**UNIVERSIDAD PRIVADA “FRANZ TAMAYO”**

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA DE SISTEMAS**

# “ANÁLISIS DE DATOS CENSALES PARA DETERMINAR LA UBICACIÓN MAS OPTIMA Y NECESARIA DE CENTROS EDUCATIVOS”

**ESTUDIANTE:** Gabriela Micaela
Durán Villafán

**ASIGNATURA:** Big Data

**DOCENTE:** Ing. Enrique Laurel

**LA PAZ – BOLIVIA**

**2025**

## Marco conceptual

### 1. Problema y motivación

El objetivo es identificar lugares óptimos para ubicar centros educativos **(**nuevas escuelas, ampliaciones, o redistribución) basándose en características sociodemográficas y espaciales del censo (población por edad, densidad poblacional, distancia a escuelas existentes, accesibilidad por red vial, vulnerabilidad socioeconómica, etc.). Usar datos censales permite estimar demanda potencial y necesidades territoriales con granularidad (unidades censales, manzanas o coordenadas agregadas).

### 2. Conceptos y teorías relevantes

#### Concepto de Análisis Espacial

El análisis espacial es una metodología que combina el análisis multicriterio con los Sistemas de Información Geográfica (SIG), constituyendo una herramienta poderosa que posibilita la integración, análisis, síntesis y difusión de conocimiento. Esta metodología permite procesar grandes cantidades de información precisa a bajo costo mediante la disponibilidad de programas informáticos gratuitos y de código abierto, combinada con la abundancia de datos georreferenciados.

El análisis espacial en el contexto educativo se define como el proceso de examinar la distribución geográfica de variables educativas para identificar patrones, relaciones y tendencias espaciales que pueden informar la toma de decisiones en planificación educativa. Esta aproximación metodológica permite visualizar desigualdades, anticipar riesgos y planificar infraestructuras de manera más efectiva.

#### Aplicaciones del Análisis Espacial en la Planificación Educativa

El análisis espacial tiene aplicaciones estratégicas fundamentales en la promoción y gestión de la planificación educativa. Cruzar datos del sistema educativo con información georreferenciada permite a quienes son responsables del planeamiento

y la gestión de la educación generar políticas muy contextualizadas que garanticen que el sistema educativo responda a las necesidades de las comunidades locales.

Las principales aplicaciones incluyen: mayor equidad en la distribución de oportunidades educativas, mejor adaptación de estas oportunidades a las necesidades de las comunidades locales, y un uso más eficiente de todos los recursos disponibles. Para determinar el mejor lugar y el momento indicado para construir una escuela, o donde se creen programas o servicios educativos que tengan en cuenta las limitaciones geográficas, las problemáticas sociales y las realidades económicas locales, el análisis espacial proporciona un marco metodológico esencial.

El microplaneamiento resulta cada vez más pertinente, preciso y receptivo mediante el uso de datos georreferenciados, y las políticas educativas se adaptan mejor a las necesidades y los contextos locales. Las metodologías desarrolladas permiten clasificar unidades espaciales mediante la utilización de indicadores de planificación a través de puntajes estandarizados, aplicados específicamente a escuelas.

**Indicadores Geoespaciales para la Detección de Carencias Educativas**

Los indicadores geoespaciales son herramientas fundamentales para identificar y cuantificar las carencias educativas en un territorio. El análisis de la distribución espacial de los resultados junto a la distribución de otras variables permite la obtención de estudios espaciales útiles para la planificación urbana en la búsqueda de eficiencia y equidad espacial.

Las regresiones ponderadas geográficamente permiten obtener mapas de interacciones entre la variable educativa dependiente y las variables explicativas, lo que permite a quienes son responsables de las políticas tener más información sobre dónde aplicar políticas específicas con prioridad. Múltiples variables explicativas influyen en los resultados educativos en diversas partes de un territorio, y es crucial para quienes planifican determinar la fuerza de estas interacciones y su localización.

