



Analytics

Análisis de Datos Predicción de tiempos de viaje promedio en Bogotá usando datos de Uber Movement y modelos de machine learning

Caso de estudio: origen en
Santa Inés (centro de Bogotá)

PRESENTADO POR

Juan Sebastian Sanchez Guarnizo

Contexto y problema

En grandes ciudades, la congestión y la variabilidad del tráfico hacen que el tiempo de viaje sea difícil de predecir. Esto afecta la experiencia de los usuarios de plataformas de transporte tipo Uber, la planificación de la movilidad urbana y la toma de decisiones de política pública.

En este proyecto analizamos el tiempo promedio de viaje en Bogotá usando datos agregados de Uber Movement. El problema se formula como una tarea de regresión supervisada: predecir el Mean Travel Time (s) desde un origen fijo (Santa Inés) hacia múltiples zonas destino, evaluando el aporte de variables espaciales e infraestructura vial (OpenStreetMap).



Importancia del Estudio (Resumen)

- Para usuarios: estimar tiempos medios de desplazamiento para planificar viajes y decisiones cotidianas..
- Para plataformas de transporte: soporte para asignación de vehículos, estimación de tiempos y análisis operativo (por ejm: estimación dinamica del precio).
- Para ciudad / planificación:
 - identificar zonas con menor accesibilidad relativa-baja (mayor tiempo medio desde el origen)
 - priorizar intervenciones en infraestructura vial y gestión de movilidad
 - apoyar análisis exploratorios para políticas basadas en datos
- ML aplicado: evaluar qué variables espaciales explican mejor el tiempo medio y qué tanto generaliza el modelo.



Objetivos del Proyecto

- **Objetivo general**

- Modelar y predecir el tiempo medio de viaje desde un origen fijo (Santa Inés) hacia múltiples zonas destino en Bogotá, usando datos agregados de Uber Movement y técnicas de Machine Learning.

- **Objetivos específicos**

- Caracterizar el dataset (estadísticos descriptivos y verificación de calidad: rangos, dispersión, valores faltantes).
- Construir variables (features) predictoras a partir de la geometría
 - centroides origen/destino
 - distancia Haversine (km)
- Extraer un indicador de infraestructura vial desde OpenStreetMap (OSMnx)
 - densidad de calles (m/km^2) por zona destino
- Entrenar y evaluar modelos Random Forest bajo diferentes conjuntos de variables (ablation)
 - M0: sin cotas (solo variables espaciales)
 - M1: con cotas, variables de dispersión del estimado (range_width, rel_width)
- Interpretar resultados con métricas (RMSE, R^2), gráficos Pred vs Real e importancia de variables, discutiendo implicaciones para movilidad urbana.



Datos (principales- base) utilizados

- Fuente: Uber Movement – Travel Times (Bogotá), versión archivada en Kaggle.
- Unidad de observación:
 - un par origen–destino con:
 - Mean Travel Time (Seconds)
 - Range – Lower Bound/Upper Bound (Seconds)
- Configuración específica:
 - Origen fijo: SANTA INÉS, 003107, Movement ID 183 (zona central de Bogotá).
 - Observaciones: 496 pares OD
 - Destinos: 496 zonas diferentes (Movement IDs 4–1160 aprox.).
 - Cobertura temporal: 30/07/2016 – 27/08/2016, promedio diario (Every day, Daily Average, datos agregados).



Estrategia General

CARGA DE DATOS



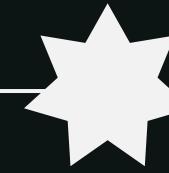
Se realiza la carga del dataset y una limpieza básica (tipos de datos, valores faltantes y consistencia de columnas y registros), asegurando un formato adecuado para el análisis.

INGENIERÍA DE CARACTERÍSTICAS



Se construyen variables predictoras a partir de la geometría: centroides de zonas, distancia Haversine (km) y variables de dispersión del estimado (range_width, rel_width). Además, se calcula densidad vial desde OpenStreetMap (OSMnx) como variable externa.

