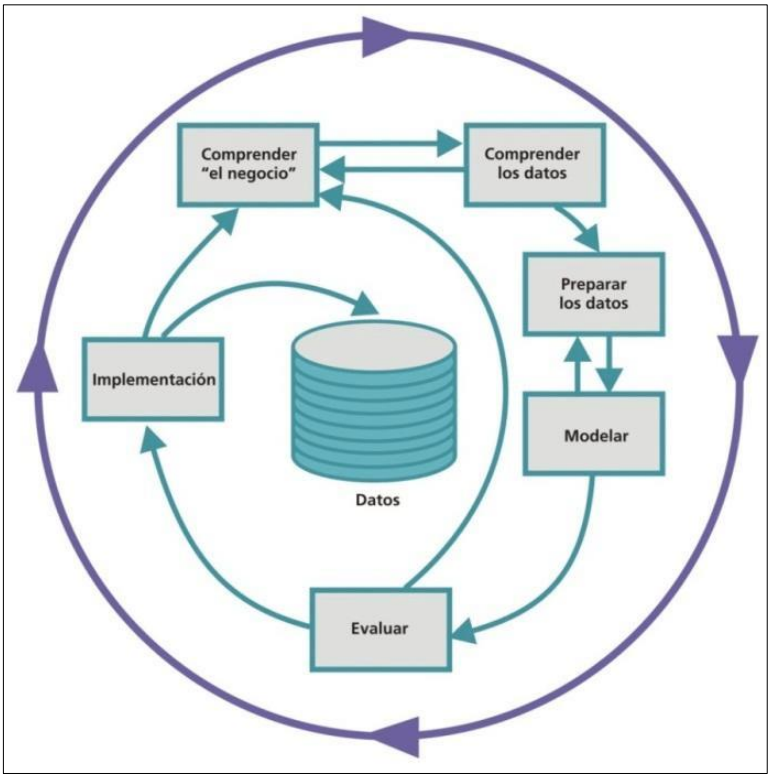


Fig. 9. Metodología CRISP – DM

Capítulo 4

Solución propuesta

Se presentan los resultados obtenidos de la aplicación de la metodología descrita en el capítulo anterior. Los hallazgos más significativos obtenidos de cada fase de la investigación, son analizados y discutidos con el fin de responder a las preguntas de investigación planteadas. Los resultados se presentan de manera organizada y relacionada con la revisión bibliográfica existente, con la intención de contribuir al conocimiento en el área de estudio.

4.1 COMPRESIÓN DEL NEGOCIO

4.1.1 Determinar los objetivos del negocio

- **Antecedentes**

El Instituto Nacional de Servicios Sociales para Jubilados y Pensionados (INSSJP) es una obra social de jubilados y pensionados, de personas mayores de 70 años sin jubilación y de excombatientes de Malvinas que opera en Argentina bajo el control estatal federal.

Fue creado en 1971 con el objetivo de brindar asistencia médica integral a las personas mayores. PAMI es la obra social más grande de Latinoamérica, con 5 millones de jubilados y sus familiares a cargo, pensionados y veteranos de Malvinas.

- **Objetivos de negocio**

Brindar una atención, alivio y contención a los afiliados lo más eficiente y eficaz posible.

4.1.2. Evaluar situación actual

- **Inventario de recursos.**

Fuente de datos: para este proyecto disponemos de un repositorio propio de datos considerable, y más allá de este volumen la importancia radica en la información que contiene. Datos sobre afiliados concerniente a sexo, edad, ubicación geográfica, patología, lugares de internación. Estos datos proceden de diferentes tipos, los encontramos en formato xls, pdf, texto plano, correos electrónicos, manuscritos.

Por otro lado, se utiliza datos de fuentes secundarias del Instituto Nacional de Estadística y Censos (INDEC).

- **Recursos humanos**

Un proyecto de esta naturaleza abarca varias disciplinas/ciencias. Y en este en particular se cuenta con el apoyo (consulta) de expertos en el tema, ver Tabla 1.

Tabla 1. Roles y funciones dentro del proyecto de datamining.

Personal	Función	¿Miembro del equipo de MD?	Obs.
Lic. Vallejos Leonardo	Analista de datos	SI	Tareas de minería de datos.
Lic. A. Kbal Lic. A. Segobia Lic. A. Sudarich	Analistas de datos sociales	NO	Nos brindan información y entendimiento sobre los programas desarrollados en territorio, sus alcances, limitaciones, características de las poblaciones, afiliados, y otros actores sociales.
Dr. J. Foguelman Dra. M. Basualdo. Dr. F. Franco. Dra. Ma. Simonit. Dra. R. Ibarra.	Analistas de datos médicos	NO	Colaboración en la clasificación de las patologías contenidas en los registros del repositorio según CIE-10 desarrollado por la Organización mundial de la salud. Además de otros conceptos o valoraciones sobre enfermedades y pacientes a tener presente en esta investigación.
Dra. S. Mariño Dra. P. Britos	Revisores académicos-científicos	NO	Guía y orientación durante el proceso de investigación y redacción del trabajo de tesis

Fuente: Elaboración propia (2024)

- **Recursos tecnológicos:**

- PC Desktop DELL Optiplex A7200.
- PC Portátil Lenovo ideadpad 3.
- Lenguaje de programación R.
- Rstudio entorno de desarrollo integrado (IDE) para el lenguaje de programación R.
- Qgis Sistema de Información Geográfica de software libre y de código abierto.
- Planilla de cálculos MS-Excel.

- **Terminología**

- UGL (Unidad de Gestión Local: Denominación para la sede central de cada provincia.
- Agencias y CAP (Centro Atención al Público): son las sedes de la obra social en el Interior de cada provincia. La diferencia se debe al número de afiliados al cual atiende la sede, siendo la agencia de mayor jerarquía que los CAP.
- Comisiones de servicios: viajes al interior de la provincia.
- INSSJP (Instituto Nacional de Servicios Sociales para Jubilados y Pensionados): nombre de la obra social, se utilizará indistintamente junto al nombre PAMI.
- Beneficiario o afiliados: son los socios de la obra social.
- Prestadores: son personas físicas como ser profesionales de la salud (odontólogo, Kinesiólogo, enfermeros, etc.) o entidades (sanatorios, hospitales, etc.) que prestan servicios a la obra social.
- Centro de jubilados: son espacios que contribuyen al fortalecimiento y bienestar integral de las personas mayores de la comunidad.
- SII – CUP: dos plataformas del instituto. SII corresponde al sistema interactivo de información y CUP clave única Pami. En ambas, se encuentran una gran cantidad de módulos (sub-sistemas informáticos), por ejemplo, el Sistema Gestión Atención que registra todas las atenciones en cada una de las sedes del organismo. Dentro del SII se encuentra el modulo “sociales” el cual contiene información relativa a afiliados, subsidios, centro de jubilados, entre otros datos más.

- **Riesgos y contingencias**

1. Tiempo: es el principal riesgo que se corre. La posibilidad de no llegar a término en el tiempo ya establecido.
2. Técnicas y algoritmos: basado en una primera aproximación, la falta de conocimientos en profundidad más su complejidad puede resultar un problema al momento de implementarlos. Plan contingencia: realizar curso de capacitación en la temática en cuestión.

3. Calidad de los datos: dada la diversidad de los datos y especialmente con los relacionados a posiciones geográficas de zonas rurales. Plan de contingencia: minimizarlos a través de consultas a referentes de la región (pueblos, parajes, colonias) como ser jefes de agencias, personal de los municipios en cuestión.
4. Interpretación de los datos: relacionado con los dos puntos anteriores. La falta de pericia en las implementaciones de algoritmos sumado a una baja calidad de datos (espaciales y atributos). Plan de contingencia: consulta tutores del proyecto. Cursos pagos de aprendizajes.
5. Privacidad de los datos: dado que este proyecto manipula datos sensibles de personas.

- **Requisitos, supuestos y limitaciones**

Por tratarse de información reservada para la gerencia, se nos ha permitido trabajar con los datos reales, pero omitiendo nombre/apellidos de los afiliados y teniendo en cuenta todas las medidas de evitar el cruzamiento de datos que permita identificarlos.

- **Análisis coste beneficio**

Para el organismo este proyecto de investigación no genera ningún costo adicional. Los datos son propiedad de la obra social, al igual que el equipo informático. Se utiliza software libre y gratuito, evitando el pago de licencias, permisos etc. Los datos de fuentes secundarias son de acceso libre también (INDEC).

El beneficio de lograrse con éxito esta investigación se concreta en información real de la situación de los afiliados. Cabe remarcar que no se hizo hasta el momento ni se dispone localmente de esta información.

- **Objetivos del datamining**

Desarrollar un proceso de minería de datos espaciales para apoyar toma de decisiones de la salud de adultos mayores. Caso de estudio de las tres enfermedades crónicas no transmisibles (ENT) más frecuentes detectadas en la ciudad de Corrientes, periodo 2019.

Objetivo N° 1: aplicación de algoritmos de agrupación (clustering) para determinar las zonas que fueron destino de las comisiones de servicios por parte del área de sociales. Se pretende realizar un análisis entre estos destinos y las características de las localidades en cuestión y el de los afiliados.

Objetivo N° 2: haciendo uso de algoritmos de agrupación analizar la distribución de los eventos de salud pertenecientes a las patologías de los afiliados de la obra social en la provincia de Corrientes.

Objetivo N° 3: aplicación de algoritmos de aprendizaje supervisado para la búsqueda de posibles asociaciones entre las variables trabajadas.

Objetivo N° 4: aplicación de algoritmos de aprendizaje supervisado para la predicción de readmisión hospitalaria de pacientes afiliados a la obra social.

- **Plan del proyecto**

Estimación cronológica del proyecto de minería de datos (ver Tabla 2).

Tabla 2. Etapas de la metodología CRISP –DM.

	Fase	Tiempo	Recursos	Riesgos
I	Comprensión del negocio	2 semanas	Analista informático y colaboradores	Disponibilidad de tiempo de personas que colaboración en la interpretación de datos.
II	Comprensión de los datos	3 semanas	Analista informático y colaboradores	Problemas de datos, Problemas tecnológicos
III	Preparación de los datos	5 semanas	Analista informático	Tiempo de análisis de base de datos. Problemas con los datos. Problemas tecnológicos.
IV	Modelado	4 semanas	Analista informático	Incapacidad para encontrar el modelo
V	Evaluación	2 semana	Analista informático	incapacidad para implementar resultados

Fuente: elaboración propia (2024)

- **Evaluación inicial de herramientas y técnicas**

  Estableceremos los criterios de selección para la herramienta datamining que se vaya a seleccionar. Elaboraremos una lista de posible software que cumplan con los criterios establecidos.

Se evaluarán también la oportunidad de uso de determinadas técnicas datamining teniendo en cuenta las necesidades del proyecto y las capacidades de la herramienta seleccionada.

4.2. COMPRESIÓN DE LOS DATOS.

En esta etapa del proceso se busca tener un primer acercamiento a los datos. Observando la naturaleza de los mismos, sus estructuras, propiedades, problemas que pueden presentar y las estrategias para eliminarlos o minimizarlos.

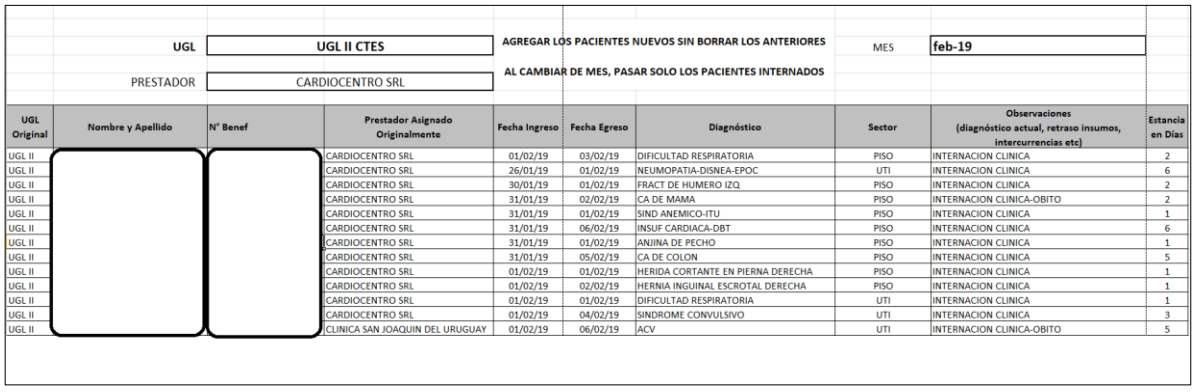
Esta fase se relaciona directamente con la calidad de los datos, lo cual representa una condición no suficiente, pero si necesaria para el excito de un proyecto de minería de datos. Se puede ampliar la información en el Anexo N° 3.

4.2.1. Recopilación de datos iniciales

En esta fase se ejecutó el proceso de carga de información, procedentes de diferentes fuentes.

Datos existentes (propiedad de la organización)

- Registro de internaciones de los afiliados en distintos centros de salud (hospitales, clínicas, sanatorios) de la provincia de Corrientes (Argentina). Estos registros pertenecen al periodo 2019. Dicha información esta almacenada en planillas de cálculos (ver Fig. 10).



UGL: UGL II CTES — AGREGAR LOS PACIENTES NUEVOS SIN BORRAR LOS ANTERIORES — MES: feb-19

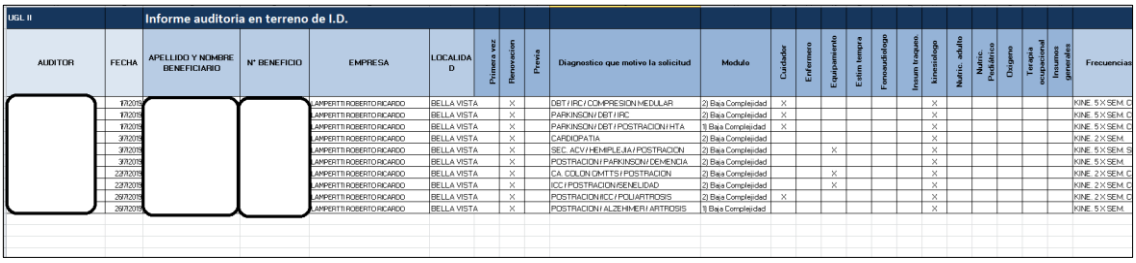
PRESTADOR: CARDIOCENTRO SRL — AL CAMBIAR DE MES, PASAR SOLO LOS PACIENTES INTERNADOS

UGL Original	Nombre y Apellido	N° Benef	Prestador Asignado Originalmente	Fecha Ingreso	Fecha Egreso	Diagnóstico	Sector	Observaciones (diagnóstico actual, retraso insumos, intercurrencias etc)	Estancia en Días
UGL II			CARDIOCENTRO SRL	01/02/19	03/02/19	DIFICULTAD RESPIRATORIA	PISO	INTERNACION CLINICA	2
UGL II			CARDIOCENTRO SRL	26/01/19	01/02/19	NEUMOPATIA-DISNEA-EPOC	UTI	INTERNACION CLINICA	6
UGL II			CARDIOCENTRO SRL	30/01/19	01/02/19	FRACT DE HUMERO IZQ	PISO	INTERNACION CLINICA	2
UGL II			CARDIOCENTRO SRL	31/01/19	02/02/19	CA DE MAMA	PISO	INTERNACION CLINICA-OBITO	2
UGL II			CARDIOCENTRO SRL	31/01/19	01/02/19	SIND ANEMICO-ITU	PISO	INTERNACION CLINICA	1
UGL II			CARDIOCENTRO SRL	31/01/19	06/02/19	INSUF CARDIACA-DBT	PISO	INTERNACION CLINICA	6
UGL II			CARDIOCENTRO SRL	31/01/19	01/02/19	ANJINA DE PECHO	PISO	INTERNACION CLINICA	1
UGL II			CARDIOCENTRO SRL	31/01/19	05/02/19	CA DE COLON	PISO	INTERNACION CLINICA	5
UGL II			CARDIOCENTRO SRL	01/02/19	01/02/19	HERIDA CORTANTE EN PIERNA DERECHA	PISO	INTERNACION CLINICA	1
UGL II			CARDIOCENTRO SRL	01/02/19	02/02/19	HERNIA INGUINAL ESCROTAL DERECHA	PISO	INTERNACION CLINICA	1
UGL II			CARDIOCENTRO SRL	01/02/19	01/02/19	DIFICULTAD RESPIRATORIA	UTI	INTERNACION CLINICA	1
UGL II			CARDIOCENTRO SRL	01/02/19	04/02/19	SINDROME CONVULSIVO	UTI	INTERNACION CLINICA	3
UGL II			CLINICA SAN JOAQUIN DEL URUGUAY	01/02/19	06/02/19	ACV	UTI	INTERNACION CLINICA-OBITO	5

Fig. 10. Registro de internaciones de los afiliados en centros de salud.

- Libro de actas de comisiones. Contiene información de las comisiones (viajes) realizadas por el personal como ser nombre de agentes, área solicitante, fecha, destino y motivo de las mismas. Se obtuvo acceso a datos que abarcan los periodos 2014 – 2019. Formato: manuscrito.
- Registro de internaciones domiciliarias. Otra fuente propia del organismo. En este caso se tiene acceso a los afiliados por cuya patología – en periodo agudo o subagudo - tienen un grado de

dependencia mayor (postrados, en proceso de rehabilitación, etc.). Algunos datos de este repositorio son nombre, patología, edad, dirección (ver Fig. 11). Formato: PDF.



UGL II		Informe auditoria en terreno de I.D.																						
AUDITOR	FECHA	APELLIDO Y NOMBRE BENEFICIARIO	N° BENEFICIO	EMPRESA	LOCALIDAD	Primera vez	Renovacion	Prorra	Diagnostico que motiva la solicitud	Modulo	Cuidador	Enfermero	Equipamiento	Estim temprana	Fonoaudiologo	Insum traqueo	Internacion domiciliaria	Nutric. adulto	Nutric. Pediátrico	Oxigeno	Terapia ocupacional	Insumos generales	Frecuencias
	1/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		DBT / IRC / COMPRESION MEDULAR	2) Baja Complejidad	X						X						KINE. 5 X SEM. C
	1/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		PARKINSON / DBT / IRC	2) Baja Complejidad	X						X						KINE. 5 X SEM. C
	1/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		PARKINSON / DBT / POSTRACION / HTA	1) Baja Complejidad	X						X						KINE. 5 X SEM. C
	3/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		CARDIOPATIA	2) Baja Complejidad							X						KINE. 2 X SEM.
	3/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		SEC. ACV / HEMIPLEJIA / POSTRACION	2) Baja Complejidad			X				X						KINE. 5 X SEM. S
	3/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		POSTRACION / PARKINSON / DEMENCIA	2) Baja Complejidad							X						KINE. 5 X SEM.
	23/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		CA. COLON C/MTTS / POSTRACION	2) Baja Complejidad			X				X						KINE. 2 X SEM. C
	23/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		ICC / POSTRACION/SENELIDAD	2) Baja Complejidad			X				X						KINE. 2 X SEM. C
	29/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		POSTRACION ICC / POLIARTROSIS	2) Baja Complejidad	X						X						KINE. 2 X SEM. C
	29/7/2019			LAMPERTI ROBERTO RICARDO	BELLA VISTA		X		POSTRACION / ALZEHIMER / ARTROSIS	1) Baja Complejidad							X						KINE. 5 X SEM.

Fig. 11. Registro de internaciones domiciliarias.

- Portal de datos abiertos de la obra social (ver Fig. 12). Dataset con información relacionada con los afiliados, nombre, rango etario y género, domicilio, prestaciones, médicos asignados.

Fig. 12. Portal de datos abiertos del instituto.

- Registro de atenciones en la agencia de General Paz e Ituzaingó, ver Fig. 13. Archivos en formato xls con información a día, fecha, tipo de solicitud, nombre afiliado. Esta información es producida por el SGA (sistema de gestión de la atención).

UO	Unidad Operativa	Area	Turno	Estado	Usuario Atención	Fecha Recepción	Inicio Atención	Fin Atención
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	113	Anulado		2/1/2019 07:18	2/1/2019 07:18	El usuario no ha finalizado
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	114	Atendido		2/1/2019 07:44	2/1/2019 07:44	2/1/2019 08:13
				Afiliado: SI		Trámite:Cambio de centro de atención médica		Obs.:solicitud Ordenes de Pr
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	115	Atendido		2/1/2019 08:13	2/1/2019 08:13	2/1/2019 08:21
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	116	Atendido		2/1/2019 09:23	2/1/2019 09:23	2/1/2019 09:29
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	117	Atendido		2/1/2019 09:31	2/1/2019 09:31	2/1/2019 09:38
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:0k
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	118	Atendido		2/1/2019 10:35	2/1/2019 10:35	2/1/2019 10:41
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	119	Atendido		2/1/2019 10:41	2/1/2019 10:41	2/1/2019 10:46
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok
20050	AGENCIA GRAL PAZ	Modelo Atención Personalizada	120	Atendido		2/1/2019 10:46	2/1/2019 10:46	2/1/2019 10:56
				Afiliado: SI		Trámite:Renovación de medicamentos sin cargo		Obs.:ok

Fig. 13. Registro de atenciones generadas desde el SGA.

Datos adquiridos

- Datos estadísticos del INDEC, (ver Fig. 14). Se procedió a la descarga en formato xls y csv de información con datos socio-económicos de la provincia de Corrientes como ser número de habitantes, índice de envejecimiento, porcentaje de alfabetización, acceso a servicios básicos (agua potable, energía eléctrica) entre otras. Esta información reviste gran importancia para conocer las características particulares de cada región de la provincia.

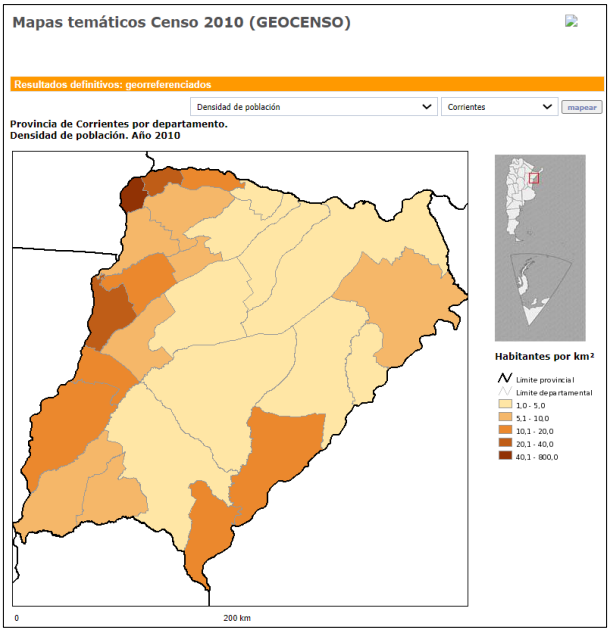


Fig. 14. Aplicativo del INDEC con datos georreferenciados.

- Listado sobre la mortalidad en la provincia de Corrientes. Planilla de cálculo con datos sobre la mortalidad Provincia de Corrientes año 2019, ver Fig. 15. Contiene información sobre el número, porcentaje y causas (según clasificación CIE-10).

	A	B	C
1	LISTA PARA TABULACIÓN DE MORTALIDAD (CODIGOS C.I.E.10)		
2	PROVINCIA DE CORRIENTES		
3	AÑO 2019		
4	TABLA Nº 8		
5			
6		2019	
7	C A U S A S	Nº	%
8	**TOTAL DEFUNCIONES GENERALES**	**7290**	**100,0**
9			
10	**A. TOTAL DE CAUSAS DEFINIDAS**	**6214**	**85,2**
11			
12	**1.- ENFERMEDADES INFECCIOSAS Y PARASITARIAS (A00-B99)**	**203**	**2,8**
13	1.1 Enfermedades infecciosas intestinales (A00-A09)	5	0,1
14	1.2 Tuberculosis, inclusive secuelas (A15-A19)	20	0,3
15	1.3 Hepatitis virales (B15-B19)	0	0,0
16	1.4 Septicemia (A41)	128	1,8
17	1.5 Enfermedad por virus de la inmunodeficiencia (B20-B24)	35	0,5
18	1.6 Tripanosomiasis (B56-B57)	2	0,0
19	1.7 Las demas enfermedades infecciosas y parasitarias	13	0,2
20			
21	**2.-TUMORES (C00-D48)**	**1501**	**20,6**
22	**2.1 Malignos**	**1355**	**18,6**
23	2.1.1 Estómago (C16)	86	1,2
24	2.1.2 Colon (C18)	178	2,4
25	2.1.3 Páncreas (C25)	81	1,1
26	2.1.4 Demás órganos digestivos y del peritoneo	144	2,0
27	2.1.5 Traquea, de los bronquios y del pulmón (C33, C34)	195	2,7
28	2.1.6 Mama (C50)	98	1,3
29	2.1.7 Utero (cuerpo,cuello y parte no especificada) (C53-C55)	119	1,6
30	2.1.8 Los demás tumores malignos	454	6,2

Fig. 15. Mortalidad en la provincia de Corrientes, 2019.

- Mapas de los barrios populares de la Rep. Argentina. Mapa interactivo con los nombres de barrios populares del país, ver Fig. 16. Fuente: Ministerio de Desarrollo Social.