Los sistemas de información geográfica permiten la representación del sistema educativo mediante un conjunto cartográfico con diversos mapas elaborados a partir de variables educativas, demográficas, ambientales, socioculturales y económicas, cuyas explicaciones e interrelaciones dan como resultado un diagnóstico de la localización actual de la oferta y demanda educativas.

**Importancia de la Georreferenciación en Estudios Sociales**

La georreferenciación es fundamental en los estudios sociales porque permite localizar y analizar fenómenos sociales en su contexto espacial específico. El planeamiento y la gestión de la educación son siempre una cuestión de contexto, por lo que determinar la ubicación precisa de escuelas, estudiantes y recursos es esencial para la planificación efectiva.

Los mapas temáticos, como herramientas cartográficas georreferenciadas, permiten representar muy diverso tipo de información localizada en el territorio. Estos mapas facilitan la visualización de desigualdades sociales, permiten anticipar riesgos, ayudan a planificar infraestructuras y revelan patrones espaciales que de otro modo serían difíciles de identificar.

La optimización de las rutas de inspección para mejorar la calidad de la educación es posible mediante el uso de datos georreferenciados junto con métodos como el problema del viajante y árboles de expansión mínima, determinando la distribución óptima de las escuelas entre los equipos de inspección y optimizando las rutas para maximizar la cantidad de escuelas visitadas a lo largo de un año escolar.

**Descripción General del Censo 2012**

El Censo Nacional de Población y Vivienda de Bolivia de 2012 (CNPV 2012) se realizó el 21 de noviembre de 2012, día que fue declarado feriado a nivel nacional. Este censo fue realizado por el Instituto Nacional de Estadística de Bolivia (INE) y constituyó históricamente el decimoprimer censo de población y el quinto censo de vivienda en toda la Historia de Bolivia.

El censo logró demostrar que en Bolivia vivían alrededor de 10.059.856 personas. Con respecto al anterior censo 2001, la población boliviana había crecido en un 21,5%. El costo económico del censo fue de 55,5 millones de dólares (380,7 millones de bolivianos). La población total censada incluyó 9.827.089 personas residiendo en viviendas particulares y 205.751 en viviendas colectivas.

El Decreto Supremo 1305, del 1 de agosto de 2012, declaró prioridad nacional el Censo Nacional de Población y Vivienda 2012, reconociendo la importancia fundamental de esta operación estadística para el conocimiento y análisis de la evolución en la composición, crecimiento y distribución de la población.

**Relevancia del Censo para Estudios Socioeducativos**

Los datos del Censo 2012 son fundamentales para conocer y analizar la evolución en la composición, crecimiento y distribución de la población, siendo esenciales para estudios socioeducativos. El censo constituye un importante insumo para definir políticas y programas adecuados a la realidad de las regiones, posibilitando el cálculo de indicadores de forma exhaustiva para todo el territorio nacional.

Aunque los censos no tienen la periodicidad de las encuestas de hogares, posibilitan el cálculo de indicadores tanto para la zona urbana como para el área rural, permitiendo establecer cuál es la distribución regional de diferentes fenómenos sociales, cuál es su distribución a nivel de departamento, municipio, y qué factores o atributos los determinan.

El Censo 2012 registró un total de 3.159.350 viviendas en el país, de las cuales 3.125.168 corresponden a viviendas particulares y 34.182 a viviendas colectivas, representando un incremento significativo respecto al Censo 2001. El número de viviendas particulares por cada mil habitantes pasó de 279,1 en 2001 a 318,0 en 2012, reflejando cambios importantes en la estructura demográfica y habitacional del país.

**Variables Censales Relacionadas con Educación y Población Escolar**

El Censo incluyó variables fundamentales relacionadas con la educación y la población escolar. Entre las variables más relevantes se encuentran: tasa de asistencia escolar de la población de 6 a 19 años de edad, nivel de instrucción alcanzado, alfabetismo, y características de acceso a servicios educativos.

