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 EN BOLIVIA”

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

Demanda educativa

Cantidad y características de la población que requiere servicio escolar (por edad, nivel, matrícula esperada). Es el insumo principal para dimensionar y ubicar centros.

Oferta educativa

Conjunto de escuelas existentes, su capacidad y recursos. Permite detectar brechas entre oferta y demanda.

Unidad de análisis (área/segmento)

División espacial usada para el análisis (manzana, sección censal, cuadrícula). Define la resolución de las decisiones.

Densidad poblacional

Número de habitantes por unidad de superficie; indicadores específicos: densidad de población en edad escolar. Ayuda a priorizar zonas con alta concentración de potenciales estudiantes.

Accesibilidad

Facilidad con la que la población llega a un servicio; puede medirse por distancia

euclidiana, distancia por red vial o tiempo de viaje. Es clave para evaluar cobertura real.

Area de influencia
Zona alrededor de una escuela desde la cual se espera que provengan alumnos. Se usa para estimar cobertura.

Áreas de servicio
Técnicas para generar zonas alrededor de puntos (escuelas) en distancia o tiempo; permiten evaluar qué población queda dentro o fuera del servicio.

Distancia en red vs distancia euclidiana
Distancia real medida sobre calles y caminos (más realista) frente a distancia en línea recta (simplificada). La elección afecta estimaciones de accesibilidad.

Problemas de ubicación-asignación
Conjunto de modelos matemáticos que buscan ubicar instalaciones (escuelas) para optimizar cobertura, minimizar distancias o maximizar equidad.

Modelo p-median y MCLP (Maximal Covering Location Problem)
Formulaciones clásicas: p-median minimiza la suma de distancias; MCLP maximiza la demanda cubierta dentro de un umbral. Son opciones para proponer ubicaciones.

Equidad y justicia espacial
Evaluación de si la distribución de escuelas es justa entre diferentes grupos o zonas (por ejemplo, barrios vulnerables vs no vulnerables).

Segmentación y priorización
Clasificación de zonas según criterios (necesidad, densidad, déficit) para priorizar intervención o inversión.

Autocorrelación espacial
Medida de si valores (p. ej. densidad o matrícula) son similares en áreas cercanas; ayuda a identificar clusters y patrones espaciales.

Clustering espacial

Agrupamiento de puntos/áreas con características similares. Herramienta para identificar zonas críticas.

Datos geoespaciales (GIS)

Datos y técnicas para almacenar, procesar y analizar información con referencia espacial (coordenadas, polígonos). Fundamental para el análisis de ubicación.

Calidad de datos

Precisión, cobertura, actualidad y completitud de los datos (p. ej. censo). Condición previa para decisiones confiables.

Integración de fuentes

Unión del censo con otras capas (red vial, escuelas, límites administrativos) para enriquecer el análisis.

Indicadores de desempeño

Métricas para evaluar propuestas: % población cubierta, distancia media a la escuela, capacidad utilizada, índice de equidad.

Escalabilidad y reproducibilidad

Prácticas para que el proceso pueda repetirse y ampliarse (scripts, formatos eficientes, documentación).

Simulación y escenarios

Creación de escenarios alternativos (p. ej. diferentes radios, crecimiento poblacional) para evaluar soluciones bajo supuestos diversos.

Definición de Big Data

Big Data: conjuntos de datos caracterizados por volumen, velocidad y variedad, que requieren técnicas y arquitecturas específicas para su procesamiento eficiente y extracción de valor. En la práctica se define como datos de gran escala o complejidad para los cuales las soluciones tradicionales de gestión y análisis no son suficientes. AIIM comunidad+1

Enlace con tu problema: aunque un único censo (2024) puede no parecer “Big Data” en el sentido de streaming continuo, en la práctica el análisis de **grandes muestras censales (millones de registros)** con atributos espaciales, la integración con otras fuentes (red vial, dispositivos móviles, escolaridad histórica) y la necesidad de procesos reproducibles y escalables justifica el enfoque y herramientas de Big Data (procesamiento por lotes, paralelización, manejo de datos geoespaciales en volumen). Gartner

Justificación del enfoque aplicado

1) Exploratorio + Descriptivo (fase inicial)

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).

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