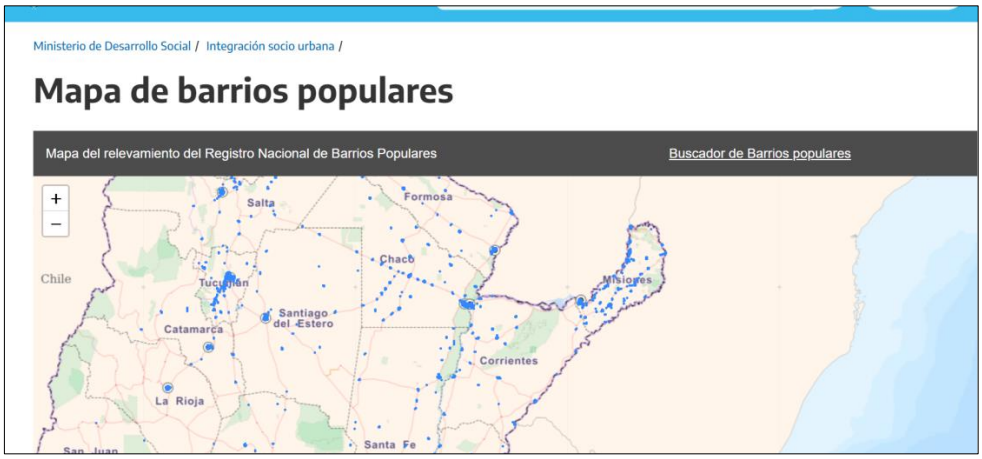


Fig. 16. Portal con mapa interactivo de barrios populares en Argentina.

- La información geoespacial. Objetos y capas de información geográfica georreferenciadas, sus atributos y sus relaciones espaciales. Fuente: Instituto Geográfico Nacional, (ver Fig. N°17). Formato: archivos Shapefile de la provincia de Corrientes.

Fig. 17. Sitio web del Instituto Nacional Geográfico.

Datos adicionales

No se realizaron tareas con el propósito de obtener información extra.

Informe de requerimientos y criterios

Realizada la primera lectura de los datos podemos determinar algunos campos que parecen relevantes para el trabajo de minería.

Para la aplicación de algoritmos de agrupación el atributo que contiene información sobre las enfermedades y edades son prometedores para ese objetivo.

En cuanto a los algoritmos predictivos, los atributos sexo, edad, patología, fecha de ingreso, estadía, son potencialmente relevantes.

Por otra parte, los atributos como nombre de afiliados, códigos de agencia pueden ser eliminados ya que no revisten importancia alguna.

En cuanto a la aplicación de algoritmos predictivos en busca del perfil de afiliados con más probabilidad de tener un reingreso a un centro de salud es posible que se requieran datos socio-económicos a partir de los sistemas de la obra social. Esta información nos puede acercar a los estratos de los cuales proceden los afiliados, y si inciden o no en el mencionado objetivo.

Como se señaló, va ser necesario fusionar datos de distintas fuentes para generar una base minable que se adapte lo mejor posible a nuestros algoritmos.

4.2.2. Descripción de los datos

En esta etapa se realizó los primeros pasos en la exploración de los datos.

Las planillas de cálculos que contienen los registros con datos de las internaciones de los afiliados en distintos centros de salud de provincia de Corrientes, se conforman con los siguientes atributos:

Registro de internaciones (ver Tabla 3)

- UGL Original: Unidad de Gestión Local, con este nombre se designan a las diferentes agencias PAMI en todo el país.
- Nombre y Apellido: nombre completo del afiliado a la obra social
- N° Benef: número identificatorio del afiliado dentro de la obra social.
- Prestador Asignado Originalmente: cada beneficiario tiene asignado un centro de salud.
- Fecha Ingreso: fecha de ingreso al centro de salud.
- Fecha Egreso: fecha de alta al centro de salud.
- Diagnóstico: patología por la cual ingreso al centro de salud.
- Sector: área dentro de centro de salud.
- Observaciones: alguna información adicional a tener en cuenta (diagnóstico actual, retraso insumos, intercurrencias etc.)
- Estancia en Días: tiempo del paciente internado.

Tabla 3.DatasetRegistro de internaciones.

Atributo	Tipo	Valores posibles	Obs.
UGL Original	string	[UGL II]	No es relevante
Nombre y Apellido	string	-	Sí. Para determinar sexo
N° Benef	string	[15000000089/00]	Sí. Para posibles cruzamiento de datos con otras fuentes.
Prestador Asignado Originalmente	string	[Clínica Madariaga]	No es relevante
Fecha Ingreso	date	[dd/mm/aaaa]	Sí. Junto a la fecha de egreso determinar estadía
Fecha Egreso	date	[dd/mm/aaaa]	Sí. Junto a la fecha de ingreso determinar estadía
Diagnóstico	string	[C.A. de pulmón, Disnea, …]	Sí.
Sector	string	[internación general – área cerrada- piso – UTI]	Sí.
Observaciones (diagnóstico actual, retraso insumos, intercurrencias, etc.)	string	-	Diagnostico actual. Insumos utilizados, etc.
Estancia días	integer		No. Se puede calcular a través de las fechas de ingreso – egreso si existen.

Fuente: elaboración propia (2024)

Libro de actas de comisiones (ver Tabla 4)

- Fecha_comision: fecha de partida de la comisión.
- Agentes: nombre del personal que realiza la comisión.
- Destino: nombre de la localidad destino del viaje.
- Área_solicitante: área/sector que realiza el pedido de la comisión.
- Motivo: motivo de la comisión.

Tabla 4. Dataset Comisiones de servicios. Periodo 2014 - 2019.

Atributo	Tipo	Valores posibles	Obs.
fecha_comision	date	[dd/mm/aaaa]	Sí. Para explorar relaciones por ejemplo entre fechas/patologías/programas desarrollados
agente	string	[nombre y apellido del personal]	No es relevante
destino	string		Sí. Mapeo de los destinos y contrastar con zonas más vulnerables, con mayor requerimiento de recursos
area_solicitante	string	[Departamento Sociales]	No. Todos los datos recabados son del Dpto. sociales
motivo	string		No es relevante. Se han obtenidos solo las comisiones realizadas por el Dpto. Sociales. Nuestro caso de interés.

Fuente: elaboración propia (2023).

Registro de internaciones domiciliarias (ver Tabla 5)

- Auditor: personal de la obra social que audita.
- Fecha: fecha en que se realiza la auditoria.
- Apellido y nombre beneficiario: nombre del afiliado.
- Empresa: empresa prestataria del servicio médico a domicilio.
- Localidad: localidad del afiliado.
- Primera vez – Renovación – Previa: indica el estado del trámite. Inicio (primera vez) o continuación de la prestación (renovación).
- Diagnostico que motivo la solicitud: patología del afiliado.
- Modulo: representado por módulos I, II, III, son las diferentes coberturas que presta la obra social.
- Cuidador, enfermería, estimulación temprana, fonoaudiólogo, insumos, traqueotomía, kinesiología, nutricionista, oxigeno, terapia ocupacional, insumos generales.
- Frecuencias: período de la prestación, 1 vez por semana, 2, 3 o 5 veces por semana.

Tabla 5. Dataset registros internaciones domiciliarias.

Atributo	Tipo	Valores posibles	Obs.
auditor	string	[nombre apellido]	No es relevante
fecha	date	[dd/mm/aaaa]	No es relevante
Nombre beneficiario	string	[nombre apellido]	Sí. Posibles cruzamiento con otras fuentes
empresa	string	[REGACE, BC SALUD, ATENDER SALUD,]	
localidad	string	[Dptos. de la prov. Ctes]	Sí. Dado que es una dato espacial
Primera vez	string	[si, no]	No es relevante
renovación	string	[si, no]	No es relevante
previa	string	[si, no]	No es relevante
Diagnostico que motivo la solicitud	string		Sí. Atributo a tener en cuenta para diferentes algoritmos
modulo	string	[I, II, III]	-
frecuencias	string	[1,2,3,5]	No es relevante

Fuente: Elaboración propia (2023)

Dataset del portal abierto de datos PAMI (ver Tabla 6)

- Nro Beneficiario: número de identificación dentro de los sistemas de la obra social.
- Apellido y Nombre: nombre del beneficiario.
- CUIT: código único de identificación tributaria del afiliado.
- Documento: número de documento de identificación del afiliado.
- Fecha Alta Padrón: fecha de alta en la obra social.
- Fecha Nacimiento: fecha de nacimiento.
- Calle: domicilio del afiliado.
- Edad: edad del afiliado.
- Grupo Etario: franja etaria a la que pertenece.
- Sexo: sexo del afiliado
- ID Agencia Beneficiario: código de identificación de las agencias dentro de la provincia.
- Localidad Beneficiario: localidad de residencia del afiliado.
- Departamento Beneficiario: departamento de residencia del afiliado.

- Médico de Cabecera: medico asignado al afiliado.

Tabla 6. Dataset padrón de afiliados. Portal datos abiertos.

Atributo	Tipo	Valores posibles	Obs. [relevancia]
nro_beneficiario	string	[15098000000/01]	Sí. Posibles cruzamiento con otras fuentes
apellido_nombre	string	-	Sí. Determinar sexo del afiliado
documento	integer		Si. Como valor para posibles búsquedas
fecha_alta_padron	date	-	No es relevante
fecha_nacimiento	date	-	Calculo de edad.
domicilio	string	-	Si reviste importancia. Tareas de georreferenciación
edad	integer	[0 - 100]	Se puede obtener como campo calculado
grupo_etareo	string		Si reviste importancia
sexo	string		
localidad_beneficiario	string	[localidades de Ctes]	Si. Por obtener la ubicación geográfica
localidad_agencia	string	[localidades de Ctes]	Si. Por obtener la ubicación geográfica
Departamento_beneficiario	string	[Dptos. de Ctes]	Si. Por obtener la ubicación geográfica
medico_cabecera	string	[nombre y apellido]	No es relevante

Fuente: Elaboración propia (2023)

SGA (Sistema Gestión de la Atención (ver Tabla 7)

En la Tabla 7 se muestra la estructura del dataset con los registros de atención al público de las agencias de General Paz e Ituzaingo.

- UO: unidad operativa, número de identificación de las diferencias agencias dentro de la provincia.

- Unidad operativa: descripción de la UO, nombre de la localidad.
- Área: Área de atención.
- Turno: fecha.
- Estado: si es atendido, o anulado el turno.
- Usuario atención: agente encargado de la atención al afiliado.
- Fecha Recepción: fecha y hora en que ingresa a la agencia. Se le entrega un número.
- Inicio atención: fecha y hora en que se sienta en el box de atención.
- Fin atención: fecha y horario de finalización de la gestión (atención).

Tabla 7. Dataset Sistema Gestión Atención. Agencias Gral Paz – Ituzaingó 2019.

Atributo	Tipo	Valores posibles	Obs. [relevancia]
UO	integer	[02000]	No es relevante
unidad_operativa	string	[General Paz, Ituzaingó]	Sí. Procedencia de afiliados de zonas más vulnerables
área	string	[mesa atención personalizada]	No es relevante
turno	integer	[1-100]	No es relevante
estado	string	[cambio médico, renov medicamentos, etc.]	Si. Tipo de prestación demandada
usuario_atencion	string		No es relevante
fecha_recepcion	date	[dd/mm/aaaa]	No es relevante
inicio_atencion	date	[dd/mm/aaaa:hs:ms]	No es relevante
fin_atencion	date	[dd/mm/aaaa:hs:ms]	No es relevante

Fuente: Elaboración propia (2023).

Datos estadísticos del INDEC (ver Anexo 4)

Desde el portal del INDEC se obtuvo acceso a mapas temáticos. Estos datos georreferenciados – los referentes a la provincia de Corrientes - fueron unificados en un archivo xls.

Id: número de identificación.

Departamento: nombre del Dpto.

Densidad_poblacion: valor en porcentaje por Dpto.

Población_total: total por Dpto.

Pobla_+65: población mayor a 65 años por Dpto.

mujeres_+65: total de mujeres mayores a 65 años por Dpto.

varones_+65: total de varones mayores a 65 años por Dpto.

inidice_enveje: porcentaje de índice de envejecimiento por Dpto.

inidice_enveje_mujeres: porcentaje de índice de envejecimiento por Dpto. de mujeres.

inidice_enveje_varones: porcentaje de índice de envejecimiento por Dpto. varones.

Analfabetismo: porcentaje de analfabetos por Dpto.

Hogares_sin_agua_vivienda: porcentaje de hogares sin agua potable dentro de la vivienda.

Uso_computadora: porcentaje de hogares con equipos informáticos dentro del hogar.

Listado sobre la mortalidad en la provincia de Corrientes 2019 – 2020. Según clasificación CIE-10. Contiene porcentajes y cantidades.

- Enfermedades infecciosas y parasitarias (A00-B99)
- Tumores (C00-D48)
- Enf. Endocrinas, nutricionales y metabólicas (E00-E90)
- Enf. Endocrinas, nutricionales y metabólicas (E00-E90)
- Enf. Del sistema nervioso (G00-G99)
- Trastornos mentales y del comportamiento (F01-F99)
- Enfermedades del sistema circulatorio (I00-I99)

Mapas interactivos del Ministerio de Desarrollo Social. ReNaBap. Registro Nacional de barrios populares de la Rep. Argentina.

Acceso al portal para verificar visualmente domicilios de pacientes y ambiente en el cual se encuentra.

Instituto Geográfico Nacional(IGN)

Se descargó en formato vectorial una serie de capas de información. Esta información es consistente con el Catálogo de Objetos Geográficos del Organismo y forma parte de la Base de Datos Geoespacial Institucional. Todos los datos se encuentran expresados en coordenadas geodésicas, utilizando el Sistema de Referencia WGS 84 y el Marco de Referencia POSGAR 07 (Código EPSG: 4326)

Referencia espacial

Sistema de coordenadas Tipo Geográfica: GCS_WGS_1984. Identificador conocido: 4326.

Tipo de geometría: Polígono.

Atributos

- Entidad: Código que hace referencia al objeto geográfico
- Objeto: Tipo de objeto geográfico.
- Fna: Nombre completo que se utiliza para designar un objeto en un mapa o carta. Está formado por el término genérico y el término específico. Ejemplo: río Mendoza.
- Gna: Parte del nombre geográfico que indica el tipo de objeto que identifica. Ejemplo: río, monte, glaciar, establecimiento.
- Nam: Parte de un nombre geográfico que acompaña al término genérico y que identifica e individualiza un objeto geográfico determinado. Ejemplo: Paraná en río Paraná; Upsala en glaciar Upsala; Las Marías en establecimiento Las Marías; Esperanza en el caso de bahía Esperanza.

Informe de atributos y volúmenes

Formatos

La mayoría de los datos se obtuvieron por medio de archivos en formato xls. Los reportes exportados desde los sistemas propios del organismo se encontraban tabulados en ese formato. Al igual que la información manuscrita en actas con los detalles de viajes de comisiones de servicio.

Los datos de los mapas temáticos del INDEC con diferentes indicadores socio-económicos fueron tabulados con MS-EXCEL.

En lo que refiere al almacenamiento de información espacial, la misma se obtuvo a través de archivos vectoriales, más específicamente los Shapefile desde el sitio web de Instituto Geográfico Nacional (IGN). Estos archivos almacenan información geométrica de diferentes elementos de una capa, por ejemplo, puntos, líneas o polígonos y cada vértice lleva implícitas sus coordenadas en un sistema de referencia concreto. Dimensiones

El dataset como mayor volumetría es el padrón de afiliados, con poco más de 100.000 registros y 12 atributos.

Luego se cuenta con planilla de afiliados internados en diferentes centros de salud de la provincia de Corrientes, organizados por meses correspondientes al periodo 2019.

Son aproximadamente entre 15 o 20 planillas por mes. Cada planilla contiene 11 atributos y el número de observaciones (registros) oscila entre 10 y 50, dependiendo si pertenece una localidad con mayor o menor densidad poblacional. Algo similar en cuanto a las características ocurre con las planillas de internaciones domiciliarias (afiliados postrados).

Los reportes extraídos por el SGA cuentan con 9 atributos y 50 registros, cada reporte.

La conversión digital de las comisiones de servicios produjo una planilla con 360 registros sobre los viajes realizados durante el periodo 2014 – 2019.

Calidad de los datos

Encontramos atributos relevantes por nuestros objetivos de minería de datos. Tales como dirección, sexo, patología y edad revisten de importancia para la aplicación de algoritmos de agrupación.

Otros atributos aportados por los dataset del Ministerio de Salud Pública, INDEC nos permiten entender la realidad sobre la cual estamos operando, conocer más en detalle el territorio.

Conocer el número de afiliado y/o documento único nos permite cruzar datos con otras fuentes, por ejemplo, conocer su dirección física. Algo fundamental para el mapeado de enfermedades.

En esta fase no se realizó cálculos básicos estadísticos sobre ningún campo. Los cuales se tiene previsto realizar en las siguientes etapas y en colaboración de analistas de datos poder encontrar los primeros datos que nos aporten conocimientos.

4.2.3 Exploración de los datos

En esta fase se inició el análisis exploratorio de datos (EDA). La misma consistió por un lado analizar el espacio geográfico en cuestión (la provincia de Corrientes) para ello se consideraron los datos del INDEC. Se produjeron mapas temáticos correspondientes a diferentes índices como ser, índice de envejecimiento, de personas mayores a 65 años, nivel de alfabetización, y totales de afiliados de la obra social por departamento, etc.

Y la segunda línea desarrollada en esta etapa es la aplicación de procedimientos estadísticos sobre datos referentes a internaciones, patologías de los afiliados, lugar de origen, entre otros datos considerados. Todo ello, para comprender la naturaleza intima de los datos con los que se está trabajando.

Como se mencionó en el inicio de este trabajo, la salud no es innata, se construye. Si bien las personas nacen con una capacidad genética, esto representa solamente el 30% de su influencia. El 70% restante depende de impactos positivos y negativos que tienen su comunidad, contextos socio-económicos, circunstancias materiales (calidad de la vivienda, posibilidad de consumo, etc.) su entorno más cercano, factores conductuales y biológicos, sistema de salud, entre otros.

Por ello se recabaron y analizaron datos con el fin de crear mapas que ilustren factores que pueden afectar a la salud de los individuos.

El primer mapa corresponde a la distribución y localización de las personas mayores de edad (mayor a 65 años) sujetos objetivos de este trabajo. En la figura N°18 se puede detectar a simple vista los departamentos con mayores índices de población mayor a 65 años: Murucuyá 9,7%, General Paz y Berón de Estrada con 10,2%, en la zona norte de la provincial. Mientras que del margen del rio Uruguay encontramos al Dpto. de Alvear con 9,6% y más hacia el sur a Sauce con 10,3%.

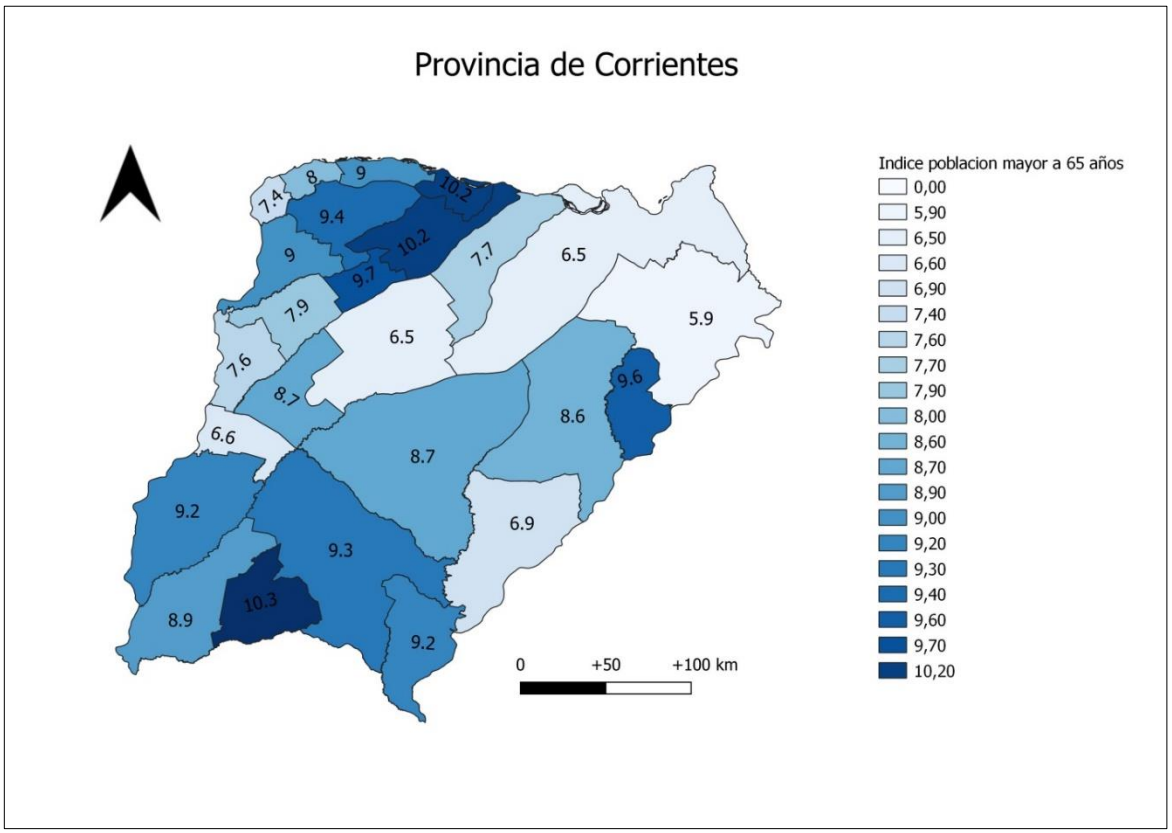


Fig. 18. Índice de personas mayores a 65 años.

Fuente: elaboración propia a partir de los datos extraídos del INDEC (2024)

Los niños y jóvenes que reciben una educación de buena calidad tienen más probabilidades de gozar de buena salud y, del mismo modo, los que gozan de buena salud están en mejores condiciones de aprender y completar su educación. Un artículo del sitio web de la UNESCO declara que las estadísticas demuestran por ejemplo que un mayor nivel de instrucción entre las madres mejora la nutrición de los niños y las tasas de vacunación, al tiempo que reduce las muertes infantiles que pueden evitarse, la mortalidad materna y las infecciones por VIH. Todo lo anterior demuestra la asociación entre educación y salud. En la Fig. 19 los dptos. de Berón de Estrada, General Paz tiene los niveles más altos de analfabetización en la región Noroeste de la provincia (10,2). Mientras que, en la zona sur, el dpto. de Sauce cuenta con el mayor porcentaje (10,3). Consultas realizadas a trabajadores sociales y autoridades sobre este punto, nos indicaban que la población particularmente en zonas rurales se manejaba con el idioma guaraní

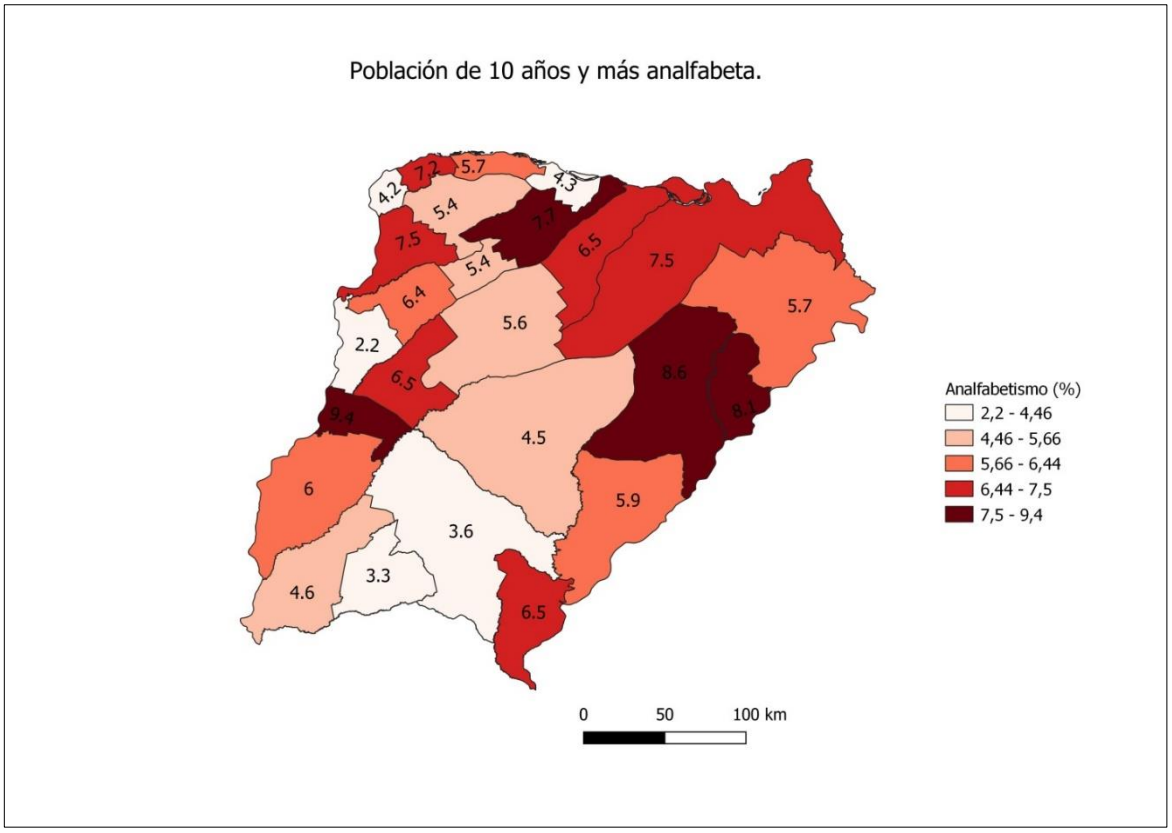


Fig. 19.Índice de analfabetismo en Ctes.