Los resultados mostraron que el 87,3% de la población de 6 a 19 años de edad asiste a una unidad educativa, con variaciones departamentales significativas. En cuanto a la tasa de escolaridad de la población femenina de 6 a 19 años de edad, se registró un incremento de 9 puntos porcentuales entre el período intercensal.

El Censo también recopiló información sobre población en edad escolar por edades simples, lo que permite un análisis detallado de la demanda educativa potencial. De la población de 6 a 19 años que trabaja, el 60,1% asiste a la escuela o colegio, siendo las niñas de 6 a 11 años el segmento poblacional que registra un mayor porcentaje con 91,5%, seguido de los niños de 6 a 11 años con 85,7%.

**Limitaciones y Potencialidades de los Datos Censales**

Los datos censales presentan tanto limitaciones como potencialidades. Entre las limitaciones, se encuentra la falta de periodicidad comparada con las encuestas de hogares, lo que puede generar desactualización de la información en períodos intercensales largos. Además, el censo tiene limitaciones inherentes en la profundidad con la que puede abordar ciertos temas específicos, privilegiando la cobertura universal sobre el detalle temático.

Sin embargo, las potencialidades son significativas. El censo permite la construcción de indicadores exhaustivos para todo el territorio nacional, posibilitando análisis a nivel agregado total nacional, tanto para la zona urbana como para el área rural. Permite establecer la distribución regional de diferentes fenómenos, su distribución a nivel de departamento, municipio, y para las principales ciudades a nivel de localidad o comuna.

Los datos censales son fundamentales para la medición y monitoreo de diferentes dimensiones del desarrollo social, incluyendo la pobreza y la desigualdad tanto en

el nivel nacional como en los niveles regional y local, y apoyan la formulación y evaluación de políticas sociales.

## 5. Variables y Criterios para Determinar Déficit Educativo

### Población en Edad Escolar

La población en edad escolar se refiere a la población que tiene la edad teórica para cursar los niveles educativos establecidos en el sistema educativo. En el contexto boliviano, esto comprende: i) Inicial en familia comunitaria escolarizada (de 4 a 5 años), ii) Primaria comunitaria vocacional (de 6 a 11 años), iii) Secundaria comunitaria productiva (de 12 a 17 años).

Según proyecciones de población para 2017, los niños/as de 0 a 11 años de edad llegaron a 2.906.000 habitantes, con 50,9% hombres y 49,1% mujeres. La mayor cantidad de población infantil se concentra en el rango de edad de seis años, con 243.200 niños aproximadamente. A nivel nacional, la proporción de la población en el rango de edad escolar (4 a 23 años) ha experimentado una disminución constante desde 2005, pasando del 45,6% al 36,7%.

En todo sistema educativo, es importante conocer la magnitud de la población escolar y su comportamiento para el diseño de políticas educativas. La tasa de cobertura neta, que se refiere al número de estudiantes en el grupo de edad esperada que se encuentran inscritos para el nivel que corresponde a su edad, brinda una mejor medida del acceso escolar. En Bolivia, aproximadamente 5 de cada 10 niños inicia la primaria en la edad que le corresponde.

### Número y Distribución de Unidades Educativas

La distribución de unidades educativas en Bolivia presenta patrones territoriales significativos. La población estudiantil perteneciente a primaria es mayor que la población del nivel secundario, por lo tanto la cantidad de infraestructura de unidades educativas públicas es destinada para el funcionamiento de la formación del nivel primario.

La mayor cantidad de Unidades Educativas se encuentra en La Paz, Santa Cruz, Potosí y Cochabamba, correlacionándose directamente con la mayor población de estudiantes y cantidad de docentes. Por la dispersión de la población escolar en el área rural, las unidades educativas rurales atienden a una menor cantidad de alumnos que las urbanas, lo que implica mayores costos per cápita y desafíos particulares de gestión.

El Sistema de Información Educativa del Ministerio de Educación proporciona información detallada sobre la concentración y distribución espacial de unidades educativas en todo el territorio nacional. Esta información es fundamental para identificar áreas con déficit de infraestructura educativa y planificar expansiones del sistema.