MODELADO Y EVALUACIÓN



Se entrenan modelos de regresión (Random Forest) bajo distintos conjuntos de variables (ablation). Se evalúa el desempeño con partición train/test 80/20 y métricas RMSE y R². Finalmente se analizan gráficas Pred vs Real e importancia de variables para interpretar resultados.

Ingeniería de características (Feature Engineering)

1) Variables espaciales (geométricas)

- Centroides de origen y destino
 - Calculados a partir de los polígonos de cada zona.
- Distancia Haversine (km)
 - Aproxima la distancia geográfica mínima entre zonas origen–destino.
 - Captura el efecto dominante del recorrido.

2) Variables de incertidumbre del tiempo

- Range width = Upper bound – Lower bound
- Relative width = Range width / Mean travel time

3) Variables de localización

- Latitud y longitud del destino
- Identificador de zona (Destination Movement ID)

4) Variable exógena (infraestructura vial)

- Densidad de calles (m/km^2)
 - Calculada con OSMnx sobre la red vial tipo drive.
 - Longitud total de vías / área del polígono.



Algoritmo y enfoque de modelado

- Modelo principal: Random Forest Regressor (regresión no lineal basada en ensamble de árboles).
- Motivación: captura relaciones no lineales y maneja interacciones entre variables sin suposiciones paramétricas fuertes.
- Diseño experimental: comparación por ablation con diferentes conjuntos de variables (M0/M1 y variantes con/sin densidad vial).
- Evaluación: partición train/test 80/20, métricas RMSE y R².



Diseño experimental (ablation study)

Se evaluaron cuatro configuraciones de modelos para aislar el efecto de distintas variables:

- M0 – Exógeno (sin cotas)
 - Variables espaciales básicas: distancia Haversine, coordenadas origen/destino, ID de destino.
 - No utiliza información de incertidumbre del tiempo de viaje.
- M1 – Base (con cotas)
 - M0 + información estadística del tiempo de viaje:: range width y relative width.
 - Representa un escenario con mayor información histórica.
- M0+ – Exógeno + densidad vial
 - M0 + densidad de calles (m/km^2) del destino.
 - Evalúa si la infraestructura urbana mejora la predicción.
- M1+ – Base + densidad vial
 - M1 + densidad de calles.
 - Evalúa si la infraestructura aporta valor adicional cuando ya existen cotas.



Herramientas y Librerías Utilizadas

PYTHON

Lenguaje de programación

PANDAS Y NUMPY

Análisis de datos

SCIKIT-LEARN

Implementación de modelos de machine learning supervisado (Regresión robusta de Huber y Random Forest), partición de datos y métricas de evaluación (RMSE, R2R^2R2).

OSMNX Y SHAPELY

Extracción y procesamiento de información espacial urbana (red vial), cálculo de densidad de calles y manejo de geometrías

MATPLOTLIB Y SEABORN

Visualización Datos Y Resultados

```
    -> var boolean
    -> */
define('PSI_INTERNAL_XML', false);
if (version_compare("5.1", PHP_VERSION, ">")) {
    die("PHP 5.2 or greater is required!!!");
} if (!extension_loaded("pcre")) {
    die("phpSysInfo requires the pcre extension to php in order to work
properly.");
}
require_once APP_ROOT . '/includes/autoload.inc.php';
// Load configuration
require_once APP_ROOT . '/config.php';
if (!defined('PSI_CONFIG_FILE')) {
    $tpl = new Template("/templates/html/error_config.html");
    echo $tpl->fetch();
    die();
}
if (isset($_GET['language'])) {
    $lang = $_GET['language'];
    if ($lang == "es") {
        header("Content-Type: text/javascript; charset: utf-8");
        echo javascript;
    } else {
        header("Content-Type: text/plain; charset: utf-8");
        echo strtolower;
    }
}
```