Fuente: elaboración propia a partir de los datos extraídos del INDEC (2024)

Según la OPS (Organización Panamericana de la Salud) el acceso a servicios inadecuados de agua, saneamiento e higiene tienen consecuencias importantes para asegurar la salud o la enfermedad de

una población. Las enfermedades causadas por el uso del agua están relacionadas con la presencia de microorganismos y sustancias químicas presentes en el agua de consumo. Entre ellas se puede citar la malnutrición, las enfermedades desatendidas, la diarrea, las intoxicaciones, entre otras.
Por ello se mapeo esta incide que tiene implicancias en la salud de la población. El resultado obtenido se puede ver en la Fig. 20.

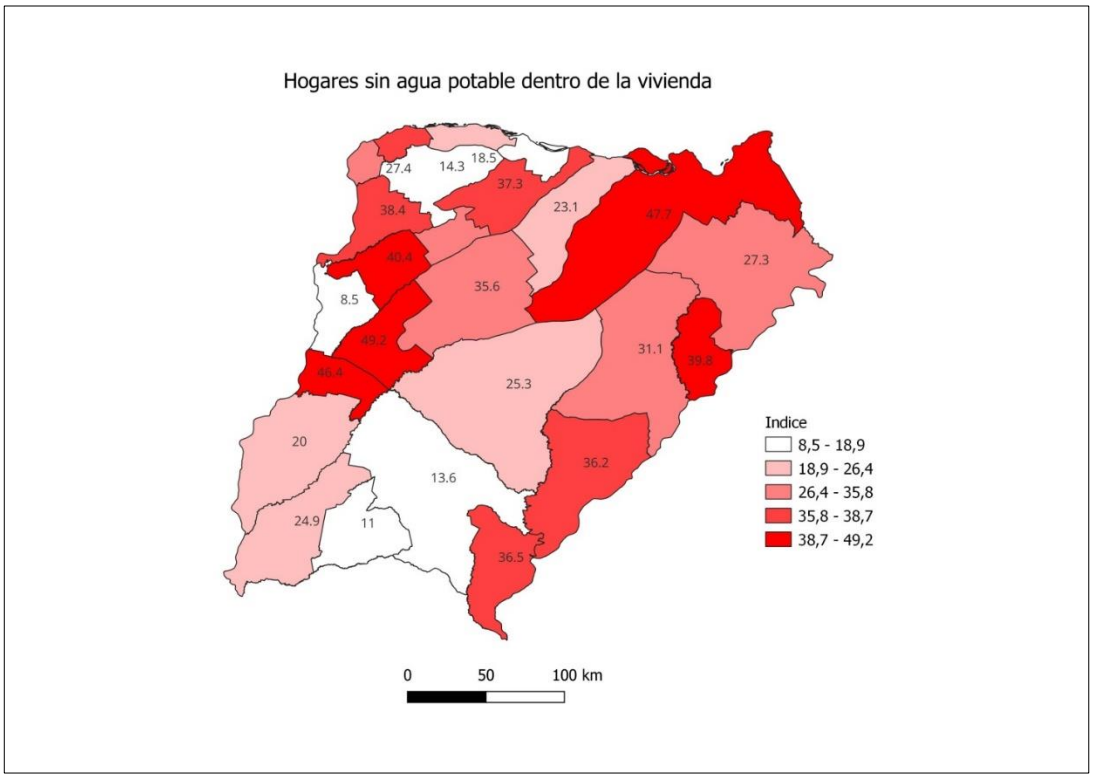


Fig. 20. Hogares sin agua potable dentro de la vivienda.
Fuente: elaboración propia a partir de los datos extraídos del INDEC (2024)

Personas de 65 años y más por cada 100 potencialmente activas**.**
Las causas de dependencia en mayores son múltiples y varían de forma notable según los casos. Llamamos dependencia al estado de carácter permanente en el que una persona mayor se encuentra ya sea por edad, enfermedad o discapacidad, impidiéndole realizar de manera autónoma actividades básicas de la vida diaria y, por tanto, necesitando ayuda en su día a día.
Son muchos y variados los factores que causan o lleva a esta situación de dependencia. La fragilidad física, los problemas de movilidad y las enfermedades. Las enfermedades asociadas a la edad hacen necesario un aumento en el consumo de medicación, lo que puede provocar efectos secundarios que contribuyen la dependencia. Limitaciones sensoriales, los problemas de visión y la sordera en la

vejez son ejemplos que influyen en gran medida en la discapacidad de las personas mayores. En Fig. 21se ilustra en porcentajes esta cuestión.

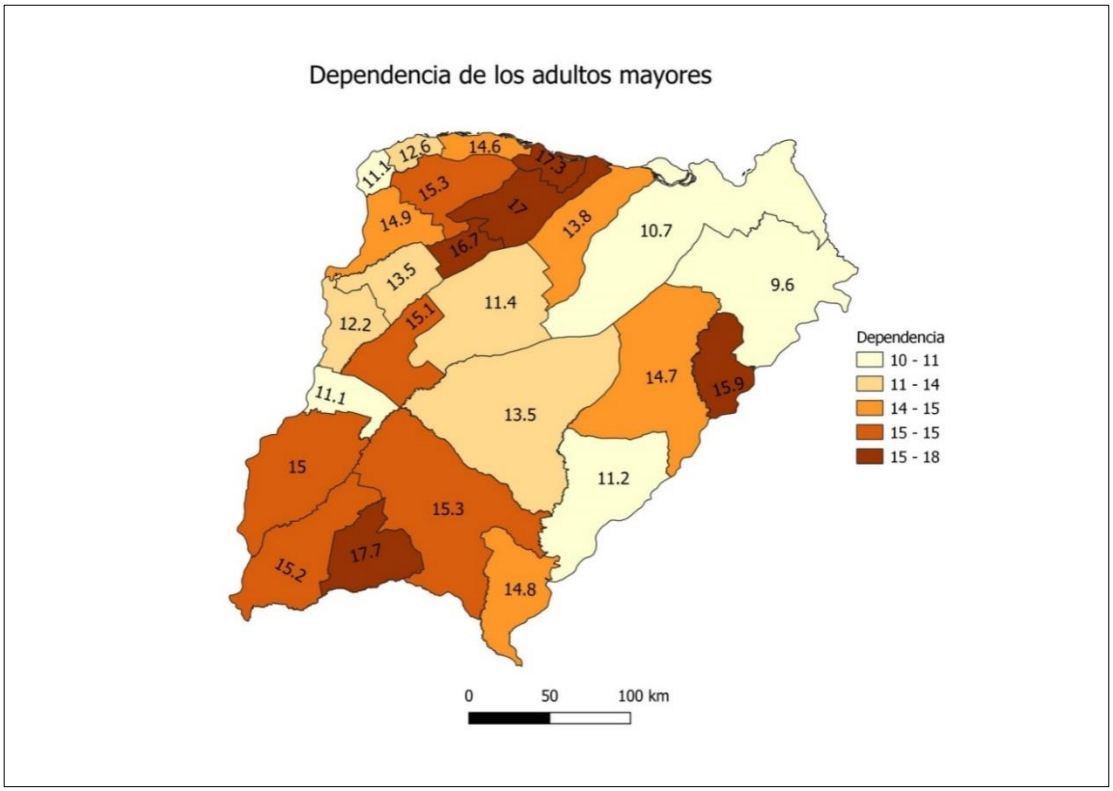


Fig. 21. Dependencia de los adultos mayores.

Fuente: elaboración propia a partir de los datos extraídos del INDEC (2024)

Finalmente se encuentra en la Fig. 22, el total de afiliados de la obra social discriminada por departamentos de Corrientes en el año 2019.

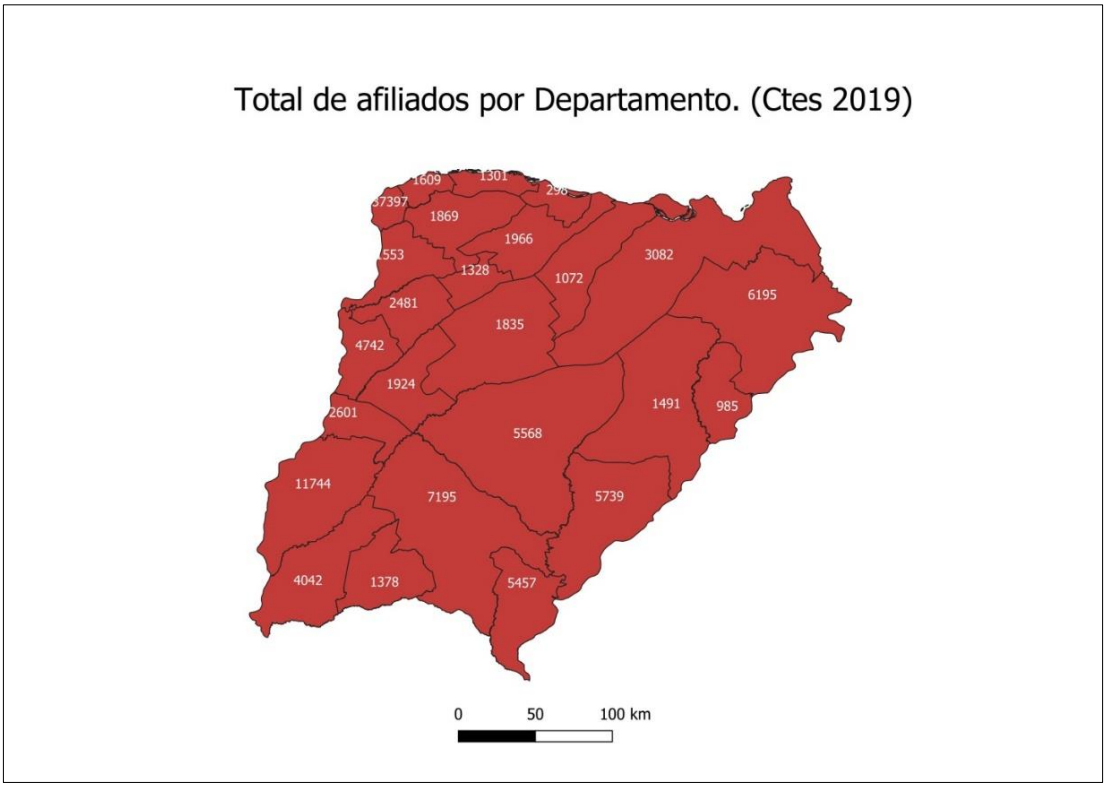


Fig. 22. Total de afiliados de la obra social en Ctes.

Fuente: elaboración propia (2024)

Datos cualitativos. Los diagnósticos.

Si bien este proyecto trabaja fundamentalmente con datos cuantitativos, el análisis de la variable diagnostico - la cual contiene las patologías registradas por cada centro de salud – sin ser sometida a un proceso de normalización que permita denominar de manera uniforme a cada una de las enfermedades, puede ser una fuente primaria de hallazgos. Los cuales nos den una primera lectura.

Para tal caso se pensó en una herramienta de visualización “Las nubes de palabras” o WordClouds en inglés. Se utiliza para representar visualmente las palabras más frecuentes en un conjunto de texto. Es importante resaltar que se trata de una primera aproximación, ya que durante el proceso de carga encontramos un sin números de inconsistencias propia de esta fase (ej., cáncer puede ser registrada como CA, C.A. tumor, carcinoma, neoplasia, etc.). A pesar de ello esta herramienta permite rápidamente encontrar lo que en marketing se suele llamar “Pain Points” o puntos de dolor. Que, aplicado a la investigación, serían los puntos de dolor de los afiliados internados. Los valores del atributo patología conforma un conjunto valioso de datos a ser estudiados. En la figura N°23 se

presenta la nube de palabra obtenida a través del lenguaje de R. En el Anexo N° 5 se presenta el código correspondiente.

Fig. 23. Nubes de palabras a partir de la información de patologías registradas por los afiliados.

Fuente: elaboración propia (2024)

Datos contenidos en el repositorio (Internaciones).

En la Tabla 8 se recogen los datos sobre las patologías de los afiliados en los diferentes nosocomios de la provincia de Corrientes. Se utilizó el estándar CIE-10 para la categorización de las mismas.CIE-10 es el estándar internacional para la generación de estadísticas, tanto de morbilidad como mortalidad publicada por la Organización Mundial de la Salud (OMS). Convierte términos de diagnósticos de palabras en códigos alfanuméricos que permiten su fácil almacenamiento y posterior recuperación para análisis de la información. Permite realizar análisis de la situación de salud en grupos de población, por ejemplo.

El dataset “internaciones 2019” contiene 10.681 registros correspondientes a los 12 meses del año 2019. Se registró una sola entrada por afiliado en cada uno de los meses relevados, se tomaron los códigos más generales para su clasificación.

De las 10.547 patologías validas registradas - se descontaron 135 registros por no contener diagnóstico - el 66% aproximadamente comprenden las llamadas “Enfermedades No

Transmisibles[4]" (ENT). No son causadas por un agente infeccioso y dan como resultado consecuencias para la salud a largo plazo.

Tabla 8. Causas de patologías registradas y sus frecuencias. Periodo 2019

ID_CIE-10	CIE-10 Descripción	Frecuencia
R00-R99	Síntomas, signos y hallazgos anormales clínicos	1712
I00-I99	Enfermedades del aparato circulatorio	1436
K00-K93	Enfermedades del aparato digestivo	1365
J00-J99	Enfermedades del aparato respiratorio	1243
C00-D48	Neoplasias	1183
E00-E90	Enfermedades endocronicas, nutricionales y metabólicas	902
S00-T98	Traumatismos, envenenamientos y algunas otras consecuencias	685
Na	Na	568
N00-N99	Enfermedades del aparato genitourinario	562
M00-M99	Enfermedades del sist. Osteomuscular y del tejido conectivo	280
A00-B99	Ciertas enfermedades infecciosas y parasitarias	249
L00-L99	Enfermedades de la piel y el tejido subcutáneo	180
F00-F99	Trastornos mentales y del comportamiento	168
O00-O99	Embarazo, parto y puerperio	67
G00-G99	Enfermedades del sistema nervioso	28
Z00-Z99	Fact. Que influyen en el estado de salud y c/ serv. De salud	23
D50-D89	Enf. de la sangre y de los órganos hematopoyéticos y otros.	22
U00-U99	Códigos para situaciones especiales	3
Q00-Q99	Malformaciones congénitas deformidades y anomalías	3
P00-P96	Afecciones originadas en el periodo perinatal	1
H00-H59	Enfermedades del ojo y sus anexos	1
	TOTAL	***10681***

Fuente: elaboración propia (2024)

En el gráfico de la Fig. 24 se visualiza la proporción que representan las ENT sobre el total de los registros de internaciones. Mientras que en la Fig. 25 de desglosa en forma gráfica las patologías referentes a cardiovasculares, cáncer, diabetes y pulmonares.

[4] Las enfermedades no transmisibles matan a 41 millones de personas cada año, lo que equivale al 71% de las muertes producidas en el mundo. Fuente: https//www.paho.org/es/temas/enfermedades-no-transmisibles.

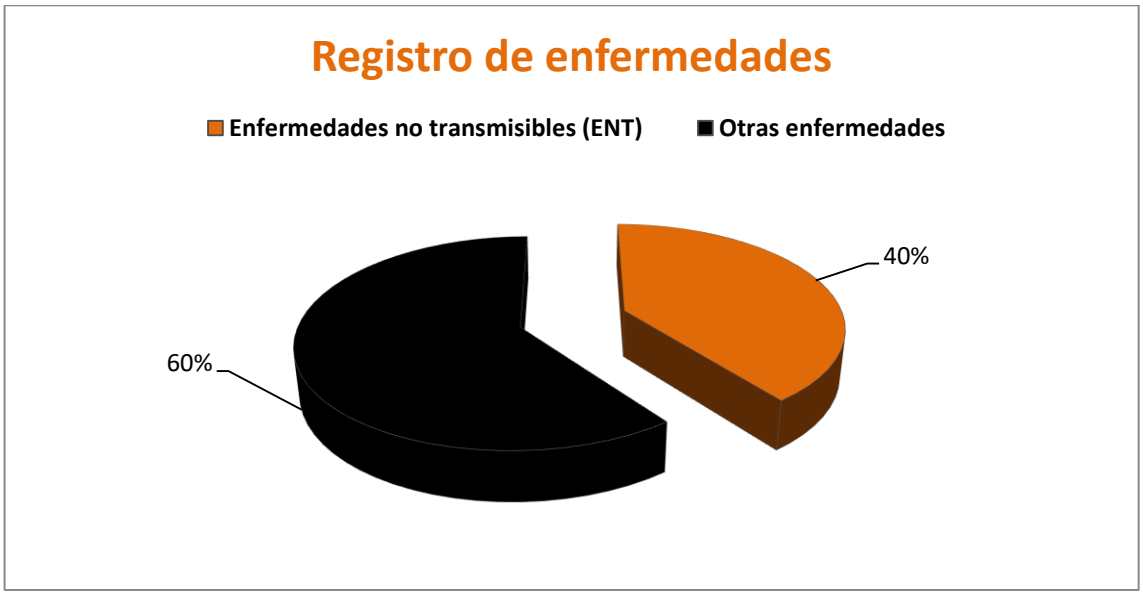


Fig. 24.ENF registradas en el repositorio.

Fuente: elaboración propia (2024)

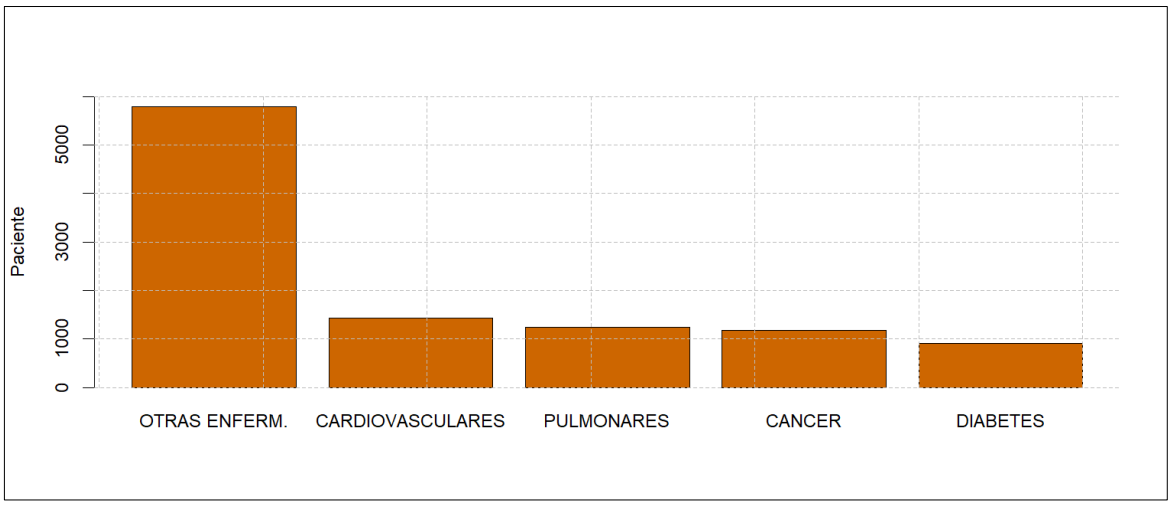


Fig. 25. Desglose de las patologías correspondientes a ENF.

Fuente: elaboración propia (2024)

En lo que refiere a infraestructura, específicamente al número y localización de centros de salud dentro de la provincia de Corrientes se elaboraron 2 mapas. Las Fig. 26 y 27 presentan dicha información.

Con respecto a las clínicas, hospitales y sanatorios de los cuales se tienen registros de las internaciones de los afiliados, son presentados en la Fig. 28 y en forma tabular en la Tabla 9.

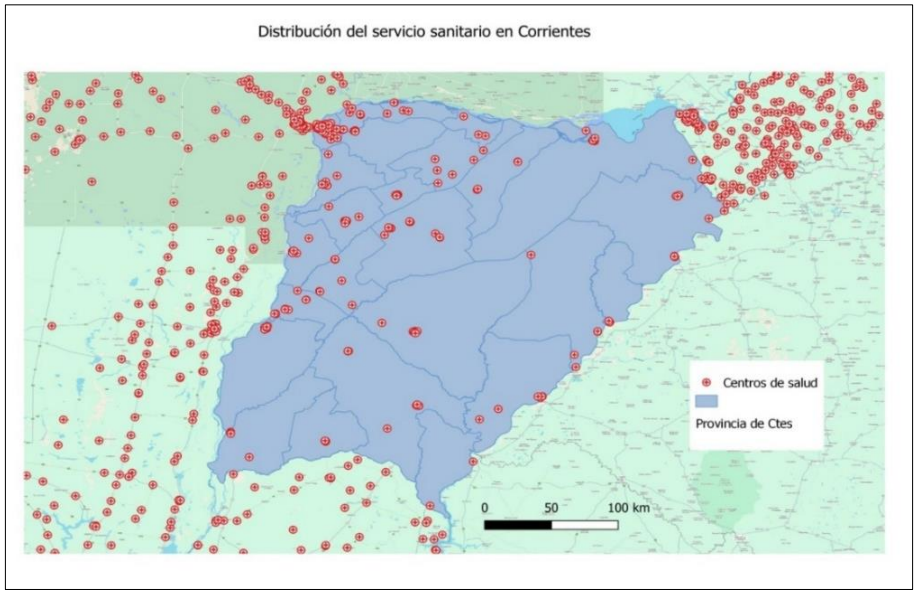


Fig. 26.Localización de centros de salud en Corrientes.

Fuente: elaboración propia (2023)

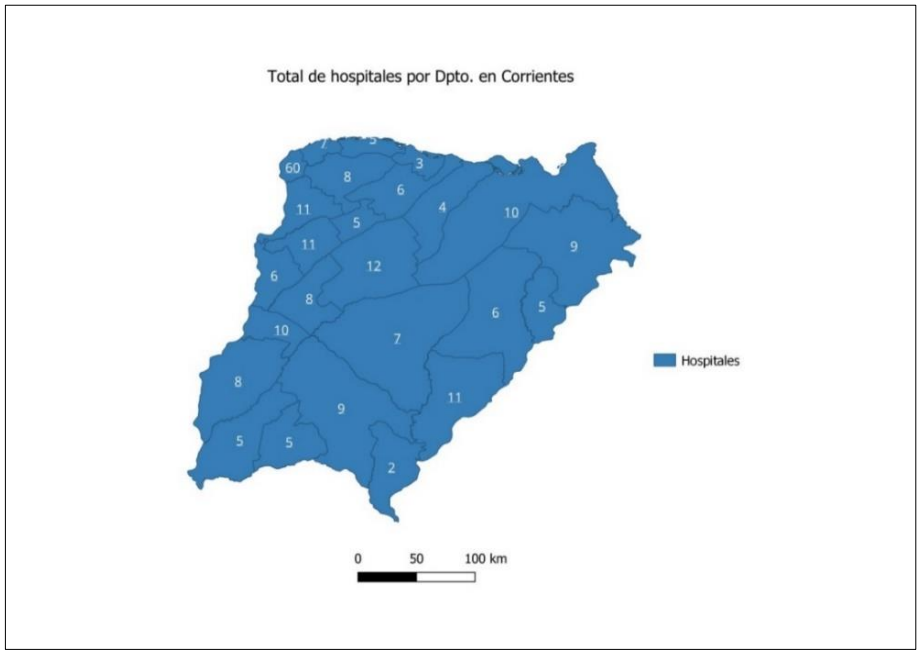


Fig. 27. Mapa con el número de centros de salud por dpto.

Tabla 9. Internaciones registradas durante 2019, (Prov. Ctes).