**Relación entre Población Escolar y Oferta Educativa**

La relación entre población escolar y oferta educativa es un indicador fundamental para determinar el déficit educativo. Esta relación considera no solo la cantidad de establecimientos disponibles, sino también su capacidad, calidad y pertinencia para atender las necesidades de la población en edad escolar.

La oferta educativa está en íntima vinculación con la demanda actual y potencial. Para el total de América Latina, casi 9% de la población con edad de asistir a la educación básica (6 a 17 años) no concurre a la escuela. En Bolivia, más del 10% de niños, niñas y adolescentes en edad escolar están fuera de la escuela.

La concentración de la oferta privada en las áreas urbanas es significativa, mientras que en las áreas rurales la oferta pública absorbe prácticamente a la totalidad de los estudiantes. El sector privado solo tiene presencia significativa en las áreas urbanas y atiende a alumnos pertenecientes a sectores socioeconómicos más altos, lo que implica que el sector público enfrenta el problema del diseño de dispositivos específicos que contribuyan al buen funcionamiento de las escuelas en diversos contextos sociales, culturales y económicos.

**Factores Socioeconómicos y Demográficos Asociados**

Los factores socioeconómicos y demográficos son determinantes fundamentales del déficit educativo. La desigualdad de ingresos ha generado una brecha considerable en el acceso a oportunidades educativas, ya que las familias de bajos ingresos priorizan la subsistencia diaria sobre la inversión en educación.

Entre los principales factores se encuentran: la pobreza, el desempleo y la inestabilidad laboral de los padres, que son las principales causas del abandono escolar. La teoría de la reproducción social de Bourdieu sostiene que las desigualdades socioeconómicas se perpetúan a través del sistema educativo, y los estudiantes de entornos desfavorecidos enfrentan mayores dificultades para acceder y permanecer en la escuela.

El nivel de instrucción de los padres es un factor crucial, ya que existe una importancia creciente de la educación de los padres en la distribución de oportunidades educativas. La composición familiar, el tamaño de la familia, y la ausencia de los padres en los hogares también son variables significativas que afectan la deserción escolar.

El área geográfica representa una circunstancia importante en el nivel de desigualdad de oportunidades educativas. Las brechas rural-urbanas son sustanciales en Bolivia, con los niveles de oportunidades educativas de los niños, niñas y adolescentes que viven en el área rural siendo mucho menores que sus contrapartes urbanas. Esta creciente desigualdad en las oportunidades educativas podría reducir la acumulación de capital humano en Bolivia a nivel nacional y regional.

## 6. Métodos y Enfoques de Análisis

### Análisis Descriptivo de Datos Educativos

El análisis descriptivo de datos educativos constituye la etapa inicial fundamental del análisis, teniendo como objetivo describir los datos encontrados en una muestra mediante valores característicos y presentarlos en forma de gráfico o tabla. Esta presentación de los datos se refiere a las variables individuales y sus

características, permitiendo resumir los datos recogidos y procesados en tablas, medidas, gráficos significativos.

La estadística descriptiva suele estar al principio del análisis de datos y a menudo se combina con otros métodos. Se utilizan tres medidas principales para describir los datos: medidas de tendencia central (media, mediana, moda), medidas de dispersión (varianza, desviación estándar), y medidas de forma de distribución.

En el contexto educativo, el análisis descriptivo permite examinar en qué medida los resultados obtenidos a partir de una muestra pueden trasladarse a la población real de interés. Antes de poder probar las hipótesis reales, hay que comprobar si la variable de la muestra se distribuye normalmente, por ejemplo, si la proporción de edad o de género refleja la de la población.

## Herramientas Tecnológicas para el Análisis de Datos

### Uso de Python en la Ciencia de Datos

Python se ha consolidado como el lenguaje de programación líder para la ciencia de datos, siendo esencial para manejar, analizar y graficar conjuntos de datos y presentar resultados para la toma de decisiones. Su popularidad se debe a su sintaxis clara, amplia biblioteca de paquetes especializados y capacidad para integrar diferentes herramientas de análisis.