Referencia

Shokoohyar, S., Sobhani, A., Malhotra, R., & Liang, W. (2020). Travel time prediction in ride-sourcing networks – A case study for machine learning applications. Journal of Business Cases and Applications, 26. Academic and Business Research Institute (AABRI). <https://www.aabri.com/copyright.html>



Metricas (RMSE y R²)

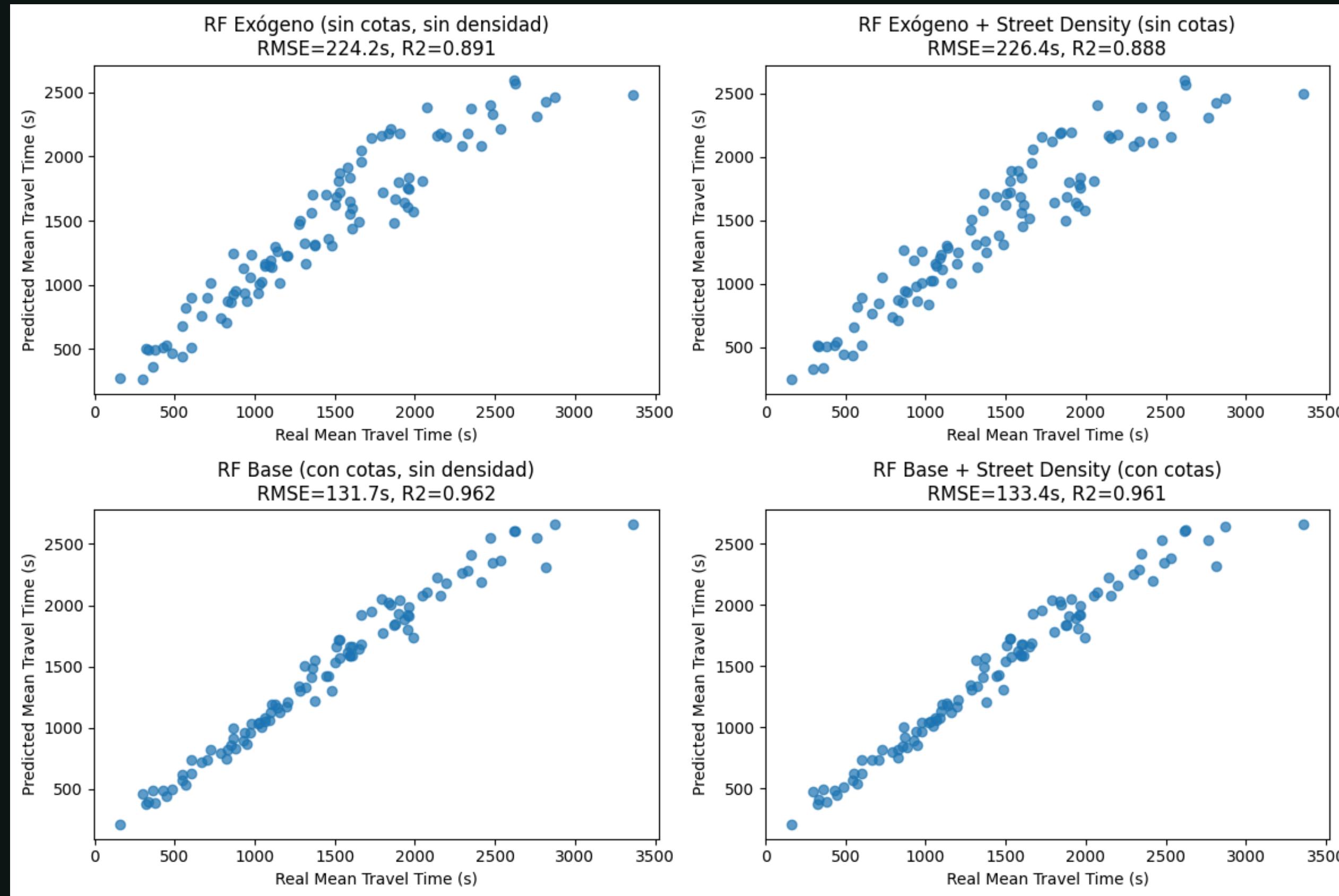
--- RESUMEN RÁPIDO ---

M0	RMSE: 224.17594595645676	R2: 0.8905157274137824
M0+	RMSE: 226.39319895685176	R2: 0.8883392687950741
M1	RMSE: 131.6618222306046	R2: 0.9622346565838217
M1+	RMSE: 133.4464162125466	R2: 0.961203946883614