Localidades	Centro de salud	Nro. de internaciones
Alvear	1	7
Bella Vista	2	796
Corrientes	12	7174
Curuzú Cuatia	1	202
Esquina	2	131
Goya	3	281
Mercedes	2	422
Monte Caseros	2	902
Paso de los Libres	1	719
San Roque	1	2
Sauce	1	39
Virasoro	1	6

Fuente: elaboración propia (2024)

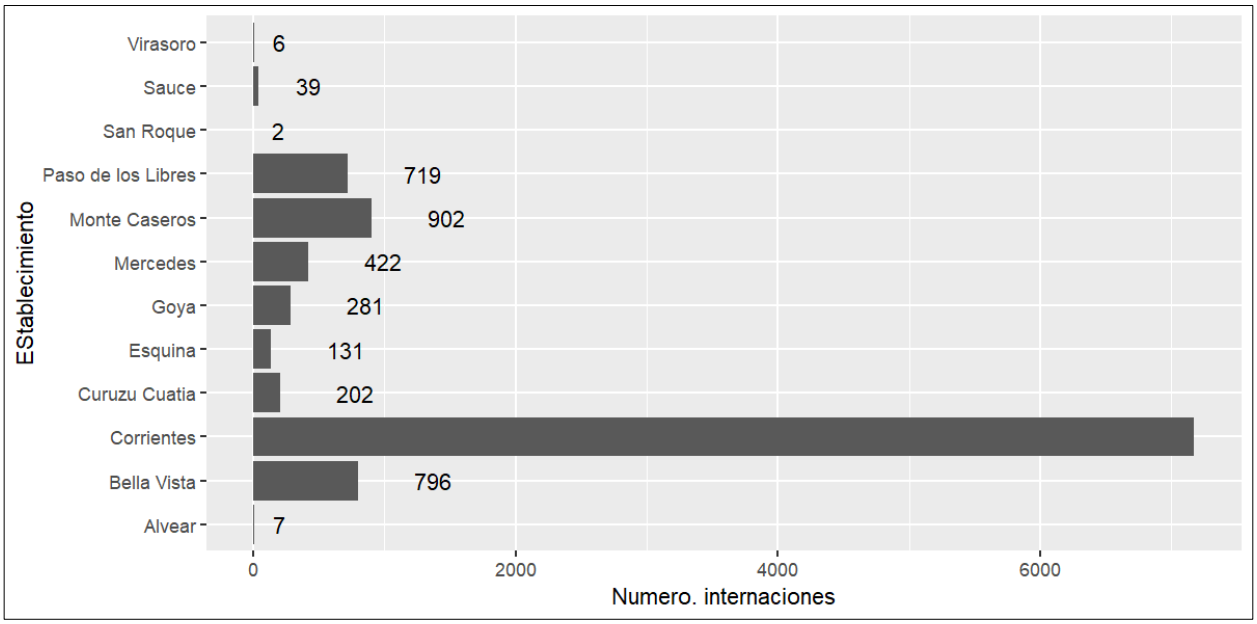


Fig. 28. Número de internaciones por centros de salud. Fuente: elaboración propia (2024)

MEDIDAS DE TENDENCIA CENTRAL

Se calcularon la media, la mediana y la moda del atributo edad de los pacientes, lo que proporciona información sobre la edad promedio, la edad central y la edad más común en la muestra seleccionada

(dataset internaciones2019). Los estadísticos calculados, media, mediana y moda fueron calculados con lenguaje R.

A continuación, se presenta el dataset con los valores de edad.

```
> sort(ds$edad)

  [1]  2  6 15 16 19 19 22 29 34 42 50 52 53 53 57 57 57 57 58 60 60 60 60 61 61 61 62 62 62 63 63
 [32] 63 63 63 63 63 63 63 64 64 64 64 64 64 64 65 65 65 65 65 65 65 65 65 65 65 65 66 66 66 66 66
 [63] 66 66 66 66 66 66 67 67 67 67 67 67 67 67 67 67 67 68 68 68 68 68 68 68 68 68 68 68 68 68 68
 [94] 68 68 68 68 68 69 69 69 69 69 69 69 69 69 69 69 69 69 69 69 69 69 69 70 70 70 70 70 70 70 70
[125] 70 70 70 70 70 70 70 71 71 71 71 71 71 71 71 71 71 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
[156] 73 73 73 73 73 73 73 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74 74 75 75 75 75 75 75 75
[187] 75 75 75 75 75 75 75 75 75 75 75 76 76 76 76 76 76 76 76 76 76 76 76 76 76 76 76 76 76 77
[218] 77 77 77 77 77 77 77 77 77 77 77 77 78 78 78 78 78 78 78 78 78 78 78 78 78 78 78 79 79 79 79
[249] 79 79 79 79 79 79 79 79 79 79 79 79 79 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80
[280] 81 81 81 81 81 81 81 81 81 81 81 81 82 82 82 82 82 82 82 82 82 82 82 82 83 83 83 83 83 83 83
[311] 84 84 84 84 84 84 84 85 85 85 85 85 85 85 85 85 86 86 86 86 86 87 87 87 87 87 87 87 87 87 88
[342] 88 88 88 88 88 88 88 88 89 89 89 89 89 90 90 90 90 90 90 91 91 91 92 92 93 93 93 94 95 95 98
```

```
>mean(ds$edad)
[1] 73.58333
> median(ds$edad)
[1] 75
>mfv(ds$edad)
[1] 68 75
>max(ds$edad)
[1] 98
> min(ds$edad)
[1] 2
```

Se encontró una edad promedio de 74 años, una mediana de 75, mientras que los valores extremos son de 98 y 2 años (máximos y mínimos). Los valores más frecuentes son de 68 y 75 años. La edad mínima (2 años) se entiende porque los hijos y nietos de los afiliados pueden estar a cargo de los titulares. Por lo tanto, el dataset cuenta con valores por debajo de los 18 años también. En la Fig. 29 se observa la distribución de las edades de los pacientes internados.

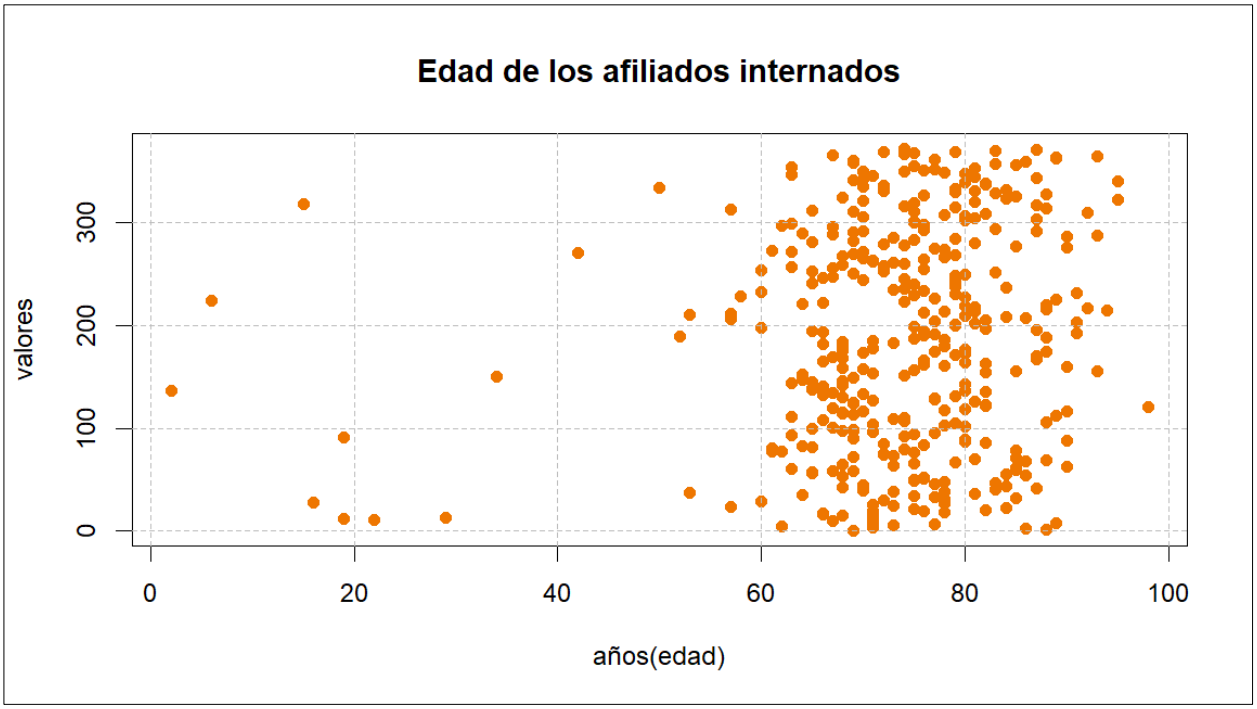


Fig. 29. Edad de los afiliados internados**.** Fuente: elaboración propia (2024)

MEDIDAS DE DISPERSIÓN

Distribución de frecuencias

Se calcularon las frecuencias de las categorías en las variables de sexo y patología para comprender la distribución de los datos.

En valores absolutos el dataset sobre internaciones en cuanto a género está conformado por 6.249 mujeres y 4.432 varones. En términos porcentuales se puede expresar como 60% de mujeres internadas y 40% de varones.

```
> table(ds$sexo)
Femenino        Masculino
6249            4432
>round(prop.table(table(ds$sexo)), 1)
Femenino        Masculino
0.6              0.4
```

Tabla de frecuencia sobre el total por sexo en cada uno de las localidades relevadas.

```
table(ds$localidad_centro_salud, ds$sexo)
                    Femenino        Masculino
```

Alvear	4	3	
Bella Vista	406	390	
Corrientes	4263	2911	
Curuzu Cuatia	116	86	
Esquina	82	49	
Goya	148	133	
Mercedes	240	182	
Monte Caseros	512	390	
Paso de los Libres	451	268	
San roque	2	0	
Sauce	24	15	
Virasoro	1	5	

Distribución de frecuencias para el código = "C00-D48", que corresponde a Neoplasias (cáncer)

```
>table(ds$CIE10_ID=="C00-D48", ds$sexo)
       Femenino  Masculino
FALSE      5614       3884
TRUE        635        548
```

Distribución de frecuencias para el código = " I00-I99", que corresponde a enfermedades del aparato circulatorio. Algunas de ellas son enfermedades hipertensivas, isquémicas del corazón, enfermedades cardiopulmonares, entre otras más.

```
>table(ds$CIE10_ID=="I00-I99", ds$sexo)
       Femenino  Masculino
FALSE      5408       3837
TRUE        841        595
```

Distribución de frecuencias para el código = " J00-J99", que corresponde a Enfermedades del aparato Respiratorio. Gripe, neumonía, Rinitis, faringitis y rinofaringitis crónica, entre otras.

```
>table(ds$CIE10_ID=="J00-J99", ds$sexo)
       Femenino  Masculino
FALSE      5527       3911
TRUE        722        521
```

Distribución de frecuencias para el código = " E00-E90", que corresponde a Enfermedades endocrinas, nutricionales y metabólicas. De este grupo de enfermedades las que más nos ocupa es la diabetes.

```
>table(ds$CIE10_ID=="E00-E90", ds$sexo)
       Femenino  Masculino
FALSE      5736       4043
 TRUE       513        389
```

Seguidamente se presente en la Fig. 30 la distribución de frecuencias de todas las patologías para cada una de las categorías sexo (femenino, masculino). Los códigos y su correspondiente descripción se encuentran en la Fig. 30.

```
> internaciones <- read.csv2("internadosB.csv")
>ds<- internaciones
> sexo <- ds$sexo
>patologia<- ds$CIE10_ID
>newdata<- ds[, c("sexo", "CIE10_ID")]
>Tabla.contingencia<- table(newdata$sexo, newdata$CIE10_ID)
>Tabla.contingencia
```

```
      A00-B99 C00-D48 D50-D89 E00-E90 F00-F99 G00-G99 H00-H59 I00-I99 J00-J99 K00-K93 L00-L99
0   0     184     635       8     513     105      17       0     841     722     768     106
1   0      63     548      14     389      63      11       1     595     521     599      74

  M00-M99 N00-N99  na O00-O99 P00-P96 Q00-Q99 R00-R99 S00-T98 U00-U99 Z00-Z99
0     187     295 334      65       0       2     993     457       2      15
1      93     267 234       2       1       1     719     228       1       8
```

Fig. 30. Distribución de frecuencias de todas las patologías. Fuente: elaboración propia (2024).

Informe de exploración de datos

El objetivo en esta fase de la metodología CRISP-DM fue explorar, descubrir, empezar a conocer los datos sobre los cuales se va a operar.

Basados en los datos procesados hasta el momento se pudo determinar las variables más prometedoras para la aplicación de los futuros algoritmos de datamining. Entre ellos encontramos a los atributos edad, sexo, domicilio y patología.

Los parámetros estadísticos sobre la distribución de las enfermedades nos dieron una idea más clara en cuanto a representación de las enfermedades no transmisibles sobre el total e enfermedades registradas en los pacientes internados.

Otras variables (motivos y destino) pertenecientes al dataset que presenta las comisiones realizadas serán tenidas en cuenta en las siguientes fases de la metodología. En ésta se procedió a describirla sin ningún otro tratamiento estadístico. En las siguientes fases habría que determinar el número de trabajadores sociales del instituto y su territorio de influencia en relación con la cantidad de afiliados.
Los datos obtenidos del Ministerio de salud de Provincia sobre las causas de muertes pueden ser cotejadas con los resultados que se obtengan al final la aplicación de los algoritmos. Especialmente en la etapa de Discusión del proyecto de investigación.
El atributo "médico de cabecera" del dataset padrón de afiliados es potencialmente importante y pudiendo ser objeto de futuras aplicaciones, aunque por el momento se descarta su empleo.
No se formuló ninguna hipótesis ni tampoco las exploraciones revelaron nuevas características de los datos.
El estudio realizado sobre los datos de internaciones plantea la posibilidad de trabajar con un subconjunto para la etapa de modelado. El fundamento pasa principalmente por la no disponibilidad de recursos y/o permisos sobre datos que nos permitan hacer un cruzamiento con otras fuentes, y así ampliar con nuevos atributos.

4.2.4 Verificaciones y gestión de la calidad

Los datos son la base de la investigación, su calidad es importante para que el trabajo tenga sentido y aporte unos resultados coherentes y útiles. Esta etapa se dedica a la verificación de los datos obtenidos.

Se sabe por definición, ningún dato es perfecto. Todo dato que se utiliza contendrá errores, y estos pueden ser desde totalmente irrelevantes hasta de una gravedad que desvirtúen por completo los resultados.

- **Identificación de problemas y soluciones**

En el proceso de revisión y análisis de los datos se identificaron los siguientes problemas.

- **Los datos perdidos**: incluyen valores vacíos o codificados como sin respuesta (por ej. $null$,? o 999). Se han encontrado en el dataset internaciones 135 registros que en el campo diagnostico no contenían valor alguno, por lo tanto, descartados.
- **Los errores de datos**: suelen ser errores tipográficos cometidos al introducir los datos. De este tipo de error se estuvo consciente desde la obtención de los mismos, en especial con el atributo

diagnóstico. La razón por la cual se procedió a normalizar utilizando la clasificación internacional de enfermedades CIE-10.

- **Los errores de mediciones**: Este error se aplica en espacial con el atributo domicilio de los afiliados que estuvieron internados. El problema surgió con las poblaciones rurales, donde encontramos muy pocas referencias por citar algunas "Parajes Punta Portillo y María Carapé s/n", "calle Ejército Argentino 000". Para subsanar o minimizar este tipo de errores se realizaron consultas con lugareños, autoridades municipales de la zona en cuestión. De todas maneras, los registros con estas características no fueron en un porcentaje relevante que afecta al trabajo desarrollado.
- **Las incoherencias**: de codificación suelen incluir unidades no estándar de medida o valores incoherentes, como el uso de M y F para expresar el género. Con categoría referente a sexo se codifico con 0 femenino y 1 masculino.
- **Los metadatos erróneos:** incluyen errores entre el significado aparente de un campo incluido en un nombre o definición de campo. No se han encontrado.

4.3 PREPARAR LOS DATOS

Esta etapa se estima que consume el 50–70% del tiempo y esfuerzo en un proyecto de minería de datos.

4.3.1. Selección de los datos.

Un dataset presentado y no trabajado, trata con los destinos de comisión de servicios. Estos datos se encontraban manuscrito y fueron digitalizado (planilla de cálculo), ver tabla N° 10.

Tabla 10. Comisiones de servicios, periodo 2014-2019.

Id	origen	frecuencia
1	Goya	56
2	Mercedes	27
3	Paso de los Libres	27
4	Esquina	25
5	Bella vista	23
6	Santo Tome	22
7	Ituzaingó	12
8	Monte Caseros	12
9	Curuzu Cuatia	11
10	La Cruz	11
11	San Roque	11
12	Empedrado	10
13	Felipe Yofre	9
14	Loreto	9
15	Mariano I Loza	9
16	Mburucuya	9
17	Chavarría	8
18	9 de Julio	7
19	Perugorría	7
20	Tabay	7

Fuente: elaboración propia (2023)

En las tablas N° 11 y 12 se presentan los datos sobre la estructura del dataset con las internaciones y algunas observaciones para su mejor comprensión. A partir de estos se crearán dos bases minables con distintos objetivos de minería.

.

Tabla 11. Estructura del dataset internaciones

Atributo	incluido	Motivo
UGL Original	NO	todos los datos se refieren la UGL de Ctes
Nombre y Apellido	SI	Se necesitan para futuros cruzamientos con otras fuentes
N° Beneficiario	SI	Se necesitan para futuros cruzamientos con otras fuentes
Prestador Asignado Originalmente	No	Es potencialmente un atributo importante pero para nuestros objetivos de datamining son descartados
Fecha Ingreso	SI	Potencialmente útiles
Fecha Egreso	SI	Potencialmente útiles
Diagnóstico	SI	Potencialmente útiles
Sector	No	No tiene relevancia para nuestros algoritmos
Observaciones (diagnóstico actual, retraso insumos, intercurrencias etc.)	NO	No es relevante para el proyecto
Estancia sin corrección	NO	No es relevante para el proyecto
Estancia en Días	NO	No es relevante para el proyecto

Fuente: elaboración propia (2023)

Tabla 12. Dataset internaciones con la carga de algunos registros.

lugar_internacion	localidad_centro_salud	nombre	fecha_ing	fecha_egre	diagnostico
Clínica de la Mujer y Niño	Corrientes	XXXXX	16/2/2019	20/2/2019	ABD AGUDO
Cardiocentro SRL	Corrientes	XXXXX	22/05/19	23/05/19	ABD AGUDO
Cardiocentro SRL	Corrientes	XXXXX	17-sep	19-sep	ABD AGUDO
Centro medico S.A.	Corrientes	XXXXX	11/01/19	16/01/19	ABDOMEN AGUDO
Centro médico SRL	Corrientes	XXXXX	22/08/19		ABDOMEN AGUDO
Centro médico SRL	Corrientes	XXXXX	13/12/19	16/12/19	ABDOMEN AGUDO
Clínica Bella Vista SRL	Bella Vista	XXXXX	21/10/19		ABDOMEN AGUDO
Clínica de la Mujer y Niño	Corrientes	XXXXX	08/11/19	13/11/19	ABDOMEN AGUDO
Clínica del Sol	Corrientes	XXXXX	31/1/2019	02/02/19	ABDOMEN AGUDO
Clínica del Sol	Corrientes	XXXXX	15/8/2019	20/08/19	ABDOMEN AGUDO
Clínica Madariaga	Paso de los Libres	XXXXX	23/01/19	26/01/19	ABDOMEN AGUDO
Clínica Madariaga	Paso de los Libres	XXXXX	26/06/19	28/06/19	ABDOMEN AGUDO
Sanatorio del Norte SRL	Corrientes	XXXXX	12/02/19	14/02/19	ABDOMEN AGUDO
Sanatorio del Norte SRL	Corrientes	XXXXX	24/03/19	25/3/2019	ABDOMEN AGUDO
Sanatorio del Norte SRL	Corrientes	XXXXX	08/08/19	16/8/2019	ABDOMEN AGUDO

Fuente: elaboración propia (2023)

4.3.2. Limpieza de datos.

Esta etapa fue llevada a cabo durante el proceso de verificación y gestión de calidad de los datos.

4.3.3. Construcción del juego de datos.

A partir de los datos de viajes al interior de la provincia, y sobre los cuales se aplicarán algoritmos, se observó la necesidad de la creación de dos campos (X e Y) de tipo geométrico que almacenaran las coordenadas de latitud y longitud. En la Tabla 13 a modo de ilustración se presentan 10 primeras observaciones con el nuevo campo geográfico.

La carga de datos en esos atributos se hizo de manera manual, empleado el servidor de aplicaciones de mapas en la web Google Maps.

Tabla 13. Dataset comisiones

id	origen	frecuencia	x	y
1	Goya	56	-59,257073	-29,152411
2	Mercedes	27	-58,057271	-29,194457
3	Paso de los Libres	27	-57,134514	-29,692501
4	Esquina	25	-59,511728	-30,027324
5	Bella vista	23	-59,041384	-28,508597
6	Santo Tome	22	-56,041298	-28,551168
7	Ituzaingó	12	-56,690096	-27,59407
8	Monte Caseros	12	-57,640489	-30,255029
9	Curuzu Cuatia	11	-58,059766	-29,783335
10	La Cruz	11	-56,653482	-29,173851

Fuente: elaboración propia (2023)

Sobre el dataset “internaciones” se construyeron nuevos datos. La base minable denominada “GEO_PATOLOGIAS” fue diseñada con la intención de aplicar algoritmos cluster. Este conjunto de datos está conformado por aproximadamente 370 observaciones (registros). Y esta misma base minable, pero sin contar con los atributos geométricos que conforman 10.800 se destinó para algoritmos predictivos, ver Tabla 14.

Tabla 14. Estructura dataset internaciones

Atributo	Nuevo dato	Valor
lugar_internacion	No	-
localidad_centro_salud	No	-
nombre	No	-
fecha_nac	No	-
edad	No	-
grupo_etareo	No	-
Sexo	No	-
coord_x	Si	Contiene dato sobre la coordenadas de longitud
coord_y	Si	Contiene dato sobre la coordenadas de latitud
fecha_ing	No	-
fecha_egre	No	-

Atributo	Nuevo dato	Valor
diagnostico	No	-
CIE10_ID	Si	Código de enfermedades según estándar CIE-10
CIE10_DES	Si	Descripción de enfermedades según estándar CIE-10
periodo	No	-

Fuente: elaboración propia (2023)

4.3.4. Integración de los datos.

A lo largo de todo este trabajo se utilizaron múltiples orígenes de datos. En esta etapa de la metodología se hace uso de las mismas.

Sobre los datos "GEO_PATOLOGIAS", se integraron otros campos. Partiendo de la información de nombre y apellido, CUIL, y numero de beneficiario se cruzaron con los contenidos en el padrón de afiliados para obtener los nuevos campos de localidad y Dpto. de afiliado, médico de cabecera. El CUIL fue importante en algunos casos en que había duda sobre el género del paciente. De esta manera se conformaron las 2 bases, la que contenía los campos geométricos (372 observaciones) y la otra sin esta adicción (10.800 registros)

4.4. **Modelado**

Este es el punto donde todo el trabajo previo cobra sentido. La preparación de los datos se incorpora a los algoritmos y técnicas seleccionadas.

Esta fase comúnmente es un proceso iterativo, donde sobre los datos seleccionados se aplican los algoritmos, se observan los resultados y se ajustan parámetros. Es posible volver a fases previas, preparando nuevamente nuestra colección de datos en pos de alcanzar los objetivos planteados al inicio.

Teniendo en vista siempre los objetivos de la investigación, se seleccionaron los modelos que a priori se estimó más conveniente, ver Tabla 15.

Tabla 15. Técnicas – algoritmos aplicados en el proyecto de investigación

Técnica / algoritmo	Clasificación	Pregunta de investigación
Densidad de Kernel (HeatMap)	Técnica de aprendizaje automático No Supervisado.	¿Cuáles son las áreas geográficas con mayores vulnerabilidades?
Vecino más próximo (KNN)	Técnica de aprendizaje automático Supervisada	¿Existe un patrón o tendencia en la distribución espacial de los datos?
Clusters K - Means	Técnica de aprendizaje automático No Supervisada	¿Existen agrupaciones de interés?
ANOVA (Análisis de la Varianza)	Técnica de análisis estadística. Supervisada.	¿Se puede encontrar relaciones entre diferentes variables?
Regresión lineal[5]	Técnica Supervisada de predicción.	¿Se puede predecir los reingresos hospitalarios?

Fuente: elaboración propia (2024)

4.4.1 TECNICA ESTIMACION DE DENSIDAD DE KERNEL

La densidad se calcula en función del número de puntos en una ubicación, con un mayor número de puntos agrupados que dan como resultado valores más grandes. Los mapas de calor permiten una fácil identificación de *puntos calientes* y la agrupación de puntos.