El propósito fundamental de usar Python en ciencia de datos incluye: manejar, analizar y graficar conjuntos de datos; realizar análisis estadísticos con facilidad; limpiar y preprocesar datos para análisis posteriores; extraer información valiosa para impulsar la toma de decisiones; y ahorrar tiempo al automatizar tareas repetitivas de análisis de datos.

Python permite transformar datos en información útil para tomar decisiones, crear modelos de Machine Learning y obtener insights valiosos. Su integración con múltiples bibliotecas especializadas lo convierte en una herramienta versátil que puede abordar desde análisis estadísticos básicos hasta complejos modelos de aprendizaje automático.

## Aplicación de Librerías: Pandas, NumPy y Matplotlib

**NumPy** es la biblioteca fundamental para el cálculo científico en Python, proporcionando funciones para crear y manipular arrays multidimensionales, así como herramientas para realizar operaciones matriciales y estadísticas. NumPy es esencial para trabajar con grandes conjuntos de datos de manera eficiente, permitiendo realizar operaciones matemáticas y estadísticas de alto rendimiento.

**Pandas** está construido sobre NumPy, facilitando la manipulación y el análisis de datos de forma rápida y sencilla, siendo ideal para trabajar con datos tabulares como los vistos en hojas de cálculo o bases de datos. Pandas proporciona estructuras de datos de alto rendimiento y fáciles de usar, así como herramientas de análisis de datos. La estructura de datos principal de Pandas es el DataFrame, que es similar a una tabla en una base de datos relacional o una hoja de cálculo de Excel.

Entre las principales funciones de Pandas destacan: capacidad de manejar datos faltantes, operaciones de agregación, filtrado y ordenamiento, y realizar cálculos y transformaciones. Pandas permite la carga de datos desde diferentes fuentes, facilitando la limpieza, transformación y análisis de datos complejos.

**Matplotlib** es la librería de referencia para la visualización de datos en Python, permitiendo transformar datos en gráficos e imágenes que comunican claramente los hallazgos. Tanto NumPy como Pandas se integran perfectamente con Matplotlib, permitiendo crear gráficos desde arrays de NumPy o directamente desde DataFrames de Pandas.

El flujo de trabajo completo integra estas tres bibliotecas: se utiliza Pandas para cargar y estructurar datos desde diversas fuentes; NumPy para realizar cálculos intensivos, como operaciones matriciales y simulaciones; y finalmente Matplotlib para visualizar los resultados de los análisis, creando gráficos que pueden comunicarse efectivamente con la audiencia.

## Entornos de Trabajo: Miniconda y Jupyter Lab

**Miniconda** es una versión minimalista de Anaconda que incluye solo lo esencial: Python y Conda (el gestor de paquetes y entornos). La idea detrás de Miniconda es que se empiece con una base ligera y luego se instalen solo las bibliotecas que se necesiten, sin el "peso extra" de todas las herramientas preinstaladas que vienen con Anaconda.

Miniconda proporciona un entorno de línea de comandos simplificado para administrar paquetes y crear entornos virtuales. La creación de entornos virtuales permite tener diferentes "espacios de trabajo" para proyectos distintos, con sus propias versiones de Python y paquetes instalados. Esto es fundamental cuando se trabaja con proyectos de ciencia de datos, ya que permite usar versiones específicas de los paquetes o del propio runtime de Python.

**Jupyter Lab** es una interfaz de usuario basada en una aplicación web para el manejo de notebooks, que también permite trabajar con editores de texto, consolas y componentes personalizados, integrando en una sola interfaz todos los elementos implicados en el análisis de datos. Los notebooks son documentos interactivos en los que se puede integrar texto, código ejecutable en diversos lenguajes de programación, así como tablas o figuras.

La interfaz de JupyterLab consta de un área de trabajo con diseño basado en pestañas que permite combinar distintos documentos y herramientas de codificación en paneles de pestañas que se pueden redimensionar o subdividir mediante arrastrar y soltar. Este sistema permite una programación más integrada y eficiente, ya que los desarrolladores pueden realizar la mayoría de sus tareas sin salir de JupyterLab.