Incluir “cotas” del tiempo reduce significativamente el error; la densidad vial no aporta mejora relevante en este dataset.



Visualización Predicción vs Real



Importancia de variables en los modelos

```
Top features M0 Exógeno (sin cotas):  
dist_km          0.860883  
dest_lat         0.069105  
dest_lon         0.046675  
Destination Movement ID  0.023338  
orig_lat        0.000000  
orig_lon        0.000000  
dtype: float64
```

```
Top features M0+ Exógeno + Density (sin cotas):  
dist_km          0.857763  
dest_lat         0.065913  
dest_lon         0.042218  
Destination Movement ID  0.019702  
street_density_m_per_km2  0.014404  
orig_lon        0.000000  
orig_lat        0.000000  
dtype: float64
```

```
Top features M1 Base (con cotas):  
range_width      0.485908  
dist_km          0.420082  
rel_width        0.051177  
dest_lat         0.027078  
dest_lon         0.011770  
Destination Movement ID  0.003985  
orig_lat        0.000000  
orig_lon        0.000000  
dtype: float64
```

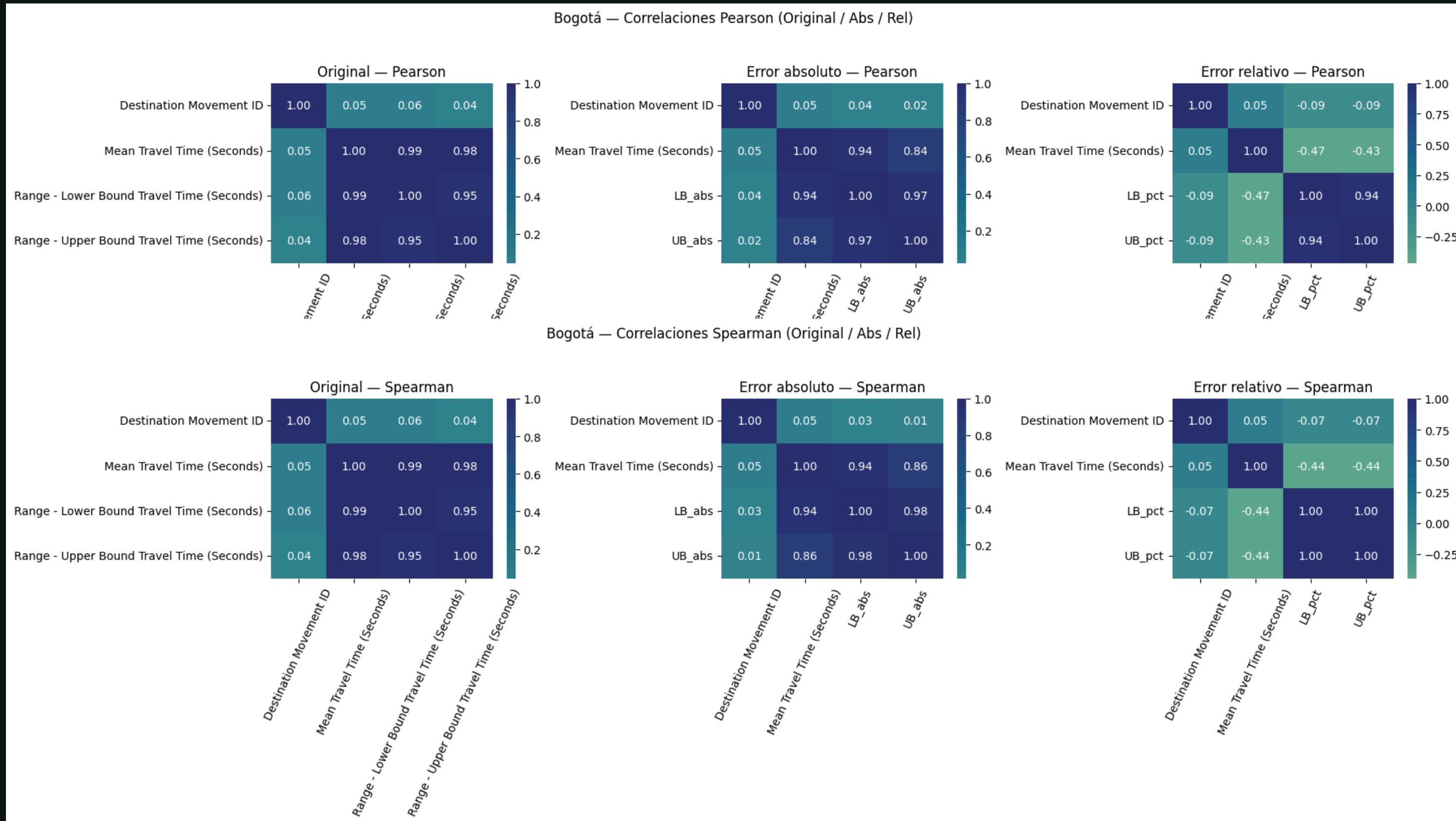
```
Top features M1+ Base + Density (con cotas):  
range_width      0.485188  
dist_km          0.419563  
rel_width        0.050213  
dest_lat         0.026728  
dest_lon         0.011377  
Destination Movement ID  0.003684  
street_density_m_per_km2  0.003248  
orig_lon        0.000000  
dtype: float64
```