La decisión del empleo de esta técnica se debe a la facilidad de interpretación de los resultados. Ofrece una visualización muy fácil de comprender. En la Fig. 31 se observan las zonas más visitadas durante los viajes realizados en el abordaje en territorio durante 2019, aquellas con un color más

[5] No se llegó a aplicar. Algoritmo a desarrollar en una futura investigación.

intenso

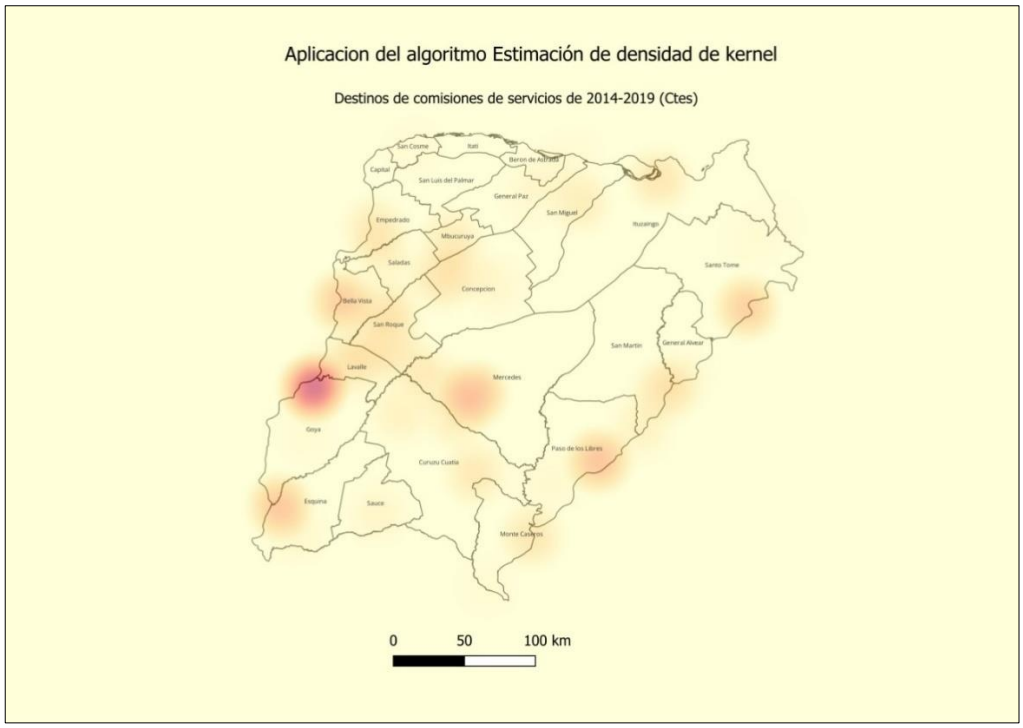


Fig. 31. Destinos de comisiones periodo 2014-2019. Mapa de calor elaborado en QGIS**.**

Usando los mismos datos, pero con el IDE Rstudio se generó una nueva gráfica, ver Fig. 32. A continuación se incluye el código correspondiente.

```
#instalamos la libreriaforeing
install.packages("foreign")
# cargamos las librerias
library("sp")
library("sf")
library("ggplot2")
library("tmap")
library("foreign")
#espacio de trabajo
getwd()
#Importar csv con los puntos de localidades
puntos <- read.csv2 ("datos_prueba_comisiones.csv", encoding = "UTF-8")
#Importamos el Shapefile con la geometría de la prov de Corrientes
ctes<- st_read("ctesdivsion.shp")
# determinamos las coordenadas
```

```
coordinates(puntos) <- ~x + y
#transformamos el objeto puntos en un sf
puntos <- st_as_sf(puntos)
#establecemos el sistema de coordenadas
st_crs(puntos)=4326
frec <- puntos[puntos$frecuencia, ]
coord_den = st_coordinates(frec)
# graficamos
ggplot(data=frec) + geom_sf(alpha=1.5)
ggplot() + geom_sf(data= frec, size=2, alpha=0.2) + stat_density2d(aes(x=coord_den[,1],
y=coord_den[,2], fill=..level.., alpha=..level..), geom="polygon") +
scale_fill_gradient(low="yellow", high="red")
```

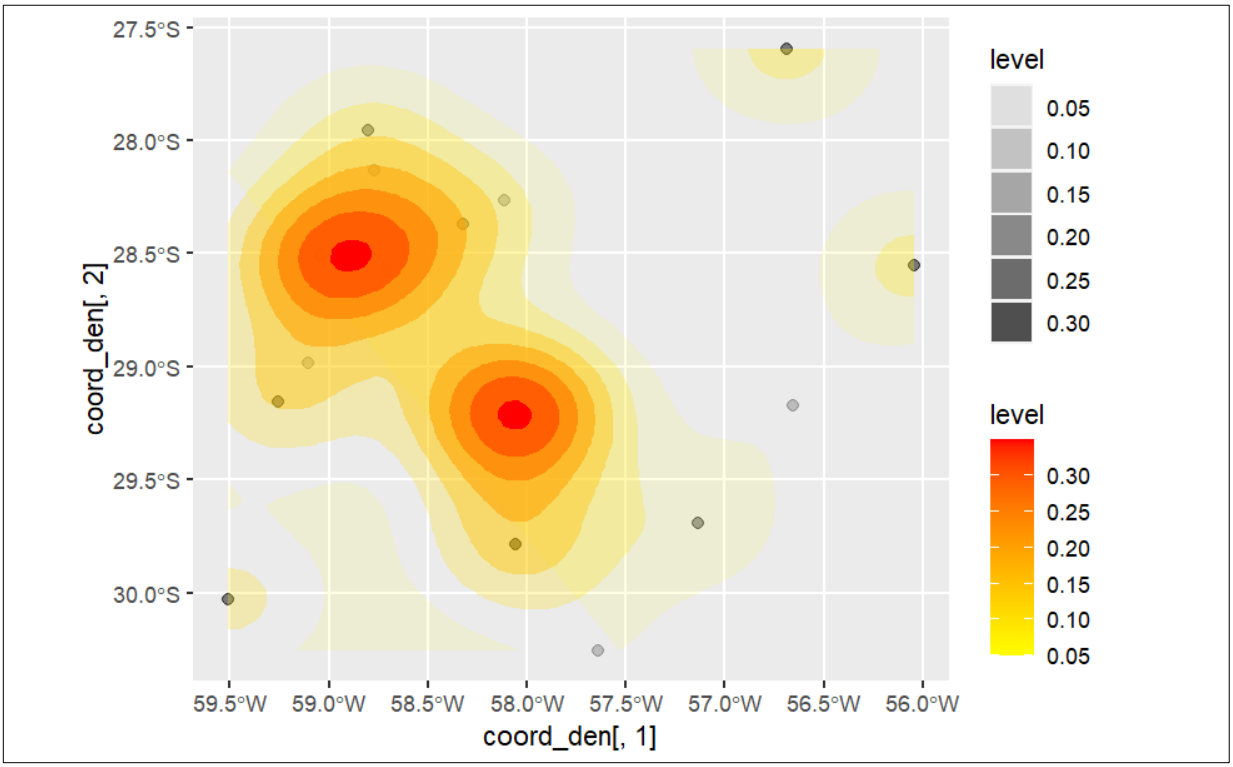


Fig. 32. Destinos de comisiones periodo 2014-2019. Mapa de calor elaborado en R

4.4.2 ANÁLISIS DE VECINOS MÁS PRÓXIMO K-NN

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN ((Nearest Neighbor), es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Si bien se puede

usar para problemas de regresión o clasificación, generalmente se usa como un algoritmo de clasificación, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro.

El índice de vecino más cercano es una medida utilizada en análisis espaciales para evaluar la distribución de puntos en un conjunto de datos. Un valor alto de índice de vecino más cercano indica que los puntos están más dispersos, mientras que un valor bajo sugiere una mayor agrupación o patrón de puntos cercanos entre sí.

En el caso de estudio, interesa determinar si los datos analizados que contienen las coordenadas de los domicilios de los pacientes internados presentan un patrón sistemático. Es decir, observar si existe un patrón de agregación, de aleatoriedad o están disperso en el espacio.

En la Fig. 33, se presenta el mapa con la distribución de los domicilios pertenecientes a los pacientes internados durante 2019.

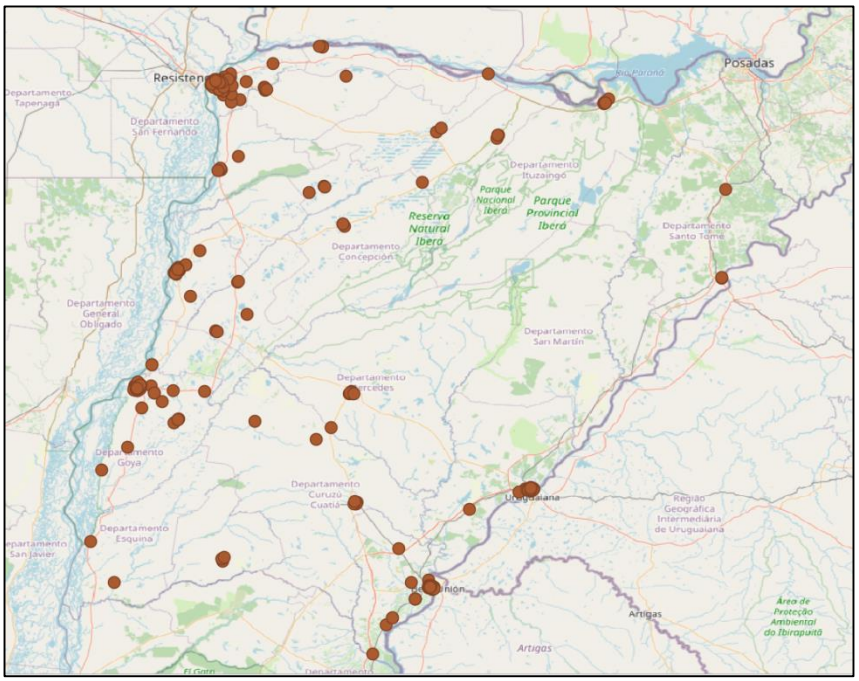


Fig. 33. Distribución de los domicilios de los pacientes

El dataset está proyectado en coordenadas geográficas (latitud - longitud), las cuales deben ser reconvertidas a coordenadas planas, previo a la aplicación del algoritmo. Para nuestro caso, el paso consiste es reproyectar la capa de EPSG 4326 a Postgar 94 faja 5. En la Fig. 34, se ilustra dicha tarea en QGis.

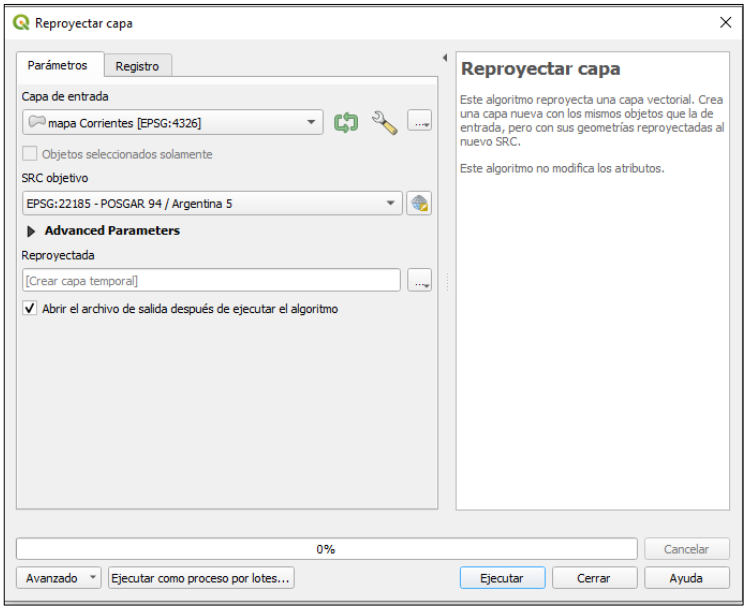


Fig. 34. Reproyectar SRC en capa Qgis.

Una vez realizado esta modificación, se procede a la aplicación del algoritmo "Vecinos más próximos" (K-NN), el cual genera los siguientes valores.

Observed mean distance:1803.18126903175

Expected mean distance:9280.56128237662

Nearest neighbor index: 0.19429657476

Number of points:371

Z-Score:-29.68882758604

Con el propósito de continuar en un futuro y profundizar el estudio, se analizó tres zonas en particular: la ciudad de Corrientes Capital, Goya y una ciudad del margen del rio Uruguay, Monte Caseros.

En la fig. N° 35 se presenta la distribución de los domicilios de los pacientes internados en la ciudad de corrientes y seguidamente los valores obtenidos por K-NN.

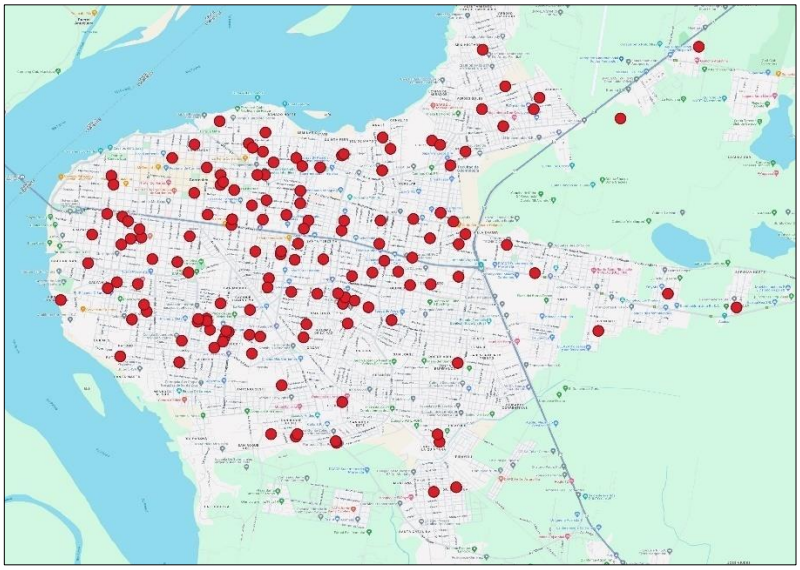


Fig. 35. Casos referenciados en la ciudad de Corrientes.

Ciudad de Corrientes

Distancia promedio observada: 485.05471279449

Distancia media esperada: 745.19461712766

Índice de vecino más cercano: 0,65091011347

Número de puntos: 149

Puntuación Z:-8,15194581186

En la Fig. 36 se localizan los casos pertenecientes a la ciudad Goya, luego los valores de K-NN.

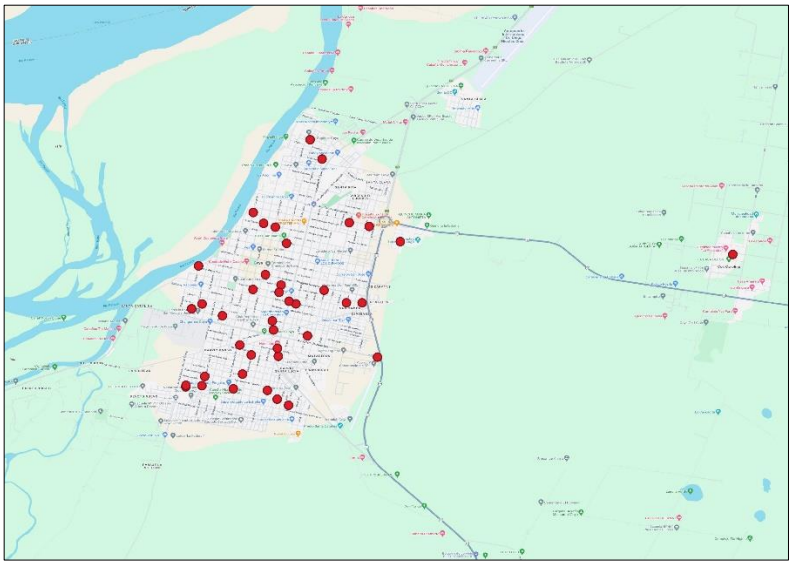


Fig. 36. Casos referenciados en la ciudad de Goya.

Ciudad de Goya

Distancia media observada: 983.61254860111

Distancia promedio esperada: 3782.42541753713

Índice de vecino más cercano: 0,52442872750

Número de puntos: 52

Puntuación Z: -6,56066960560

Y finalizando la etapa de aplicación de K-NN a las zonas de estudios seleccionadas se obtiene el mapa de la ciudad de Monte Caseros y sus casos georreferenciados, ver Fig. 37.

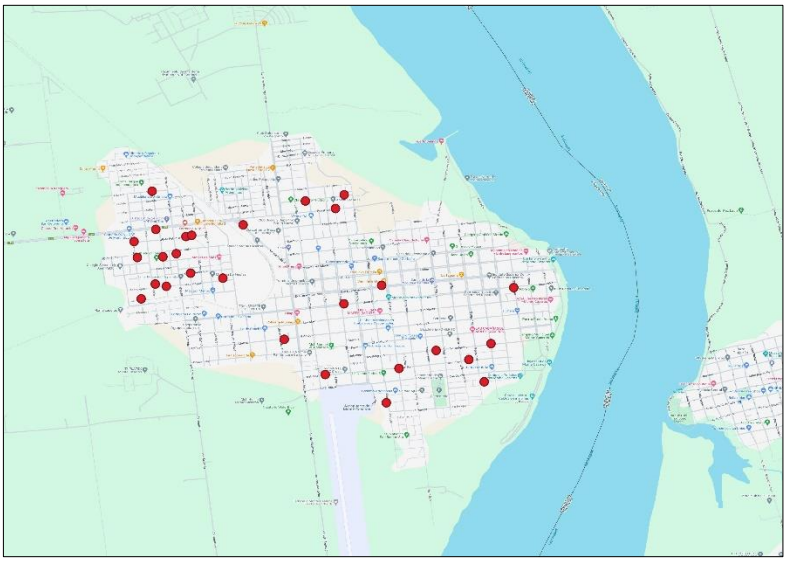


Fig. 37. Casos referenciados en la ciudad de Monte Caseros.

Ciudad de Monte Caseros

Distancia media observada: 1713.28650393297

Distancia media esperada: 3708.83125399780

Índice de vecino más cercano: 0.46194781768

Número de puntos: 39

Puntaje-Z: -6.42817340376

4.4.3 ALGORITMO DE CLUSTERING K - MEANS

Es un algoritmo de agrupamiento que se utiliza para dividir un conjunto de datos en k grupos o clusters basados en características similares. El objetivo es minimizar la suma de las distancias cuadradas de cada punto al centroide de su cluster asignado. Es un algoritmo de aprendizaje no supervisado y se utiliza comúnmente para la segmentación de datos.

Durante el desarrollo de la investigación surgieron nuevas fuentes de datos que no formaron parte de las colecciones iniciales pero dada su potencial importancia fueron agregados en esta etapa. Es el caso de un dataset con información sobre los subsidios otorgados en el año 2019. Su estructura consiste en 31 atributos, de los cuales se seleccionó localidad, número de casos y se añadió una columna con los valores de índice de dependencia (INDEC). Esta ayuda económica se destina a los afiliados con algún grado de dependencia bajo la denominación de subsidio PADyF, programa de apoyo a la dependencia y vulnerabilidad. El conjunto de datos fue extraído desde uno de los sistemas internos de la obra social, en Fig. 38 se presenta su estructura.

La importancia de este dataset es evidente por cuanto está destinado a uno de los grupos de afiliados más vulnerables.

Fecha de carga	Apellido y Nombre afiliado	Nro de Beneficio	Sexo	Agencia	Lugar de residencia
03/10/2019			FEMENINO	0000 - UGL II - CORRIENTES	
30/12/2019			MASCULINO	0000 - UGL II - CORRIENTES	
31/10/2019			FEMENINO	0001 - AGENCIA GOYA	
21/10/2019			MASCULINO	0003 - CAP SANTO TOME	
30/09/2019			FEMENINO	0002 - CAP PASO DE LOS LIBRES	
28/08/2019			MASCULINO	0001 - AGENCIA GOYA	
10/12/2019			MASCULINO	0050 - CAP GRAL PAZ	
07/11/2019			FEMENINO	0009 - CAP ESQUINA	
05/08/2019			FEMENINO	0001 - AGENCIA GOYA	
12/11/2019			MASCULINO	0000 - UGL II - CORRIENTES	
26/12/2019			MASCULINO	0001 - AGENCIA GOYA	
27/08/2019			FEMENINO	0000 - UGL II - CORRIENTES	
25/11/2019			MASCULINO	0003 - CAP SANTO TOME	
15/10/2019			MASCULINO	0008 - CAP GOBERNADOR VIRASORO	

Fig. 38. Listado de afiliados que perciben subsidio PADYF.

Para analizar y determinar si se establecen grupos con ciertas características similares se aplicó el algoritmo de Clustering K-Means. Este permite identificar patrones en las localidades en función de las variables que están disponibles en el dataset.

```
getwd()
setwd("H:/WORKING")
# Cargar los datos
datos <- read.csv2("subsidios.csv")
#transformacion de las variables casos y afiliados
datos$casos <- as.numeric(datos$casos)
```

```
datos$afiliado <- as.numeric(datos$afiliado)
#normalización de los datos
datos_norm <- datos[, c("casos", "afiliado", "indice_dependencia")]
# Aplicar K-means con 3 clusters
kmeans_result <- kmeans(datos_norm, centers = 3)
# Añadir la columna de cluster asignado a los datos
datos$cluster <- kmeans_result$cluster
# Mostrar los resultados
print(kmeans_result)
# Crear un dataframe con las localidades y su cluster asignado
localidades <- data.frame(localidad = 1:nrow(datos), cluster = kmeans_result$cluster)
# Mostrar las localidades y su cluster asignado
print(localidades)
```

```
K-means clustering with 3 clusters of sizes 8, 4, 4
Cluster means:
casos  afiliado indice_dependencia
1           4.50       3.646875             13.550
2           19.25      2.970000             13.700
3           99.00      13.998000           11.325
Clustering vector:
 [1] 3 3 3 3 2 2 2 2 1 1 1 1 1 1 1 1
Within cluster sum of squares by cluster:
[1]   99.95629  146.69429 7874.79413
 (between_SS / total_SS= 75.5 %)
Available components:
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter
"          "ifault"
```

Los resultados obtenidos del algoritmo K-means muestran que se han identificado 3 clusters en los datos, con tamaños de 8, 4 y 4 observaciones respectivamente. A continuación, se detallan algunas de las métricas obtenidas:

- Medias de cada cluster para las variables "casos", "afiliado" e "indice_dependencia".
  - Cluster 1: casos=4.50, afiliado=3.646875, indice_dependencia=13.550
  - Cluster 2: casos=19.25, afiliado=2.970000, indice_dependencia=13.700
  - Cluster 3: casos=99.00, afiliado=13.998000, indice_dependencia=11.325

- Vector de clustering: muestra a qué cluster ha sido asignada cada observación. Las primeras 4 observaciones pertenecen al cluster 3, las siguientes 4 al cluster 2, y las últimas 8 al cluster 1.
- Suma de cuadrados dentro de cada cluster: indica cuánta variabilidad hay dentro de cada cluster en comparación con la variabilidad total en los datos.

En general, estos resultados permiten entender cómo se agrupan las observaciones en función de las variables analizadas. Otra línea o tarea de investigación podría analizar más a fondo cada cluster para identificar patrones o características distintivas en cada uno.

```
>print(localidades)
   localidad cluster
1         1       3
2         2       3
3         3       3
4         4       3
5         5       2
6         6       2
7         7       2
8         8       2
9         9       1
10       10       1
11       11       1
12       12       1
13       13       1
14       14       1
15       15       1
16       16       1
```

```
# Calcular las medias de las variables de interés para cada cluster
cluster_means<- aggregate(datos[, c("casos", "afiliado", "indice_dependencia")], by = list(localidades$cluster), FUN = mean)
# Renombrar las columnas
colnames(cluster_means) <- c("cluster", "media_casos", "media_afiliado", "media_indice_dependencia")
# Mostrar las medias por cluster
```

```
print(cluster_means)
```

```
># Mostrar las medias por cluster
>print(cluster_means)
  cluster media_casosmedia_afiliadomedia_indice_dependencia
1       1        4.50       3.646875                  13.550
2       2       19.25        2.970000                  13.700
3       3       99.00      13.998000                   11.325
```

```
# Obtener los nombres de las localidades correspondientes al Cluster 3 en el dataframe original
'datos'
nombres_cluster3 <- datos$localidad[cluster3_localidades]
# Mostrar los nombres de las localidades en el Cluster 3
print(nombres_cluster3)
> # Mostrar los nombres de las localidades en el Cluster 3
>print(nombres_cluster3)
[1] "CORRIENTES" "GOYA"       "SANTO TOME" "VIRASORO"
```

En la Fig. 39 se presenta el grafico correspondiente a la aplicación del algoritmo K-Means en Rstudio.