## Visualización y Comunicación de Resultados

### Representación Gráfica de los Datos Educativos

La representación gráfica de datos educativos es fundamental para comunicar efectivamente los resultados del análisis. Un gráfico estadístico es una representación visual de una serie de datos estadísticos que capta la atención, facilita la comprensión y permite identificar patrones y tendencias.

Los gráficos estadísticos más utilizados en contextos educativos incluyen: gráficos de barras (simples y múltiples), gráficos de líneas, pictogramas, diagramas de puntos, y gráficos de sectores. Se observa el predominio del gráfico de barras en los libros de texto educativos, lo que coincide con investigaciones en diferentes países y con las recomendaciones establecidas en las orientaciones curriculares.

La complejidad semiótica de los gráficos puede clasificarse en diferentes niveles: nivel 1 (lectura de datos simples), nivel 2 (representación de relaciones entre variables), nivel 3 (representación de una distribución), y nivel 4 (análisis de relaciones complejas). La mayoría de los gráficos educativos corresponden al nivel 3, representación de una distribución, lo que facilita el análisis e interpretación de la información que comunican los datos.

**Mapas Temáticos y Gráficos de Distribución**

Los mapas temáticos son herramientas cartográficas basadas en otros tipos de mapas, como los topográficos o políticos, para representar sobre ellos fenómenos o hechos geográficos o sociales pertenecientes a una categoría específica. Estos aspectos son representados mediante símbolos o colores, permitiendo visualizar la distribución espacial de variables educativas.

La cartografía temática puede ser cualitativa o cuantitativa. Los mapas cualitativos muestran la distribución espacial o la situación de un grupo de datos nominales, mientras que los mapas cuantitativos muestran aspectos espaciales de datos numéricos. Esta distinción es fundamental para seleccionar el tipo apropiado de representación según el objetivo del análisis.

Los componentes de un mapa temático incluyen: una base geográfica (mapa base) y una capa de contenido temático. El usuario de un mapa temático debe ser capaz de integrarlas, visual e intelectualmente, durante la lectura del mapa. La simbología utilizada puede incluir símbolos de datos puntuales (identifican características y su situación), símbolos de datos lineales (representan detalles que tienen una línea definida), y símbolos de datos superficiales (proporcionan información sobre una cierta superfcie).

Los mapas temáticos facilitan la visualización de desigualdades sociales, permiten anticipar riesgos, ayudan a planificar infraestructuras y revelan patrones espaciales. Su objetivo no es la descripción exhaustiva del espacio físico, sino la comunicación de información analítica a partir de datos georreferenciados, aplicando criterios de clasificación, generalización y diseño visual adecuados a la escala, el propósito y el público destinatario.

**Justificación del enfoque aplicado**

**1) Exploratorio y Descriptivo**

**Qué es (en 1 frase):** mirar y resumir los datos para entender qué hay — dónde vive la gente, cuántos niños hay, qué zonas parecen desatendidas, y si faltan o hay errores en los datos.

- **¿Por qué se eligió?**
  Porque antes de proponer ubicaciones tienes que conocer la realidad: detectar errores, ver patrones espaciales y priorizar zonas. Sin esto cualquier modelo dará malas respuestas.
- **Herramientas recomendadas y por qué:**
  - **Python**: pandas para tablas, geopandas para datos espaciales, matplotlib para gráficos y mapas.
  - JupyterLab para explorar de forma interactiva (gráficos + código). Estas herramientas son fáciles de usar, muy usadas en análisis y permiten ver resultados rápido.
- **Arquitectura / cómo organizarlo:**
  - Carpeta data_raw/ (datos originales) y data_processed/ (versiones limpias).
  - Notebooks en notebooks/ para gráficos y descubrimientos.
  - Scripts en src/transform.py para las limpiezas reproducibles.

- Control desde main.py para ejecutar transformaciones automáticas.