Importancia de variables en los modelos

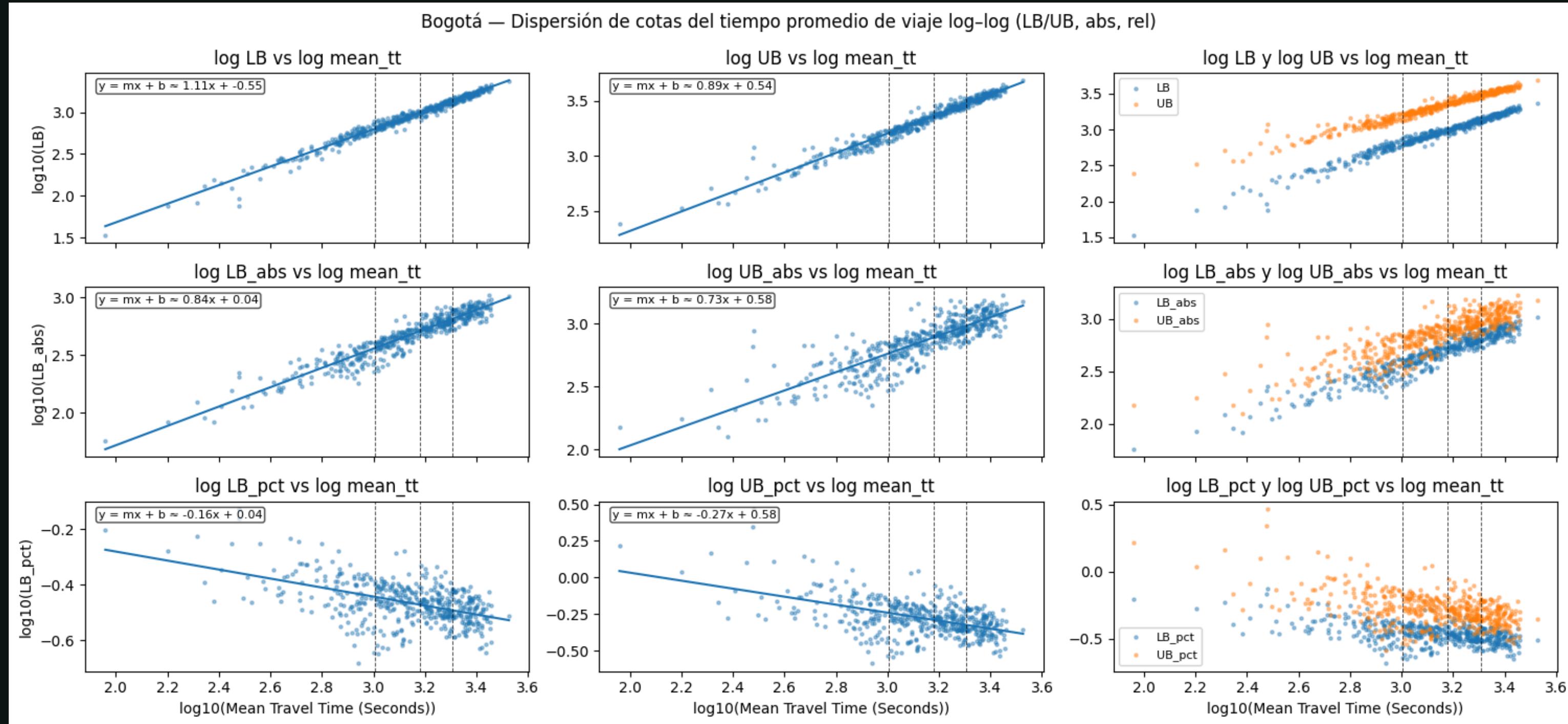
- En modelos netamente exogenos quedamos depende casi exclusivamente de la geometría espacial.
- La incertidumbre temporal contiene información predictiva clave, pues nos mejora mucho el modelo de regresion (al comparar con los meramente exogenos).