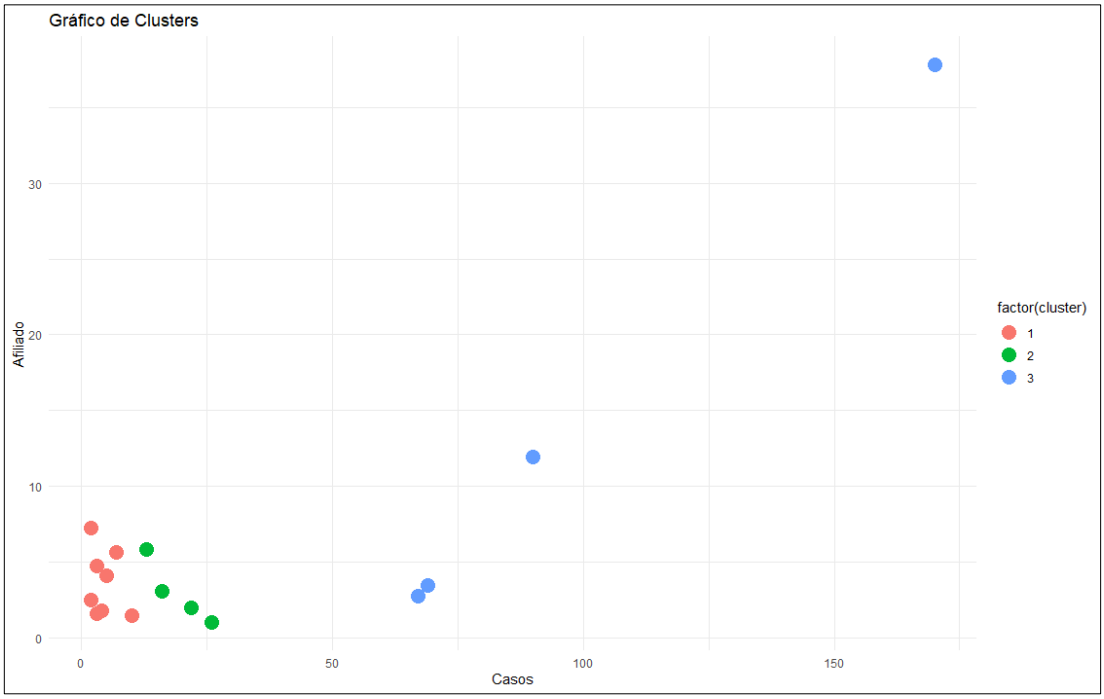


Fig. 39. Cluster de localidades agrupadas por índice de vulnerabilidad

CLUSTERING según atributo edad de afiliados

En este caso tomando el dataset que contiene las patologías, pero haciendo foco en el comportamiento del atributo *edad* de los pacientes. Seguidamente se presenta código en R y su correspondiente gráfica, ver Fig. N°40.

```
getwd()
setwd("E:/WORKING")
install.packages("stats")
library(stats)
datos <- read.csv2("GEO_PATOLOGIAS2.csv")
datos$edad<- as.numeric(datos$edad)
kmeans_model<- kmeans(datos$edad, centers = 3)
between_ss<- sum(kmeans_model$betweenss)
edades_df<- data.frame(edad = datos$edad, grupo = as.factor(kmeans_model$cluster))
ggplot(edades_df, aes(x = edad, color = grupo)) +
geom_density() +
labs(title = "Agrupamiento de edades por k-means")
```

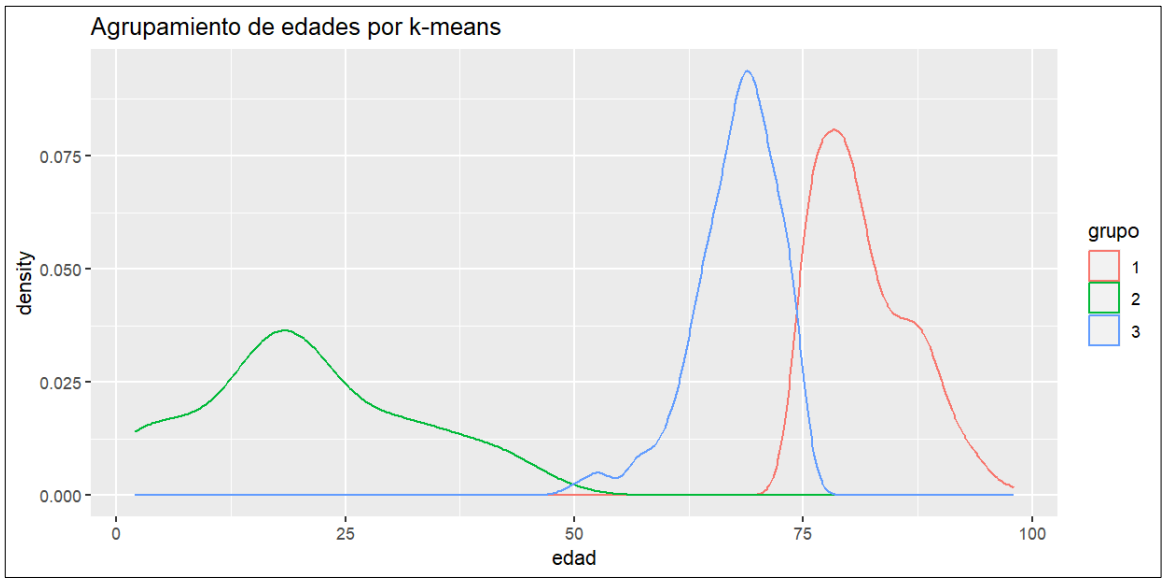


Fig. 40. Gráfico de agrupación según la edad del afiliado.

4.4.4 TÉCNICA ANOVA (ANÁLISIS DE LA VARIANZA)

La técnica ANOVA es una técnica de análisis estadístico supervisado que se utiliza para comparar las medias de tres o más grupos diferentes en función de una variable dependiente.

Para aplicar este algoritmo a nuestros datos - edad y sexo de pacientes con ENT - y así poder determinar si hay diferencias significativas en la edad de los pacientes entre los cuatro grupos de enfermedades, primero debemos verificar si se trata de una distribución normal, (ver Fig. 41).

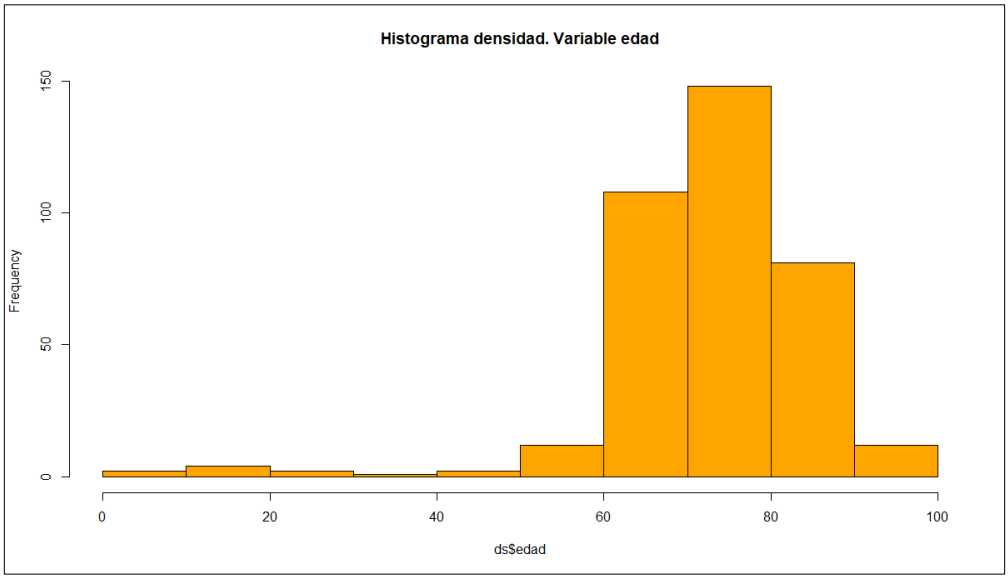


Fig. 41. Histograma de la variable edad.

A simple viste se puede observar en la gráfica anterior, que la edad no tiene una distribución normal. Por ello en un primer intento de acercar a dicha distribución se tomaron - de manera arbitraria - en consideración las edades mayores a 49 años. Resultando la Fig. 42.

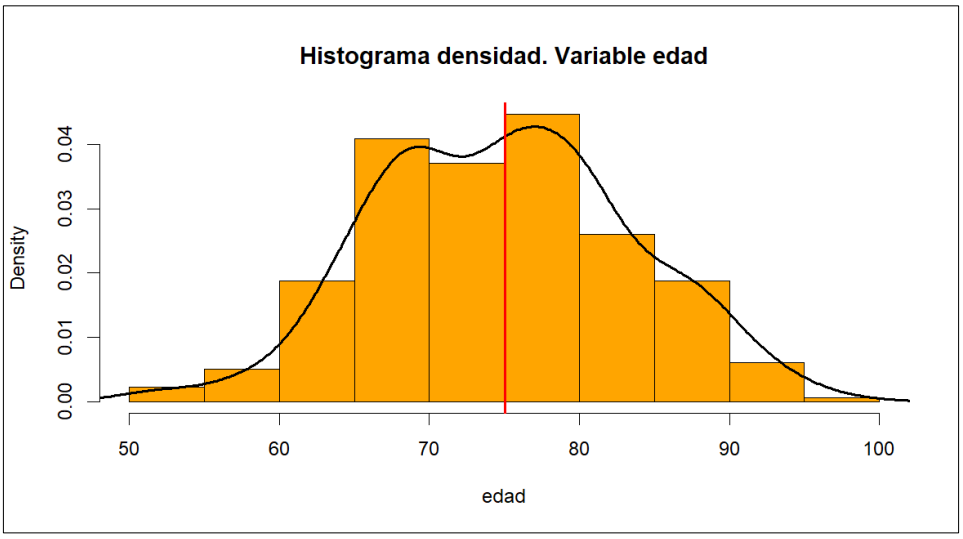


Fig. 42. Histograma de la variable edad de afiliados mayores a 49 años.

Normalidad de los datos: como se mencionó, antes de aplicar ANOVA, es necesario verificar si la distribución de la edad en cada grupo de enfermedad se aproxima a una distribución normal. Se utilizó la prueba de normalidad de Shapiro-Wilk y gráficos como el histograma y el gráfico Q-Q para

evaluar la normalidad de los datos, ver Fig. N°43. Se puede encontrar más información sobre estas pruebas en el Anexo N° 6.

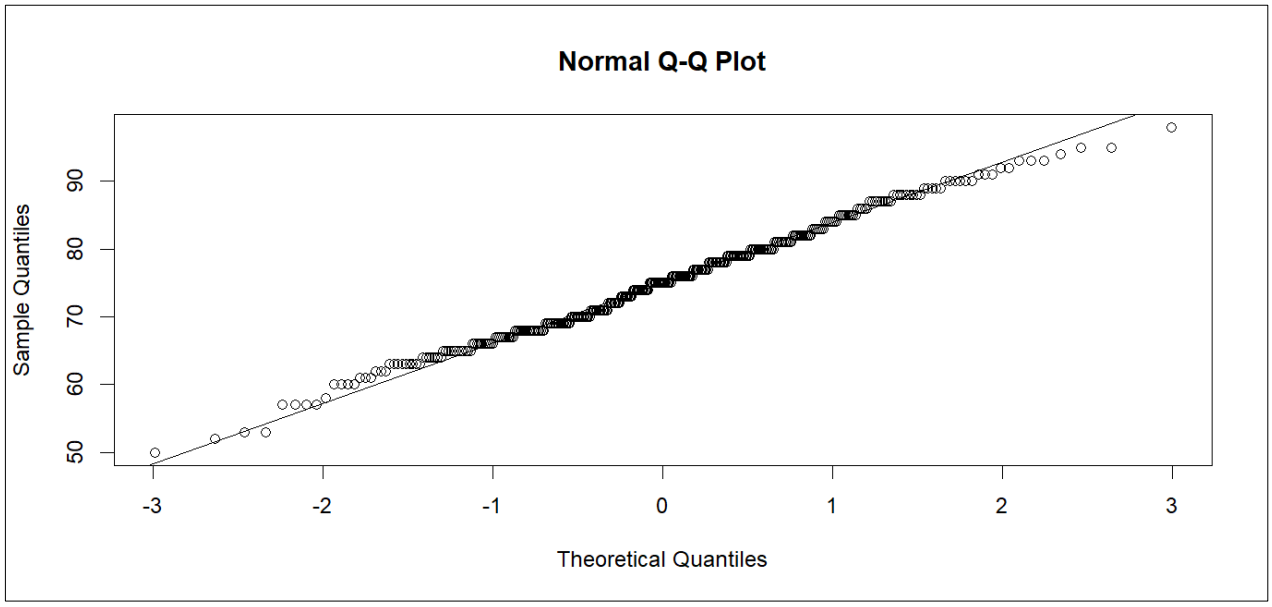


Fig. 43. Gráfico Q-Q para evaluar la normalidad de los datos.

Seguidamente se presenta líneas de código en R aplicando el test de Shapiro – Wilk sobre el atributo edad.

```
shapiro.test(edad)
        Shapiro-Wilk normality test
data:  edad
W = 0.99402, p-value = 0.1665
```

Con los valores obtenidos de p, se puede suponer que la variable edad tiene una distribución aproximadamente normal.

Ahora verificamos la segunda condición para poder aplicar ANOVA, la ***homogeneidad*** de la varianza. Mediante el empleo del Test de Levene podemos analizar este punto. Como se puede observar el fragmento del script en R, se obtiene un p = 0.1006. al ser mayor a un α=0,05. Se concluye que las varianzas son similares.

```
> #visualizamos las primeras 6 observaciones del dataset
```

```
> head(data2)
data.edaddata.Sexo data.CIE10_ID
1       76       0       C00-D48
2       73       1       C00-D48
3       74       0       C00-D48
4       84       0       C00-D48
5       79       0       C00-D48
6       79       1       C00-D48
> #creamos los grupos
>grupo <- data2$data.CIE10_ID
> #convertimos a numerico la edad para poder utilizar en el Test Levene
>data2$data.edad <- as.numeric(data2$data.edad)
> #cargamos las edades en el objeto datos
>datos <- data2$data.edad
> #aplicamos el test
>leveneTest(datos, grupo)
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group3 2.1172 0.1006
     144
```

Aplicación de Anova con las variables sexo, edad y patologías (data$CIE10_ID)

```
modelo_anova<- aov(data$edad ~ data$Sexo * data$CIE10_ID, data = data)
> summary(modelo_anova)
Df Sum Sq Mean Sq F value Pr
(>F)
data$Sexo                       1    87   86.87  0.677  0.412
data$CIE10_ID                   3    217   72.24   0.563  0.640
data$Sexo:data$CIE10_ID  3   522 174.09  1.3570.259
Residuals 140    17964128.31
```

En el análisis de varianza (ANOVA) realizado en RStudio, la tabla resumen proporciona información sobre la significancia de los efectos de los diferentes factores y sus interacciones en el modelo.

- [data$Sexo]: El factor "Sexo" tiene un valor de F de 0.677 y un valor p de 0.412. Un valor p mayor a 0.05 indica que no hay una diferencia significativa en las edades entre los grupos de diferentes sexos.
- [data$CIE10_ID]: El factor "CIE10_ID" tiene un valor de F de 0.563 y un valor p de 0.640. Un valor p mayor a 0.05 sugiere que no hay una diferencia significativa en las edades entre los diferentes grupos de diagnóstico.
- [data$Sexo:data$CIE10_ID]: La interacción entre "Sexo" y "CIE10_ID" tiene un valor de F de 1.357 y un valor p de 0.259. Un valor p mayor a 0.05 indica que no hay una interacción significativa entre el sexo y el diagnóstico en términos de la edad.
- [Residuales]: La variabilidad dentro de los grupos, representada por los residuos, es alta con una suma de cuadrados de 17964 y un promedio de 128.31. Esto sugiere que hay una cantidad significativa de variabilidad no explicada por los factores en el modelo.

4.5. EVALUACIÓN DEL MODELO

En esta fase del proyecto, se completó el desarrollo de los diferentes algoritmos. Y se procede a evaluar los resultados obtenidos en función de los objetivos del negocio. De lo anterior se deduce que se requiere una clara comprensión del negocio.

En relación a los resultados obtenidos.

- Los algoritmos y técnicas seleccionadas expresan con claridad y de forma sencilla los resultados. Estos serán presentados a la(s) gerencia(s) que correspondan como apoyo en las decisiones.
- Se han realizados descubrimientos especialmente importantes que cabe resaltar. Por ejemplo, demostrando la falta o poca planificación al momento de decidir los destinos de comisiones de servicios. Problemática que a la luz de los números viene de larga data y tiene como consecuencia una ausencia de la obra social en zonas geográficas con índices importantes de vulnerabilidad.
- En función de su capacidad de aplicación, el conocimiento generado por los modelos es perfectamente llevado a la práctica.
- En términos de negocio, los modelos desarrollados generan un gran aporte en lo concerniente a lograr mejores planificaciones, uso eficiente de los recursos disponibles, etc. estando en línea con lo establecido en (Declaracion-Lisboa-IIS-InnovacionPublica-ES-05-2023)[23], que promociona modelos de gestión para la innovación pública, considerando la participación activa

de los servidores públicos en la toma de decisiones, en la construcción, en la implementación y evaluación de soluciones.

En la Tabla 16, se presenta información sobre el desarrollo de cada fase. Es una etapa para reflexionar sobre el trabajo llevado adelante, donde se plasman los aciertos y errores. La metodología CRISP – DM brinda como estrategia fundamental aprender desde la propia experiencia. Persiguiendo la meta de generar proyectos de minería de datos más efectivos.

Tabla 16. Proceso de revisión

Etapas	Contribuyo a dar valor de los resultados finales	Existen forma de simplificar	Fallos o errores cometidos	Se produjeron sorpresas (buenas o malas)
Compresión del negocio	SI	Si. Reuniones con personas que tiene poder de decisión o expertos en el área	Asumir presunciones basado en los números obtenidos sin tener en cuenta o disponer de la opinión de expertos	(+) Entrevistas con personal que esclarecieron puntos específicos sobre la dinámica del trabajo social-medico
Compresión de los datos	SI	Si. Investigar métodos que automaticen tareas como extracción de datos, limpieza	Falta de conocimientos o profundización de ellos, en varias temáticas. Por ej. estadísticas, estudios sociales	(+) Obtención de una considerable cantidad de información importante. (-) El trabajo con datos espaciales fue bastante laborioso
Preparar los datos	SI	Si. Investigar métodos que automaticen tareas como extracción de datos, limpieza	Etapas de mala organización en la forma de trabajo que resultaron en retrocesos.	(+) El tiempo. Hubo procesos que nos requirieron más de lo planeado
Modelado	SI	-	Problemas con los datos espaciales (entendimiento)	(+) El tiempo. Hubo procesos que nos requirieron más de lo planeado
Evaluación del modelo	SI	-	-	-
Implementación del modelo *	No aplica	No aplica	No aplica	No aplica

Fuente: elaboración propia (2024)

Capítulo 5

Discusión

En esta sección se analizan los resultados obtenidos en el estudio y se establecen las conclusiones derivadas de los mismos. Se abordan los hallazgos a la luz de la literatura existente y se discuten las implicaciones de los resultados.

En primer lugar, se presenta un análisis de los datos recopilados durante la investigación, destacando las tendencias y patrones identificados. Se discuten las relaciones entre las variables estudiadas y se evalúa el cumplimiento de los objetivos planteados en la investigación.

5.1 ¿Cuáles son las áreas geográficas con mayores vulnerabilidades?

Para dar respuesta a este interrogante y tal como se desarrolló en el capítulo anterior, se utilizó un HeatMap (mapa de densidad de Kernel) generado en la fase anterior, ver Fig. 3.

En el proceso de anonimización, se sustituirán los nombres de las ciudades por un código alfanumérico arbitrario para proteger la privacidad de los datos geoespaciales, ver Anexo N° 7.

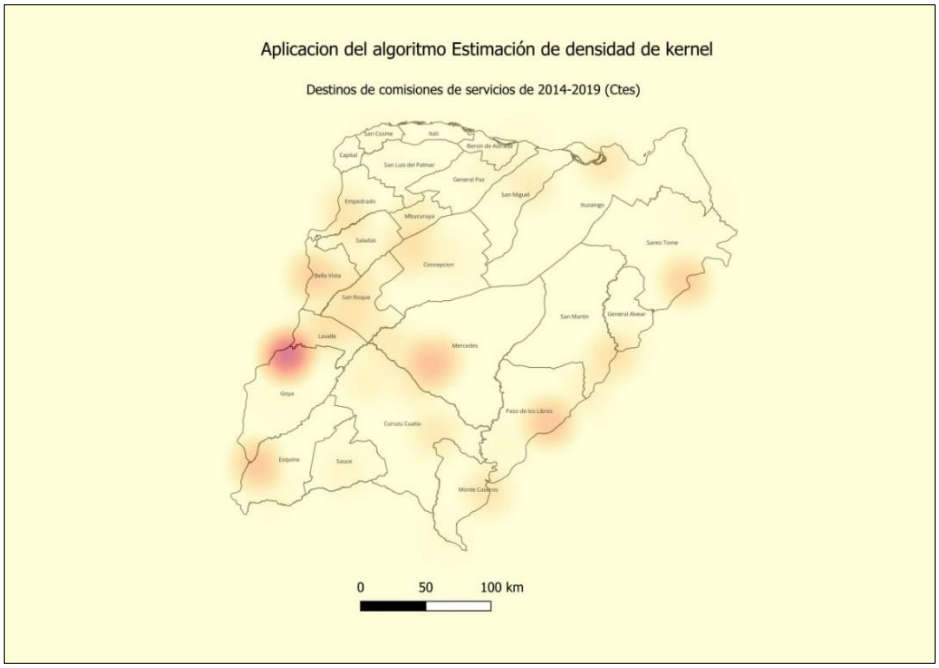


Fig. 31. Destinos de comisiones periodo 2014-2019. Mapa de calor elaborado en QGIS**.**

Es fácil ubicar visualmente las zonas más frecuentadas y las menos, además la ubicación de trabajadores sociales en el territorio provincial, ver Fig. 44. La zona conformada por los Departamentos D09, D02, D22, D12 requirió una especial atención.

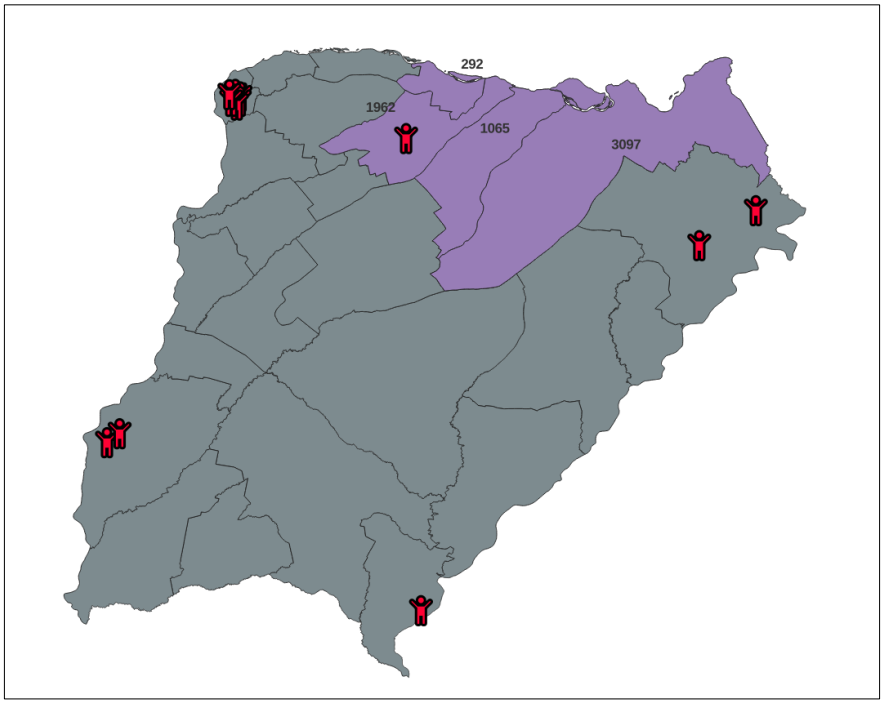


Fig. 44. Distribución de los trabajadores sociales en la provincia de Corrientes.

Abordaje territorial. (en base al mapa con datos de centros de salud en Corrientes)

Análisis de la situación en la Región NOROESTE[6]

(-) Los Dptos. de D02 Y D22 **NO cuentan con agencias de la obra social.**

(-) De ***399*** viajes realizados al interior ***24*** tuvieron como destinos los mencionados Dptos. D12 (12), D29 (9) y D09 (3).

(-) D02 3 establecimientos públicos y ningún privado.

(-) D22 6 públicos y ningún privado.

(-) D09 8 establecimientos públicos y ningún privado.

(-) D12 22 en total, 17 públicos y 5 privados

Cobertura Sanitaria.

[6] Para más información sobre la división de la provincia de Corrientes en regiones consultar el Anexo N° 6.

La zona cuenta con un solo trabajador social (D09), y según datos del padrón de afiliados de 2020, se contabilizaba al momento:

- D09: 1.962 afiliados.
- D12: 3.092 afiliados.
- D02: 292 afiliados.
- D22: 1065 afiliados.

Son aproximadamente 7300 afiliados, en una región de la provincia con las siguientes particularidades: en relación al **índice de analfabetismo** D09 posee 7,7%, D12 7,5% y D22 6,5%. ***Ubicándose las 3 entre las 9 localidades de la Prov. con mayores índices de analfabetismo.***

Dependencia del adulto mayor: D02 y D09***se ubican en el 2do y 3er lugar con los mayores porcentajes de dependencia,*** 17,3% y 17% respectivamente.

Hasta aquí podríamos decir que son datos que muestran si el organismo estuvo o no presente en el territorio. Ahora el análisis refiere a si el afiliado se acerca a la obra social.

Con el objetivo de analizar más en detalle los números -pero esta vez en relación a la actividad de las agencias en esta región - se analiza los datos extraídos del sistema de gestión en la atención (SGA). Es decir, datos sobre las solicitudes generadas en las agencias. Esto generó un dataset de 11.203 registros pertenecientes a los 12 meses del 2019 en las agencias de D09 y D12. A continuación, el código en R sobre el tratamiento de los datos.

```
>aten.itu.gral<- read.csv2("atenciones_gralpaz_itu_2019.csv")
> #ver la estructura del dataset con los registros de atenciones
>str(aten.itu.gral)
'data.frame':        11203 obs. of  3 variables:
$ afiliado :chr  "060319046001 - 00" "060325924406 - 00" "060338581803 - 00" "060338581803 - 00" ...
 $ periodo  :int  1 1 1 1 1 1 1 1 1 1 ...
 $ localidad: chr  "General Paz" "General Paz" "General Paz" "General Paz" ...> head(aten.itu.gral)
afiliado                    periodo        localidad
1 060346001 - 00            1                       D09
2 06924406 - 00             1                       D09
3 068581803 - 00            1                       D09
4 06081803 - 00             1                       D09
5 0653357801 - 16           1                       D09
```

Resumen de los datos sobre atenciones en las agencias D09 y D12.

Total, de solicitudes realizadas en las agencias de D09 y D12: 11.203

Solicitudes agencias D09: 4.188 (**40 de afiliados procedentes de D02**).

Solicitudes agencias D12: 7.015 (**343 afiliados procedentes de D22**).