- **Garantizar reproducibilidad / escalabilidad / claridad:**
  - Guardar pasos de limpieza como scripts (no solo acciones manuales en el notebook).
  - Registrar versiones de datos (ej. censo2012_v1.parquet).
  - Usar environment.yml para reproducir el entorno y documentar cada variable.
  - Si crece el volumen, cambiar a Parquet y/o Dask para procesar en paralelo.

**2) Optimización espacial (prescriptivo)**

**Qué es (en 1 frase):** usar modelos matemáticos para proponer ubicaciones de escuelas que maximicen cobertura o minimicen distancia de la población objetivo.

- **¿Por qué se eligió?**
  Porque además de saber dónde hay necesidad, queremos **tomar decisiones concretas**: dónde poner escuelas nuevas para beneficiar a más niños con el menor costo/distancia.
- **Herramientas recomendadas y por qué:**
  - **Python**: pulp / ortools para formular y resolver modelos (p-median, MCLP).
  - geopandas + shapely para operaciones espaciales (buffers, intersecciones).
  - osmnx/networkx si se quiere distancia por red vial (más realista). Estas librerías permiten tanto calcular cobertura espacial como resolver la optimización.
- **Arquitectura / cómo organizarlo:**

- Módulo src/model_opt.py que reciba datos procesados (puntos de demanda, candidatos de ubicación, capacidades) y devuelva soluciones.
- Guardar matrices de cobertura precomputadas (quién queda dentro de cada buffer) en archivos para acelerar pruebas.
- main.py --run model_opt --k 5 para ejecutar con parámetros diferentes.

- **Garantizar reproducibilidad / escalabilidad / claridad:**
  - Parametrizar el modelo (k, umbral de distancia, weights) en un config.yaml.
  - Versionar resultados con sello de tiempo y parámetros usados.
  - Tests para confirmar que la matriz de cobertura se construye correctamente.
  - Si los cálculos son grandes, usar solver que acepte streaming o dividir el problema por zonas.

**3) Predictivo (opcional)**

**Qué es (en 1 frase):** usar modelos estadísticos o de ML para estimar cómo cambiará la demanda escolar en el futuro (crecimiento, migración).

- **¿Por qué se eligió (o no)?**
  Es útil sólo si necesitas planear a futuro. Con un censo estático (2012) la predicción requiere asumir tasas de crecimiento o usar otras fuentes. Es **opcional** porque agrega incertidumbre si no hay series temporales.
- **Herramientas recomendadas y por qué:**
  - **scikit-learn**, **xgboost** para modelos de regresión/árboles; **statsmodels** para modelos estadísticos.

    - Si hay series temporales, **prophet** o modelos ARIMA. Usar estas librerías porque son estándar, con buenas prácticas y fáciles de validar.

- **Arquitectura / cómo organizarlo:**
    - Módulo src/predict.py que tome datos históricos o supuestos y devuelva escenarios (optimista/medio/pesimista).
    - Mantener los escenarios como archivos separados para alimentar el modelado de ubicación.
- **Garantizar reproducibilidad / escalabilidad / claridad:**
    - Documentar supuestos (tasas de crecimiento, migración) y fuente de datos.
    - Evaluar modelos con métricas claras (MAE, RMSE) y validación cruzada.
    - Versionar modelos y conservar semillas aleatorias para reproducir resultados.

**4) Enfoque por fases (resumen integrador)**

Recomiendo trabajar por etapas claras: **(1) ETL y limpieza → (2) Análisis exploratorio espacial → (3) Modelado de accesibilidad y optimización → (4) Validación y visualización**.

- **¿Por qué esta secuencia?**
  Porque cada fase depende de la anterior: no puedes optimizar bien si los datos están sucios; no tiene sentido predecir sin entender patrones actuales.
- **Herramientas para todo el pipeline:**
    - Python (pandas, geopandas, PySAL, scikit-learn, osmnx, pulp/ortools).
    - Jupyter para EDA; scripts + main.py para producción.