¿Por que?

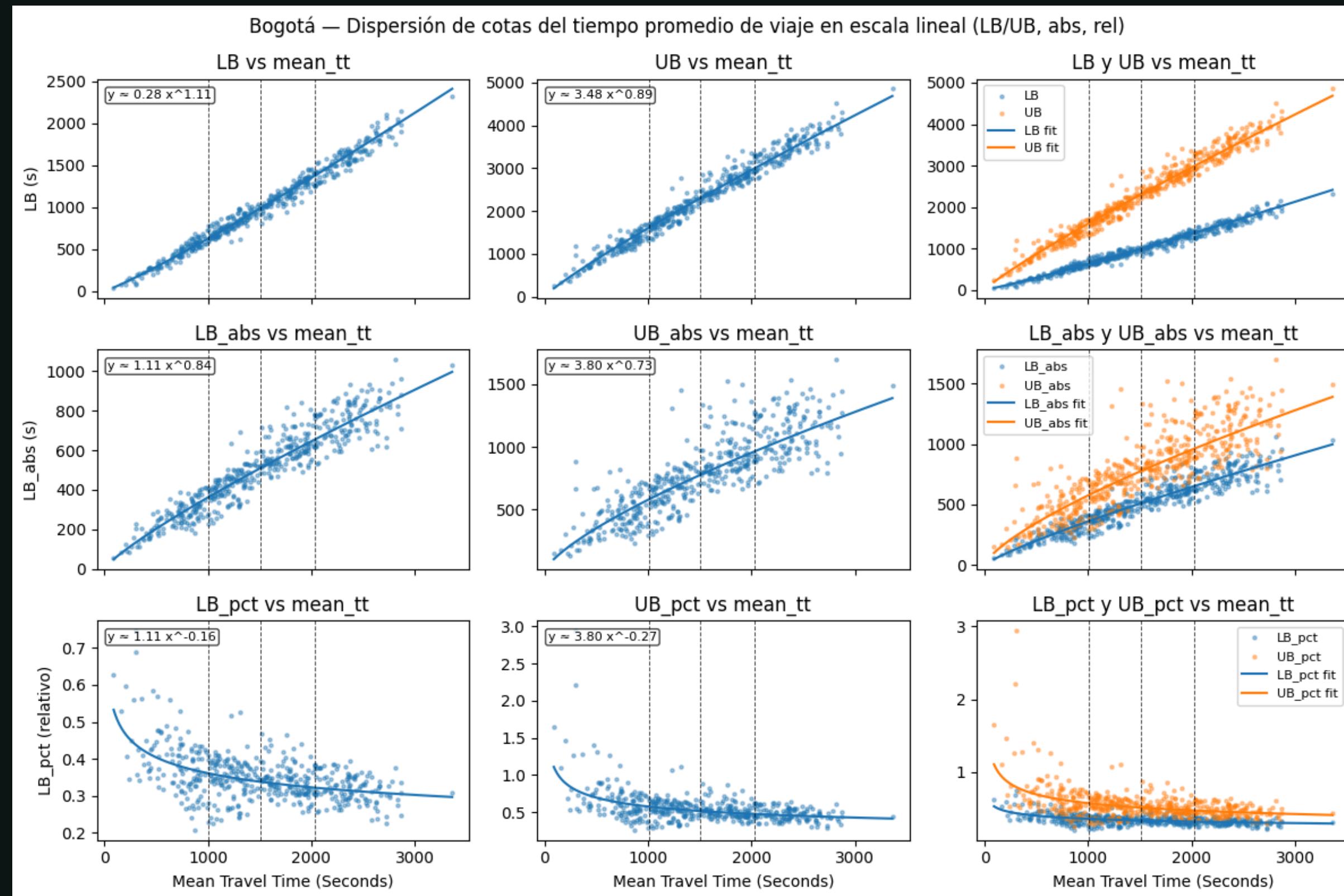
Una posible explicacion-motivo-razon



Una posible explicacion-motivo-razon



Una posible explicacion-motivo-razon



Discusion

- Efecto de las cotas temporales
La inclusión de rangos inferior y superior del tiempo de viaje reduce significativamente el error de predicción.

Estas variables capturan variabilidad operativa y condiciones implícitas del tráfico que no están presentes en la geometría pura. Particularmente, parece que logran capturar efectos multiplicativos que el modelo necesita y que dominan para tiempos de viaje promedio largos-grandes

- Rol de la distancia espacial
La distancia Haversine sigue siendo un predictor clave en todos los modelos, reflejando la estructura radial de la movilidad urbana desde el centro de Bogotá.
- Densidad vial: aporte limitado
La densidad de calles no mejora el desempeño del modelo de forma significativa. Esto puede deberse a que:
 - la métrica es demasiado agregada a nivel de zona,
 - no captura condiciones dinámicas (congestión, semáforos, eventos),
 - la información temporal ya está parcialmente contenida en las cotas.



Limitaciones

- Datos agregados temporalmente
Los tiempos de viaje corresponden a promedios diarios agregados, sin información por hora del día ni día específico de la semana.
- Variables exógenas incompletas
en relación con el punto anterior (días laborales y fines de semana)
- Métricas espaciales simplificadas
La distancia Haversine y la densidad vial son aproximaciones que no capturan completamente la complejidad de la red ni la congestión real.
- Tamaño y estructura del dataset (en relación con los dos primeros puntos)



Conclusiones

- En este caso de estudio, la información temporal agregada resulta más informativa que los indicadores estructurales estáticos de la red vial.
- Para el modelo mejorar posiblemente necesitaría de otra variable (feature) que de razon de fenomenos de componente multiplicativa e incluso de “incidentes” por asi decirlo (para ya no solo mejora del exogeno sino incluso del que incluye info de cotas).
- Modelos simples bien informados pueden ser altamente efectivos: un Random Forest con pocas variables relevantes logra un desempeño elevado sin necesidad de datos altamente complejos.