Cantidad total de afiliados D02: 292

Cantidad total de afiliados D22: 1.065

Lo anterior demuestra que durante todo el 2019, se acercaron a la agencia de D09 40 afiliados de D02 (la agencia más cerca a ese dpto.) y en D12 1.065 personas de D22.

Siguiendo el análisis de esta zona se estudió los 511 casos registrados para el otorgamiento del subsidio PADYF[7] dentro del sistema informático SII (Sistema Interactivo de Información). En la Fig. 45, se visualiza su estructuras y datos.

Fecha de carga	Nro de incidencia	Nro de solicitud	Estado	Apellido y Nombre afiliado	Edad	Sexo	Contexto	Tipo de evaluación
05/11/2019	642901	799526	PRESTACION CANCELADA		16	MASCULINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
02/07/2019	580387	727651	SOLICITUD CANCELADA		0	FEMENINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
30/12/2019	1261761	1421940	SOLICITUD CANCELADA		0	MASCULINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
27/12/2019	1261177	1421332	SOLICITUD CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
27/12/2019	1261021	1421154	PRESTACION CANCELADA		0	MASCULINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
10/12/2019	1254569	1413732	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
10/12/2019	1254517	1413674	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
10/12/2019	1254418	1413566	PRESTACION CANCELADA		0	MASCULINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
10/12/2019	1254334	1413473	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
09/12/2019	1253555	1412591	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
05/12/2019	1252751	1411670	SOLICITUD CANCELADA		0	FEMENINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
28/11/2019	1246537	1404881	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
25/11/2019	652258	810315	PRESTACION CANCELADA		0	MASCULINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
13/11/2019	648112	805431	SOLICITUD CANCELADA		0	FEMENINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
11/11/2019	646139	803193	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
07/11/2019	644753	801609	PRESTACION CANCELADA		0	FEMENINO	VISITA DOMICILIARIA	GESTION OTORGAMIENTO PRESTACION SOCIAL
07/11/2019	644613	801459	PRESTACION CANCELADA		0	FEMENINO	UGL-AGENCIA	GESTION OTORGAMIENTO PRESTACION SOCIAL

Fig. 45. Dataset con los subsidios PADyF.

Este programa otorgado por la obra social es un indicador interesante porque se aplica específicamente a la población más vulnerable de afiliados. Al poder acceder a estos datos, estamos realizando una trazabilidad del trabajo desarrollado en territorio por el instituto.

Los datos adquiridos fueron resumidos en la Tabla 17, añadiendo los datos del INDEC sobre índice de dependencia.

[7] Programa de asistencia integral a la dependencia y fragilidad, es un programa asistencial de PAMI.

Tabla 17. Otorgamiento subsidios de apoyo a la vulnerabilidad

localidad	casos	afiliado	índice de dependencia
D03	170	37.824	11,1
D10	90	11.949	15
D24	69	3.491	9,6
D26	67	2.728	9,6
D08	26	1.008	15,9
D09	22	1.962	17
D12	16	3.097	10,7
D17	13	5.813	11,2
D28	10	1.495	14,7
D15	7	5.617	13,5
D07	5	4.116	15,2
D27	4	1.819	11,4
D01	3	4.772	12,2
D19	3	1.629	12,6
D05	2	7.222	15,3
D18	2	2.505	13,5

Fuente: elaboración propia(2023)

Con la misma metodología se trató los 503 registros obtenidos del SII en lo concerniente al otorgamiento del subsidio PAS (Programa asistencia Socio - Sanitaria) el resultado se plasma en la Tabla 18.

Tabla 18. Subsidios económicos.

Localidad	subsidios	afiliados
D03	164	37.824
D10	100	11.949
D24	77	3.491
D26	68	2.728
D08	54	1.008
D12	15	3.097
D09	8	1.962
D12	5	5.813
D01	4	4.772
D27	2	1.819
D28	2	1.495
D16	2	5.457
D19	1	1.609
D07	1	4.116

Fuente: elaboración propia(2023)

El analfabetismo, además de limitar el pleno desarrollo de las personas y su participación en la sociedad, tiene repercusiones durante todo su ciclo vital, afectando el entorno familiar, restringiendo el acceso a los beneficios del desarrollo y obstaculizando el goce de otros derechos humanos, [24]. En términos de salud hablamos de personas que pueden tener dificultades para leer y completar formularios médicos, entender las indicaciones de los profesionales de la salud, o navegar por el sistema de salud en general. Esto puede resultar en retrasos en la atención médica o en la falta de seguimiento adecuado de tratamientos. Con una implicancia en mayores riesgos de enfermedades prevenibles o en un manejo inadecuado de condiciones crónicas.

[25] señala otro aspecto que incide en la salud de las personas adultas mayores, que es "su dependencia". Esto puede tener un impacto significativo, tanto física como emocional. Por ejemplo, un aumento en el riesgo de deterioro físico. La dependencia puede llevar a una disminución en la movilidad y la autonomía de los adultos mayores, lo que a su vez puede aumentar el riesgo de caídas y lesiones. La falta de actividad física contribuye a la pérdida de masa muscular, disminución de la fuerza y problemas de salud como la osteoporosis.

Las personas que dependen de otros para sus cuidados pueden ser más vulnerables a enfermedades infecciosas y crónicas. La falta de autonomía puede dificultar la adopción de hábitos saludables y el seguimiento de tratamientos médicos.

La salud mental no escapa de su impacto, esto se puede ver reflejado en sentimientos de pérdida de control, baja autoestima, ansiedad, depresión y aislamiento social. La falta de independencia y autonomía puede afectar negativamente la calidad de vida y el bienestar emocional.

En la Fig. 42 se presenta a modo de resumen las variables/factores analizados.

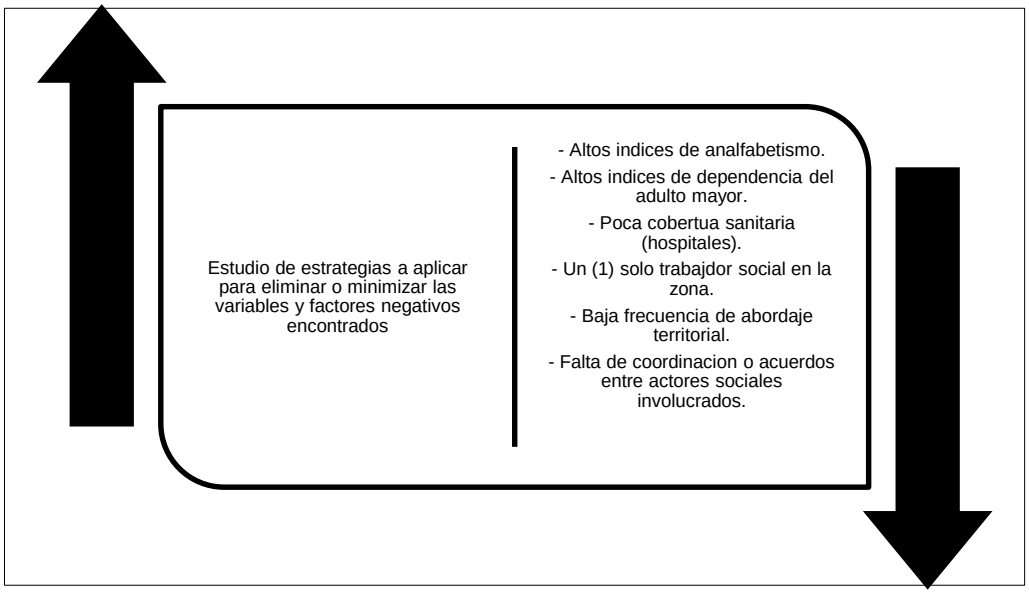


Fig. 46. Factores detectados en la región de estudio

A partir de todo lo expuesto surgen interrogantes ¿Por qué el abordaje en esta zona presenta una marcada diferencia en relación con otros puntos de la provincia? Principalmente teniendo en cuenta los índices de analfabetismo, dependencia y vulnerabilidad, que se hallaron.

¿Qué sucede con los afiliados de departamentos como D02 y D22, donde no existen agencias de la obra social? ¿Dónde se atienden? Hay escasa cobertura sanitaria en la zona. No se pudo determinar el número de internaciones porque no existen registros que nos permitiera caracterizar el territorio. Un solo trabajador social en la zona.

¿Qué rol juegan o no otros actores de la sociedad como centro de jubilados, municipios, clubes, hospitales, entre otros?

Para obtener respuestas que aclaren esta situación, se llevaron a cabo entrevistas con el personal del área social del instituto. Desde su análisis nos indicaron varios puntos:

- ✓ A diferencia de otros puntos de la provincia donde el afiliado es más demandante, por ejemplo, exigen visitas de profesionales, actividades recreativas, bajadas de programas en el territorio, entre otros pedidos, el perfil del ciudadano de esta región en general es más "conformista", menos exigente.
- ✓ Falta de coordinación o acuerdos entre actores de la sociedad (obra social, municipio, centro de jubilados, etc.)

La evidencia mostrada y las respuestas proporcionadas abren la posibilidad para futuros análisis y trabajos. No obstante, se cumplió con el objetivo de detectar zonas vulnerables en la provincia. Lugares donde es preciso aumentar el acceso a la seguridad social, mejorar los servicios de salud, apoyo social, y fortalecer la infraestructura comunitaria. Si bien se focalizo la región mencionada existen otras zonas de la provincia a prestar atención y las cuales deben ser abordadas.
Además, es preciso remarcar la utilidad de este tipo de mapas que generan nuevo conocimiento a la obra social.

5.2 ¿Existe un patrón o tendencia en la distribución espacial de las patologías?

Para dar respuesta a esta pregunta de investigación se aplicó K-NN. Estudiando en detalle los valores obtenidos se desprende lo siguiente:

- Distancia media observada: 1803.18 es la distancia promedio entre cada punto y su vecino más cercano en tus datos. En este caso, la distancia media observada es de aproximadamente 1803.18 unidades de medida.

- Distancia media esperada: 9280.56 esta es la distancia promedio que se esperaría entre cada punto y su vecino más cercano si los puntos estuvieran distribuidos de manera aleatoria. En este caso, la distancia media esperada es de aproximadamente 9280.56 unidades de medida.

- Índice de vecino más cercano: 0.194 este valor nos indica la relación entre la distancia observada y la distancia esperada. Un valor de índice de vecino más cercano inferior a 1 (como en este caso) sugiere que los puntos están más agrupados de lo que se esperaría al azar.

- Número de puntos: 371: este valor indica la cantidad total de puntos en nuestro dataset.

- Puntaje Z: -29.69: el puntaje Z se utiliza para determinar si la distribución de los puntos es significativamente diferente de una distribución aleatoria. Un puntaje Z tan negativo indica una fuerte agrupación de puntos, mucho más de lo que se esperaría al azar.

En resumen, los valores indican que los puntos están significativamente más agrupados de lo que se esperaría al azar, como lo sugiere el índice de vecino más cercano de 0.194 y el puntaje Z negativo de -29.69. **Podemos confirmar de que se trata de una distribución clusterizada.**

Se sugiere como futuro trabajo, continuar con la exploración visual de la distribución de puntos en los diferentes mapas en busca de posibles razones detrás del agrupamiento observado.

5.3 ¿Existen agrupaciones de interés?

En principio podemos responder afirmativamente a esta pregunta, atento a los valores arrojados por K-Means. Se visualizó a las localidades que se agrupan como las vulnerables siendo ellas: D03, D10, D24 y D26.

```
> # Mostrar las medias por cluster
>print(cluster_means)
  cluster media_casosmedia_afiliadomedia_indice_dependencia
1       1        4.50       3.646875                 13.550
2       2       19.25        2.970000                 13.700
3       3       99.00       13.998000                 11.325
```

Basándonos en las medias de las variables en cada cluster, podemos ver que el Cluster 3 tiene el mayor número medio de casos (99.00), la menor afiliación media a la seguridad social (13.998) y el índice de dependencia medio más bajo (11.325) en comparación con los otros clusters. Esto nos sugiere que las localidades en el **Cluster 3 son las más vulnerables en términos de casos, afiliación a la seguridad social e índice de dependencia.**

```
# Obtener los nombres de las localidades correspondientes al Cluster 3 en el dataframe original 'datos'
nombres_cluster3 <- datos$localidad[cluster3_localidades]

# Mostrar los nombres de las localidades en el Cluster 3
print(nombres_cluster3)
```

```
> # Mostrar los nombres de las localidades en el Cluster 3

>print(nombres_cluster3)

[1] "D03" "D10"      "D24"          "D26"
```

Este resultado obtenido a partir de K-means **no son definitivos**, ya que se ha detectado ciertas anomalías o particularidades en los números que requieren un análisis más profundo.

Para ejemplificar tomemos los datos de la agencia D05 y D26.

D05	2	7.222	15,3
D26	67	2.728	9,60

En una agencia como D05 con 7.222 afiliados y índice importante de dependencia del adulto mayor se tiene solamente 2 casos, mientras que en la agencia D26 con 2.728 afiliados existen 67 casos. ¿Hay menos demanda? o ¿poco abordaje? ¿Existen otras explicaciones?

Clustering por atributo: edad

En la etapa anterior se obtuvo esta agrupación según las edades de los afiliados, Fig. 40. Este dato resulta importante tenerlo en cuanto al momento de la planificación de los viajes al interior de la provincia y el tipo de programas que se pueden "bajar" en terreno.

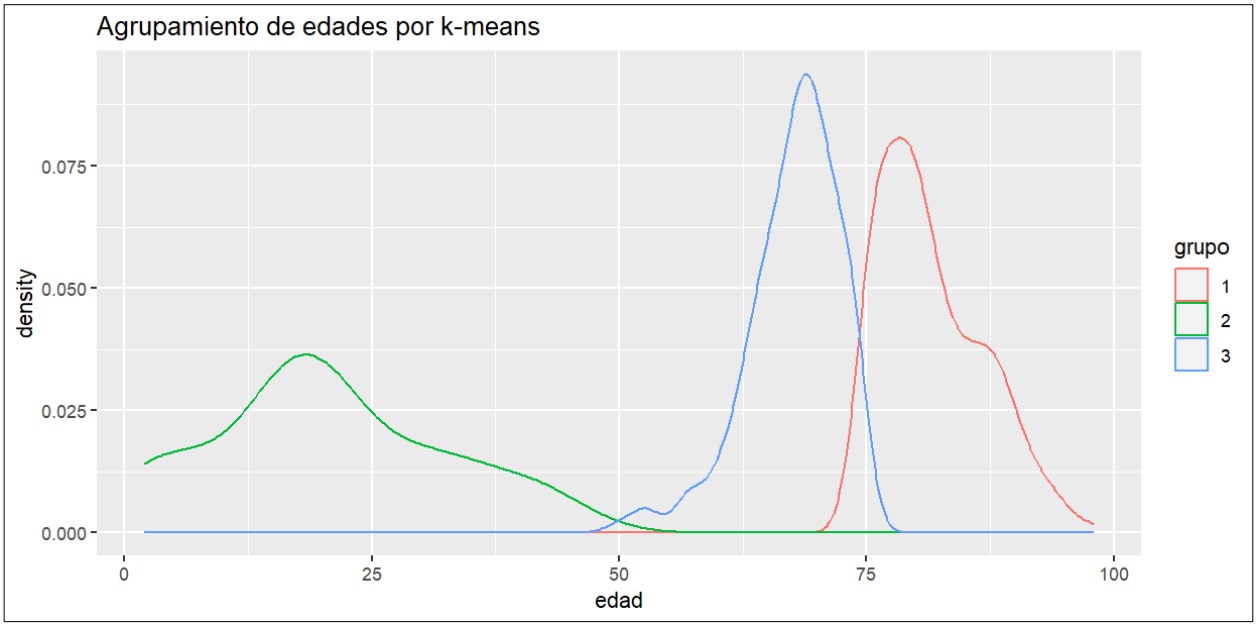


Fig. 40. Gráfico de agrupación según la edad del afiliado.

Este conocimiento es potencialmente útil en la asignación eficiente de los recursos de atención médica en función de la agrupación de patologías y edades identificadas por k-means.

Las enfermedades no transmisibles en afiliados de la obra social representan un 40% aproximadamente, un valor significativo dentro del universo de patologías que afectan al ser humano

y se corresponden con los datos sobre defunciones de la provincia de Corrientes (Dirección de Planificación y Estadísticas de Salud) ver Tabla 19.

Se sabe que las enfermedades tienen diferentes características en función de la edad del paciente. La prevalencia varía en función de la edad y el contexto clínico. Estos clusters encontrados deberían ser tenidos en consideración en próximas planificaciones de programas destinados a la prevención y tratamiento farmacológicos de enfermedades no transmisibles.

Tabla 19. Defunciones en la Prov. de Ctes (2019)

C A U S A S	2019 Nº	%
TOTAL DEFUNCIONES GENERALES	**7290**	**100,0**
A. TOTAL DE CAUSAS DEFINIDAS	**6214**	**85,2**
1.- ENFERMEDADES INFECCIOSAS Y PARASITARIAS (A00-B99)	**203**	**2,8**
1.1 Enfermedades infecciosas intestinales (A00-A09)	5	0,1
1.2 Tuberculosis, inclusive secuelas (A15-A19)	20	0,3
1.3 Hepatitis virales (B15-B19)	0	0,0
1.4 Septicemia (A41)	128	1,8
1.5 Enfermedad por virus de la inmunodeficiencia (B20-B24)	35	0,5
2.-TUMORES (C00-D48)	**1501**	**20,6**
2.1 Malignos	**1355**	**18,6**
2.1.1 Estómago (C16)	86	1,2
2.1.2 Colon (C18)	178	2,4
2.1.3 Páncreas (C25)	81	1,1
2.1.4 Demás órganos digestivos y del peritoneo	144	2,0
2.1.5 Tráquea, de los bronquios y del pulmón (C33, C34)	195	2,7
2.1.6 Mama (C50)	98	1,3
2.2 Carcinoma in situ, tumores benignos y de comportamiento incierto	**146**	**2,0**
o desconocido (D00-D48)		
3.- ENF. ENDOCRINAS, NUTRICIONALES Y METABOLICAS (E00-E90)	**351**	**7020,0**
3.1 Diabetes Mellitus (E10-E14)	299	5980,0
3.2 Desnutrición (E40-E46)	14	280,0
3.3 Resto de enfermedades endocrinas, nutricionales y metabólicas	38	760,0
4.- ENF. DEL SISTEMA NERVIOSO (G00-G99)	**81**	**1620,0**

C A U S A S	2019 Nº	%
4.1 Meningitis (G00-G03)	3	60,0
4.2 Enfermedad de Alzheimer (G30)	21	420,0
4.3 Resto de enfermedades del sistema nervioso	57	1140,0
5.- TRASTORNOS MENTALES Y DEL COMPORTAMIENTO (F01-F99)	**85**	**1700,0**
6.- ENFERMEDADES DEL SISTEMA CIRCULATORIO (I00-I99)	**1371**	**3636,3**
6.1 Enfermedades hipertensivas (I10-I15)	181	3620,0
6.2 Enfermedades isquémicas del corazón (I20-I25)	286	3,9
6.3 Insuficiencia cardíaca ((I50)	156	2,1
6.4 Las demás enfermedades del corazón (I01-I09)(I26-I49)(I51-I52)	204	2,8
6.5 Enfermedades cerebrovasculares (I60-I69)	500	6,9
7.-ENFERMEDADES DEL SISTEMA RESPIRATORIO (J00 - J99)	**1283**	**17,6**
7.1 Infecciones respiratorias agudas (J00-J22)	767	10,5
7.2 Enfermedades Crónicas de las vías respiratorias inferiores (J40-J47)	179	2,5
7.3 Las demás enfermedades del sistema respiratorio	337	6740,0
8.- ENFERMEDADES DEL SISTEMA DIGESTIVO (K00-K93)	**343**	**6860,0**
8.1 Ciertas enfermedades crónicas del hígado y cirrosis(K70);(K76)(K73-K74)	105	2100,0
8.2 Apendicitis, Hernia cavidad abdominal y Obst. intestinal(K35-K46) y K56	11	220,0
8.3 Resto de enfermedades del sistema digestivo	227	4540,0
9.- ENF. DEL SISTEMA GENITOURINARIO (N00-N98)	**154**	**3080,0**
10.-EMBARAZO, PARTO Y PUERPERIO (O00-O99)	**8**	**160,0**
11.-CIERTAS AFECCIONES ORIGINADAS EN EL PERIODO PERINATAL (P00-P96)	**121**	**1,7**
12.-MALFORMACIONES CONGENITAS,DEFORMIDADES Y ANOMALIAS CROMOSOMICAS (Q00-Q99)	**79**	**1,1**
13.-CAUSAS EXTERNAS (V01-Y98)	**483**	**6,6**
13.1 Accidentes de transporte (V01-V99)	87	1,2
13.2 Otras causas externas de traumatismos accidentales (W00-X59)	191	2,6
13.3 Lesiones autoinflingidas intencionalmente (X60-X84)	111	1,5
14.-DEMAS CAUSAS DEFINIDAS	**151**	**2,1**
B. TOTAL CAUSAS MAL DEFINIDAS Y DESCONOCIDAS (R00-R99)	**1076**	**14,8**
1. Signos, síntomas y afecciones mal definidas y desconocidas	1076	14,8
(Inconsistencias e incongruencias)		

Fuente: Ministerio de Salud Pública de Corrientes

5.4 ¿Se pueden encontrar relaciones entre las variables estudiadas?

Según los resultados del ANOVA presentados en la tabla de resumen, ***no se encontraron diferencias significativas en las edades entre los grupos de diferentes sexos, diagnósticos ni en la interacción entre sexo y diagnóstico***.

Esto nos obliga a repensar el modelo, con otras variables que puedan aportar un mejor entendimiento de la realidad que está bajo análisis. Se disponen de más atributos prometedores los cuales no fueron trabajados en esta investigación por diferentes motivos por estar fuera de nuestro alcance, no obstante, se presentan como estímulos para futuros trabajos.

Recomendaciones

Establecidas las conclusiones de esta investigación, se recomienda:

A luz de la evidencia presentada en datos, la cual contrasta con los objetivos que debe perseguir toda intervención en territorio, *se sugiere disponer de este tipo de información durante el proceso de planificación de los abordajes*. Es decir, contar con información actualizada de las distintas variables e indicadores sobre el territorio a operar. Los cual permitirá una mejor asignación de los recursos materiales, financieros y humanos del Instituto, evitando desvíos innecesarios.

Trabajos futuros

A partir de la investigación llevada a cabo quedan diversas líneas de investigación abiertas y en las que es posible continuar trabajando.

Algunas de ellas son el análisis más detallado de los clusters detectados. Creemos que con la incorporación de nuevas variables significativas pueden ayudar describir mejor subconjunto de poblaciones, descubrir patrones o tendencias, etc.

Otra línea que surgió durante la investigación está en relación con la aplicación de algoritmos de regresión. Pueden ser útiles para la predicción de reingresos hospitalarios. Durante este tiempo se ha ido revisando material teórico, analizado investigaciones que desarrollaron distintas herramientas tratando este tema en particular. Dada el número, más que interesante de que se dispone e incorporando por ejemplo variables económicas se pueden obtener resultados atrayentes para las gerencias.

Para todos los trabajos futuros es esencial, el trabajo en territorio realizado por los trabajadores sociales. Son parte de lo que en epidemiologia se llama sistema de vigilancia, tomando datos, aportando información y conocimientos.

Referencias

[1] E. K. Cromley y S. L. McLafferty, ***"Spatial Data," on GIS and Public Health.*** Guilford Press, 2002.

[2] D. Li, S. Wang y D. Li, ***"Methods and Techniques in SDM," on Spatial Data Mining: Theory and Application.*** Springer, 2018.

[3] A. B. Lawson, ***Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology***. Taylor Francis Group, 2013.

[4] P. J. Diggle y E. Giorgi, ***Model-Based Geostatistics for Global Public Health: Methods and Applications***. Taylor Francis Group, 2019.

[5] H. Hütt Herrera y O. Hernández Cruz, ***"Participación Ciudadana: un nuevo paradigma en la gestión pública"***, Estud. Gest., n.º 15, pp. 79–99, diciembre de 2023. Accedido el 10 de marzo de 2024. [En línea]. Disponible: https://doi.org/10.32719/25506641.2024.15.4

[6]S. Canale, H. De Ponti y M. Monteferrario, ***"Obras sociales provinciales: indicadores de consumo y gasto en atención médica",*** Doc. Aportes En Adm. Publica Gest. Estatal, n.º 26, pp. 225–250, agosto de 2016. Accedido el 10 de marzo de 2024. [En línea]. Disponible: https://doi.org/10.14409/da.v16i26.5941

[7] Logicalis Spain. ***"Minería de datos y calidad de vida: data mining en el sector salud".*** Tips de Logicalis. Accedido el 10 de marzo de 2024. [En línea].
Disponible: https://blog.es.logicalis.com/analytics/mineria-de-datos-y-calidad-de-vida-data-mining-en-el-sector-salud

[8] G. Cuaya-Simbro, E. Ruiz, A. Muñoz-Meléndez y E. F. Morales, ***"Análisis de readmisión hospitalaria de pacientes diabéticos mediante aprendizaje computacional"***, Res. Comput. Sci., vol. 138, n.º 1, pp. 147–157, diciembre de 2017. Accedido el 10 de marzo de 2024. [En línea]. Disponible: https://doi.org/10.13053/rcs-138-1-15

[9] E. Rodríguez Jiménez, "***Asignación de recursos a áreas de salud. Entre las propuestas, lo posible y lo necesario",*** vol. 14, p. 7, 2006.