    - Opcional: Docker para mantener el entorno idéntico en diferentes máquinas.

- **Arquitectura propuesta (concreta y simple):**

```
project/
├─ data_raw/
├─ data_processed/
├─ notebooks/
├─ src/
│  ├─ extract.py
│  ├─ transform.py
│  ├─ eda.py
│  ├─ model_opt.py
│  └─ main.py
├─ config.yaml
├─ environment.yml
└─ results/    # soluciones, gráficos, logs
```

**Cómo garantizar reproducibilidad, escalabilidad y claridad (prácticas concretas):**

- **Environment**: environment.yml o Dockerfile para que cualquier equipo instale las mismas librerías.
- **Parámetros**: todo configurable en config.yaml (umbral, k, rutas).
- **Data**: conservar data_raw/ inmutable; todo cambio queda en data_processed/ con nombres/versiones.
- **Código**: funciones en src/ con tests mínimos (pytest). No realizar limpieza sólo en notebooks.
- **Formato**: usar Parquet para grandes tablas y GeoPackage/Shapefiles para capas espaciales.
- **Escalar**: si pandas no alcanza, reemplazar por Dask (mismo API) o procesar por zonas.
- **Trazabilidad**: logs y archivos results/ con timestamp y config para saber cómo se generó cada salida.

**Herramientas propuestas y por qué**

- **Python (JupyterLab, pandas, geopandas, PySAL, scikit-learn, networkx / osmnx, folium/Kepler/plotly):**
    - Pandas: manipulación tabular.
    - GeoPandas + Shapely: manipulación espacial (uniones espaciales, buffers).
    - PySAL / esda: estadística espacial (Moran, LISA).
    - OSMnx / NetworkX: cálculo de rutas y tiempos reales sobre red vial.
    - scikit-learn / xgboost: clustering, regresión si se requiere.
    - JupyterLab es ideal para exploración reproducible y notebooks narrativos.
- **R / QGIS / ArcGIS** también son opciones válidas, pero usando Python mantienes todo integrado en notebooks y scripts reproducibles (además de integrar fácilmente con pipelines).

**Variables para ubicar la población por zona**

| Variable | Significado | Uso principal |
|---|---|---|
| **USUAREA** | Lugar de residencia habitual (zona, barrio o área censal) | Identifica la zona geográfica para agrupar la población |
| **_RESDEPT** | Departamento de residencia habitual | Permite agrupar por departamento |
| **_RESPROV** | Provincia de residencia habitual | Permite agrupar por provincia |

**Variables para contar familias o hogares**

| Variable | Significado | Uso principal |
|---|---|---|
| **_DWNUM** | Número de vivienda | Identifica cada vivienda o hogar |
| **_NSH** | Número de personas en el hogar | Permite conocer el tamaño del hogar |

## Variables para contar población y niños

| Variable | Significado | Uso principal |
|---|---|---|
| **_PERNUM** | Número de persona dentro del hogar | Identifica a cada persona única en la base |
| **_AGE** | Edad | Permite calcular niños, adultos, adultos mayores o rango de edades |
| **_SEX** | Sexo (masculino/femenino) | Para análisis demográficos desagregados por género |

## Bibliografía


- Doug Laney, “3-D Data Management: Controlling Data Volume, Velocity and Variety” (2001) — definición de las 3 V’s. AIIM comunidad
- Gartner, “Definition of Big Data” (glossary). Gartner
- Instituto Nacional de Estadística (INE), Censo Nacional de Población y Vivienda 2024 (Bolivia). INE Bolivia
- Church, R. L., & ReVelle, C. S. (1974). *The Maximal Covering Location Problem.* Papers of the Regional Science Association. (location-allocation). SpringerLink
- Anselin, L. (1995). *Local Indicators of Spatial Association—LISA.* Geographical Analysis (estadística espacial: LISA / Moran's I). Wiley Online Library
- Hansen, W. G. (1959). *How Accessibility Shapes Land Use.* Journal of the American Planning Association (accesibilidad). Taylor & Francis Online