[10] J. M. Fernández y C. Minuesa, ***Estadística básica para ciencias de la salud***, Ed. Universidad de Extremadura. Servicio de Publicaciones, Cáceres, España, 2018

[11] R. G. Henao, ***Introducción a la Geoestadística. Teoría y Aplicación***. Bogotá: Fac. Ciencia. Dpto. Estadística. Univ. Nac. Colombia. [En línea]. Disponible: https://geoinnova.org/wp-content/uploads/2021/08/LIBRO_-DE-_GEOESTADISTICA-R-Giraldo.pdf

[12] P. J. Diggle, ***"Point Process Methods in Spatial Epidemiology," on Statistical Analysis of Spatial and Spatio-Temporal Point Patterns***. Taylor Francis Group, 2013.

[13] C. H. Rotela, ***Epidemiología panorámica: Introducción al uso de herramientas geoespaciales aplicadas a la salud pública.*** Ciudad Autónoma de Buenos Aires: Comisión. Nacional. Actividades. Espaciales Dirección Epidemiologia, 2014.

[14] OPS ***Sistemas de información geográfica en salud. Conceptos básicos***. Washington: OPS, 2002.

[15] I. Del Bosque González y C. Fernández Freire, ***Los sistemas de información geográfica y la investigación en ciencias humanas y sociales***. Madrid: Confed. Española Cent. Estud. Locales (CSIC), 2012.

[16] P. Rigaux, ***Spatial Databases: With Applications to GIS.*** San Francisco: Morgan Kaufmann, 2002.

[17] J. C. Martínez LLario, ***PostGIS análisis espacial avanzado***, Ed. CreateSpace Independent Publishing Platform, 2018.

[18] ***"Introducción a la planificación en salud",*** Centro Redes Unidad Asociada al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, 2022.

[19] ***"Introducción a la Epidemiología",*** Centro Redes Unidad Asociada al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, 2022.

[20] E. Aránguez Ruiz y M. Arribas García, ***Salud y territorio. Aplicaciones prácticas de los sistemas de información geográfica para la salud ambiental***. Madrid: Soc. Española Sanid. Ambiente., 2012.

[21] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, "***CRISP-DM 1.0", Step-by-step data mining guide***, 2000

[22] L. Flores, ***"Integración de procesos de explotación de información y tecnología GIS: Aplicación para el hallazgo de patrones de robos y hurtos de la Ciudad de Corrientes",*** *Tesis de maestría, Fac. Cs. Exactas Naturales y Agrimensura, Corrientes*, 2021

[23] Centro Latinoamericano de Administración para el Desarrollo, ***Declaración-Lisboa-IIS-Innovación Publica***, 2023

[24] R. Martínez y A. Fernández, ***"Impacto social y económico del analfabetismo: modelo de análisis y estudio piloto"***, Comisión Económica para América Latina y el Caribe (CEPAL), 2010

[25] T. Duran Badillo, C. J. Domínguez Chávez, ***Dejar de ser o hacer: significado de dependencia funcional para el adulto mayor***, Acta Universitaria, 28(3), 40-46. doi10.15174/au.2018.1614 2018

[26] M. Belló y V. M. Becerril-Montekio. ***Sistema de salud de Argentina***. Salud Publica Mex, 53, 13, 2011

[27] V. Olaya, **"La calidad de los datos espaciales," en Sistemas de Información Geográfica,** 2014

Anexos

ANEXO N° 1. Sistema de salud argentino

El sistema de salud de Argentina está compuesto por tres sectores: público, seguro social y privado. El sector público está integrado por los ministerios nacionales y provinciales, y la red de hospitales y centros de salud públicos que prestan atención gratuita a toda persona que lo demande, fundamentalmente a personas sin seguridad social y sin capacidad de pago. Este sector se financia con recursos fiscales y recibe pagos ocasionales del sistema de seguridad social cuando atiende a sus afiliados.

El sector del seguro social obligatorio está organizado en torno a las Obras Sociales (OS), que aseguran y prestan servicios a los trabajadores y sus familias. Además, el Instituto Nacional de Servicios Sociales para Jubilados y Pensionados brinda cobertura a los jubilados del sistema nacional de previsión y sus familias. Las provincias cuentan con una OS que cubre a los empleados públicos de su jurisdicción. La mayoría de las OS operan a través de contratos con prestadores privados y se financian con contribuciones de los trabajadores y los patrones.

El sector privado está conformado por profesionales de la salud y establecimientos que atienden a demandantes individuales, y a los beneficiarios de las OS y de los seguros privados. Este sector también incluye entidades de seguro voluntario llamadas Empresas de Medicina Prepaga que se financian con primas que pagan las familias o las empresas y con recursos derivados de contratos con las OS. Los servicios que ofrecen se prestan en consultorios e instalaciones privados.

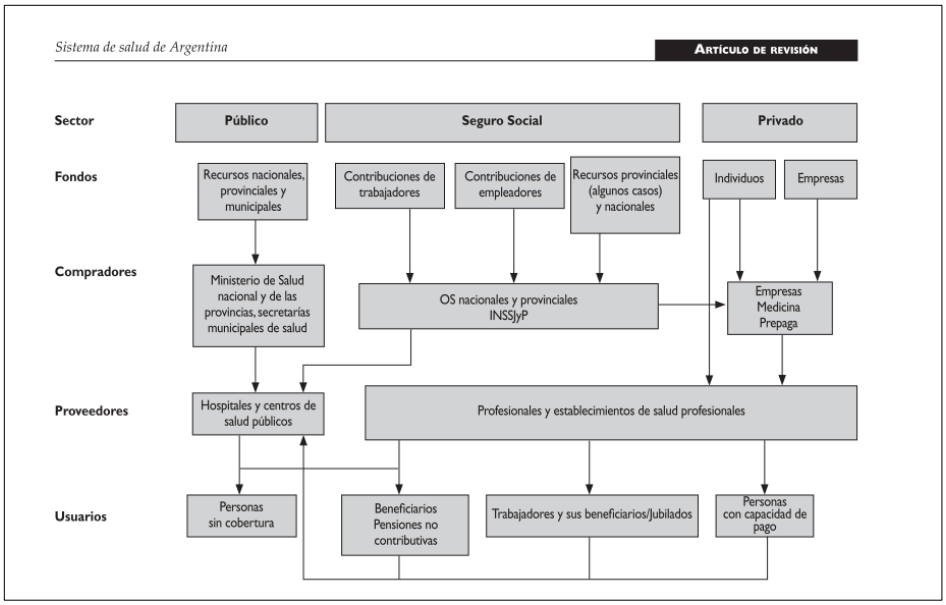


Fig. 47. Estructura del sistema de salud argentino. Fuente: [26]

ANEXO N° 2. Tipos de métodos de investigación.

En diferentes libros y trabajos que abordan la aplicación y desarrollo de los métodos de investigación, se han propuesto diversos esquemas para agrupar y caracterizar a los distintos tipos de estudio, los cuales se han clasificado de acuerdo con:

Según la estructura en: Metodologías cualitativas o Metodologías cuantitativas

- El tipo de asignación de la exposición o variable en estudio: Experimentales u Observacionales (no experimentales).
- Según la finalidad o búsqueda de causalidad: Descriptivos o Analíticos
- El número de mediciones que se realiza en cada sujeto de estudio para verificar la ocurrencia del: Transversales o Longitudinales.
- Según el grado de intervención: Controlados / No controlados.
- En función de la variable independiente: Simple / Factorial.
- Según la temporalidad del inicio de la exposición o de la ocurrencia del evento: Retrospectivos o Prospectivos.
- Según la unidad de análisis donde se mide el evento en estudio: Basados en individuos o Estudios ecológicos (o conglomerados).

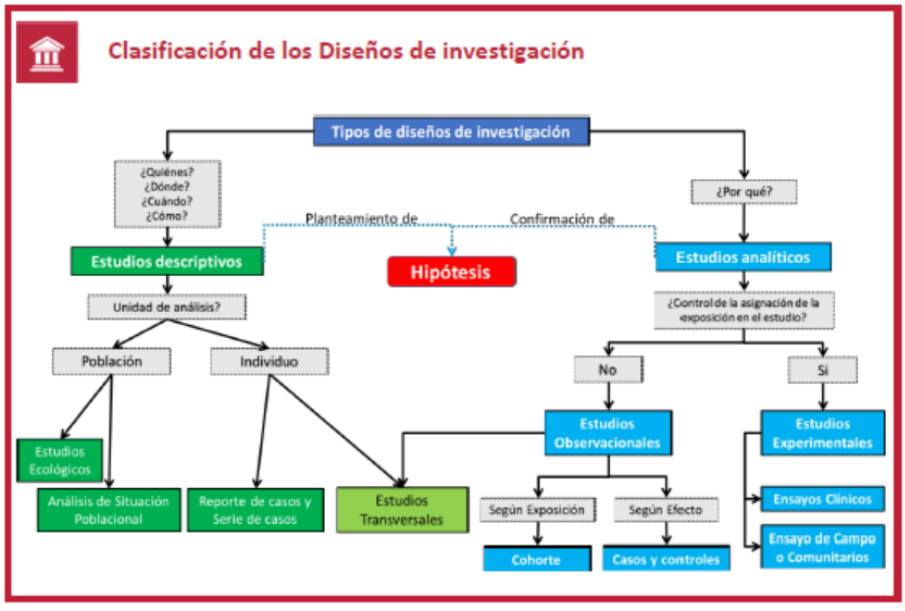


Fig. 48. Tipos de investigaciones. Fuente: [10]

ANEXO N° 3. La calidad de los datos espaciales

En este anexo se presenta información adicional relacionada con la calidad de los datos, específicamente el dato espacial. Siendo la calidad el conjunto de propiedades y de características de un producto o servicio que le confieren su aptitud para satisfacer unas necesidades explícitas e implícitas.

Sabemos por definición, que ningún dato es perfecto. Todo dato que utilicemos va a contener errores, y estos pueden ser desde totalmente irrelevantes para el desarrollo de un proceso de análisis hasta de tal magnitud que desvirtúen por completo los resultados de dicho análisis.

Destacar que no solo es importante contar con datos de calidad en los que los errores sean mínimos, sino conocer el tipo de error que existe en nuestros datos y la magnitud de estos.

A pesar de su gran importancia, la calidad de los datos espaciales no ha sido una preocupación hasta hace relativamente poco tiempo.

Conceptos y definiciones sobre calidad de datos.

El concepto básico es el error, que no es sino la discrepancia existente entre el valor real (puede ser un valor de posición, de un atributo, o cualquier otro), y el valor recogido en una capa. El error puede ser de dos tipos: sistemático y aleatorio.

Dos términos importantes en el estudio de la calidad son la precisión y exactitud.

La precisión indica el nivel de detalle con el que se recoge la información. Una capa en la que las posiciones se han medido con 5 valores decimales es más precisa que una en la que se han medido con un único decimal.

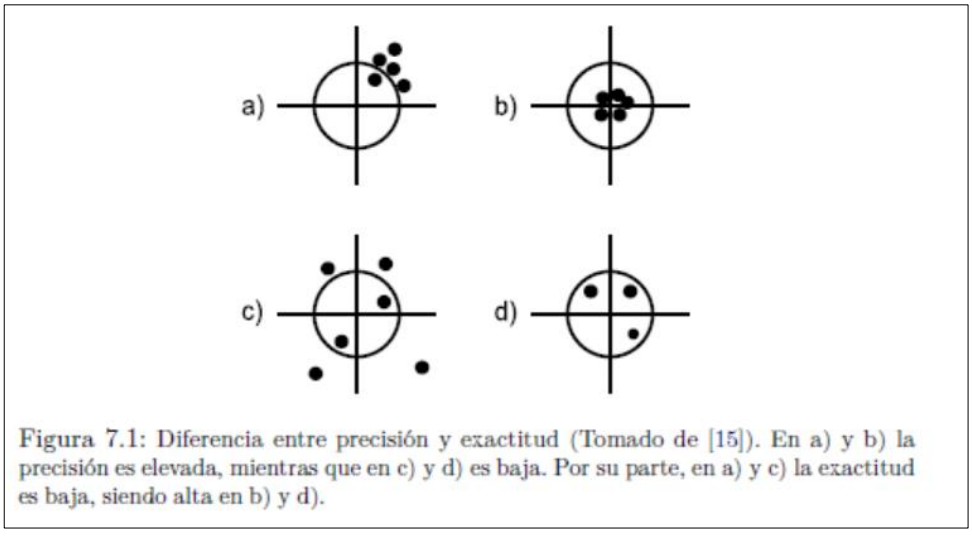


Figura 7.1: Diferencia entre precisión y exactitud (Tomado de [15]). En a) y b) la precisión es elevada, mientras que en c) y d) es baja. Por su parte, en a) y c) la exactitud es baja, siendo alta en b) y d).

Fig. 49. Diferencia entre precisión y exactitud. Fuente: [27]

Dependiendo del uso que se pretenda dar a una capa de datos geográficos, se requerirá una u otra precisión. Un trabajo geodésico requerirá medir la localización de un punto con precisión milimétrica, mientras que para un muestreo para inventario forestal es suficiente localizar las parcelas correspondientes con una precisión mucho menor.

Fuentes y tipos de errores

Errores de concepto y modelo: al recoger la información espacial utilizamos algún modelo de representación (ráster, vectorial), el cual siempre tiene alguna deficiencia. La realidad y las tareas que pretendemos realizar con una capa de información espacial no se adaptan por completo a ninguno de los modelos de representación, y el hecho de optar por uno u otro conlleva la introducción de algún error, o condiciona para la aparición de unos u otros errores en las etapas posteriores.

Errores en las fuentes primarias: el dato vectorial del que disponemos proviene originariamente de una fuente primaria, la cual puede contener errores. Si esta fuente contiene errores, estos aparecerán también en los datos que se deriven de este. Así, si digitalizamos en base a un mapa escaneado y la hoja original es errónea, también lo serán las capas que creemos en esa digitalización.

Errores en los procesos de creación de la capa: los procesos que realizamos para crear la capa pueden incorporar errores en el resultado. Por ejemplo, en el proceso de digitalización en base a ese mapa escaneado pueden aparecer errores por razones tales como un mal trabajo del operario, ya sea al digitalizar las entidades sobre una tableta o al teclear los valores de los atributos. Otros procesos, como pueden ser los de conversión entre los modelos ráster y vectorial, también pueden tener como consecuencia la aparición de errores.

Errores en los procesos de análisis: un dato espacial puede derivar de un proceso de análisis, y en él pueden aparecer errores debidos principalmente a dos razones: o bien la capa original objeto de análisis contiene de por sí errores, o bien el proceso no es por completo correcto.

Las componentes de la calidad

Algunos de los componentes principales de la calidad del dato espacial:

Exactitud posicional: todo dato espacial tiene asociada una referencia geográfica. La precisión con la que se toma esta condiciona la calidad del dato.

Exactitud en los atributos: si la componente espacial puede tener errores, estos también pueden aparecer en la componente temática.

Consistencia lógica y coherencia topológica: los datos espaciales no son elementos independientes, sino que existen relaciones entre ellos. Un dato de calidad debe recoger fielmente estas relaciones, siendo la topología la encargada de reflejar este tipo de información.

Compleción. El dato espacial no recoge todo lo que existe en una zona dada.

Calidad temporal. Aunque los datos espaciales son «imágenes» estáticas de la realidad, el tiempo es importante en muchos sentidos, pues afecta directamente a su calidad. El paso del tiempo puede degradar la calidad del dato espacial en mayor o menor medida.

Procedencia. Un dato espacial puede provenir de una fuente más o menos fiable, o haber sido generado a través de uno o varios procesos, en cada uno de los cuales se puede haber introducido algún tipo de error.

Detección y gestión de errores

Detectar los errores puede realizarse de forma visual o bien de forma analítica, pudiendo automatizarse en este segundo caso. El error medio cuadrático es la medida más habitual del error en el caso de variables cuantitativas, mientras que la matriz de confusión es empleada para variables cualitativas.

Modelar el error y su propagación puede emplearse para conocer de forma más adecuada la validez de los resultados obtenidos a partir de un dato espacial. La realización de simulaciones condicionales mediante el método de Monte Carlo es la técnica más habitual para la modelación de errores.

ANEXO N° 4. Tabla resumen con indicadores del INDEC, en tabla N° 20.

Tabla 20. . Datos obtenidos del INDEC.

Departamento	Índice de envejec. %	Población mayor a 65 años %	Analfab. %	Hogares s/ agua dentro viv. %	Hogares s/ agua dentro viv.	Hogares c/ agua dentro viv.
Bella Vista	28,2	7,4	2,2	8,5	8.201	88.792
Berón de Estrada	18,2	5,9	4,3	18,5	3.081	13.583
Capital, Corrientes	25,1	7,6	4,2	27,4	2.689	7.134
Concepción	23,8	7,9	5,6	35,6	2.005	3.633
Curuzú Cuatiá	22,2	6,9	3,6	13,6	1.805	11.449
Empedrado	30,4	9,7	7,5	38,4	969	1.555
Esquina	20,2	6,5	4,6	24,9	2.082	6.285
General Alvear	32	9,4	8,1	39,8	1.757	2.660
General Paz	26,1	8,7	7,7	37,3	1.747	2.937
Goya	31,5	9,3	6	20	2.493	9.953
Itatí	31,7	9,6	5,7	19	453	1.934
Ituzaingó	20,8	7,7	7,5	47,7	1.219	1.334
Lavalle	17,4	6,5	9,4	46,4	2.414	2.791
Mburucuyá	31,2	9,2	5,4	27,2	6.625	17.719
Mercedes	28,5	8	4,5	25,3	984	2.898
Monte Caseros	29,9	9	6,5	36,5	1.479	2.576
Paso de los Libres	33,9	10,2	5,9	36,2	1.493	2.626
Saladas	27,8	8,9	6,4	40,4	3.275	4.833
San Cosme	33,1	10,2	7,2	37,6	254	422
San Luis del Palmar	27,1	8,3	5,4	14,3	1.599	9.573
San Martín	32,6	10,3	8,6	31,1	783	1.731
San Miguel	25,9	8,6	6,5	23,1	846	2.815
San Roque	19,1	6,6	6,5	49,2	3.594	3.718
Santo Tomé	30,2	9	5,7	27,3	671	1.785
Sauce	32,7	9,2	3,3	11	1.164	9.379

Fuente: elaboración propia (2024)

ANEXO N° 5. Código de nubes de palabras

```
### Autor: VALLEJOS, LEONARDO A.
### WordClouds
### Análisis de la información contenida en la variable diagnostico
##### Configurar espacio de trabajo #####
setwd("C:/nubes")
# Instalación de los paquetes
install.packages("webshot2", repos ="https://cran.r-project.org/web/packages/webshot2/index.html")

install.packages("pdftools", repos ="http://cran.us.r-project.org")
install.packages("tm", repos ="http://cran.us.r-project.org")
install.packages("SnowballC", repos ="http://cran.us.r-project.org")

# paquete para tareas de minería de texto
install.packages("wordcloud2", repos ="http://cran.us.r-project.org")

# Cargar las libreriaswordcloud y RColorBrewer
library("wordcloud2")
library("SnowballC")
library("tm")
library("pdftools")
library("webshot2")


# importar el texto a analizar. Para ello se usaremos la función pdf_text()
texto <-pdf_text("C:/nubes/diagnostico.pdf")

# el español lleva tildes y otros signos de puntuación usamos la función iconv() para quitarlos y no tener pr
oblemas más adelante.
texto <-iconv(texto, "UTF-8")
texto =iconv(texto, to="ASCII//TRANSLIT")
# Utilizando la función Corpus(), indicamos la fuente de nuestro texto
docs<-Corpus(VectorSource(texto))
# Para verificar que el archivo se cargó correctamente, utilizamos la función
# inspect()
inspect(docs)
toSpace<-content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs<-tm_map(docs, toSpace, "\r\n")

docs <-tm_map(docs, toSpace, "/")
docs <-tm_map(docs, toSpace, "@")
docs <-tm_map(docs, toSpace, "\\|")
```

```
####### Limpieza del texto #####

# Convertir el texto a minúsculas
docs<-tm_map(docs, content_transformer(tolower))

# Quitar los números
docs<-tm_map(docs, removeNumbers)

# Quitar las palabras comunes en español
docs<-tm_map(docs, removeWords, stopwords("spanish"))

# Quitar palabras communes que consideres bajo tu criterio
# especificando un vector de palabras comunes a ser eliminadas
docs<-tm_map(docs, removeWords, c("palabrascomunes1", "palabrascomunes2"))

# Quitar signos de puntuación
docs<-tm_map(docs, removePunctuation)

# Eliminar espacios en blanco
docs<-tm_map(docs, stripWhitespace)

###### Construirunamatriz term-document #####
mtd<-TermDocumentMatrix(docs)
m <-as.matrix(mtd)
v <-sort(rowSums(m),decreasing=TRUE)
d <-data.frame(word =names(v),freq=v)

##### Generar la nube de palabras #####
wordcloud2(data = d, size =0.5, shape ="cloud",color="random-dark", ellipticity =0.5)
```

ANEXO N° 6. El test de Shapiro-Wilk

El test de Shapiro-Wilk es una prueba de normalidad que se utiliza para determinar si un conjunto de datos sigue una distribución normal. En RStudio, se puede interpretar los resultados del test de Shapiro-Wilk de la siguiente manera:

1. Valor p: El resultado más importante del test de Shapiro-Wilk es el valor p. Si el valor p es mayor que el nivel de significancia (usualmente 0.05), entonces no se rechaza la hipótesis nula y se concluye que los datos se distribuyen normalmente. Por ejemplo, si el valor p es 0.07 y el nivel de significancia es 0.05, se acepta la hipótesis nula de normalidad.

2. Estadístico de prueba: Además del valor p, el test de Shapiro-Wilk también proporciona un estadístico de prueba. Este valor es útil para evaluar la magnitud de la desviación de la normalidad, pero la interpretación principal se basa en el valor p.

ANEXO N° 7. Regiones de la provincia de Corrientes.

Debido a sus distintas particularidades económicas, climáticas y territoriales, fue necesario establecer una regionalización de la Provincia. Estas son:

Región 1. Capital: Corrientes; Riachuelo

Región 2. Tierra Colorada: Alvear; Gobernador Ingeniero Valentín Virasoro; Santo Tomé; Ituzaingó; La Cruz; Colonia Carlos Pellegrini; Colonia Liebig's; Estación Torrent; Garruchos; Guaviraví; José Rafael Gómez; San Antonio de Apipé; San Carlos; Tapebicuá; Villa Olivari; Yapeyú

Región 3. Centro Sur: Bonpland; Curuzú Cuatiá; Mercedes; Monte Caseros; Mocoretá; Paso de los Libres; Parada Pucheta; Sauce; Tapebicuá; Colonia Libertad; Felipe Yofre; Juan Pujol; Mariano I. Loza; Perugorría Región 4. Rio Santa Lucia: Bella Vista; 3 de Abril; Esquina; Goya; Santa Lucía; Chavarría; Colonia Carolina; Colonia Pando; Cruz de los Milagros; Gobernador Martínez; Lavalle; 9 de Julio; Pedro R. Fernández; Pueblo Libertador; San Roque; Yataytí Calle

Región 5. Humedal: Saladas; Colonia Santa Rosa; Concepción; Empedrado; Mburucuyá; Pago de los Deseos; San Miguel; Loreto; San Lorenzo; Tabay; Tatacuá

Región 6. Noroeste: Berón de Estrada; Caá Catí; Itatí; Paso de la Patria; San Cosme; San Luis del Palmar; Herlitzka; ItáIbaté; Lomas de Vallejos; Palmar Grande; Ramada Paso; Santa Ana de los Guácaras.

Fig. 50. División en regiones de la Prov. de Ctes[8]. Fuente: información extraída del Ministerio del Interior

[8] Según la ley 5960; y Dto. 143/11 art 39, encontramos en la Provincia de Corrientes 6 regiones.

ANEXO N° 8. Tablas de departamentos con su correspondiente código alfanumérico como proceso de anonimización, ver Tabla 21

Tabla 21. Tabla de Dptos. anonimizados. Fuente: elaboración propia (2024)

Departamento	código
Bella Vista, Corrientes	D01
Berón de Estrada, Corrientes	D02
Capital, Corrientes	D03
Concepción, Corrientes	D04
Curuzú Cuatia, Corrientes	D05
Empedrado, Corrientes	D06
Esquina, Corrientes	D07
General Alvear, Corrientes	D08
General Paz, Corrientes	D09
Goya, Corrientes	D10
Itatí, Corrientes	D11
Ituzaingó, Corrientes	D12
Lavalle, Corrientes	D13
Mburucuyá, Corrientes	D14
Mercedes, Corrientes	D15
Monte Caseros, Corrientes	D16
Paso de los Libres, Corrientes	D17
Saladas, Corrientes	D18
San Cosme, Corrientes	D19
San Luis del Palmar, Corrientes	D20
San Martín, Corrientes	D21
San Miguel, Corrientes	D22
San Roque, Corrientes	D23
Santo Tomé, Corrientes	D24
Sauce, Corrientes	D25
Gobernador Virasoro	D26
Santa Rosa	D27
La Cruz	D28