Optimizing Cell Tower Placement through Data-Driven Analysis

Alex Wu (tw2883), Guanshi Wang (gw2310),
Yueyan Lu(yl6211), Yuchen Zhao(yz8759)

November 2023

1 Project Objectives

The goal of this project is to employ big data analysis to optimize the placement of cell towers in the US, taking into account known geographic elevation, population density distribution, and building locations and shapes. We aim to predict potential cell tower locations and assess the adequacy of current tower placements. The proposed project will enhance network connectivity and coverage, improving the overall quality of service for mobile users.

2 Data Sources Description

Gridded Population of the World (GPW), v4

- Link: Year 2020_30 Second (approx. 1km)_GeoTiff: https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11/data-download
- Profile and clean by: Guanshi Wang (gw2310)
- Data is 30 arc-second resolution (~1km) in GeoTIFF format. We will use only the data from a single year 2020, and the resulting file size is 352MB. The information we need includes: **Longitude, Latitude, Population-Density**
- GeoTrellis library in Scala will be used to read data and convert data into normalized textfile.

US Building Footprints

- Link: US Building Footprint: https://github.com/microsoft/USBuildingFootprints/tree/master
- Profile and clean by: Yueyan Lu (yl6211)

- DataSet Description: The dataset contains the 129 million computer generated building footprints derived using computer vision algorithms on the satellite imagery. The data is in GeoJSON format and covers the polygon geometries divided by the 50 US states. The total data size is roughly 34 GB.
- Circe library will be used for JSON processing together with GeoTrellis library for geospatial data processing.
- The data schema contains Building **Polygon Coordinates**, and **Centroid, Areas** may require calculation.

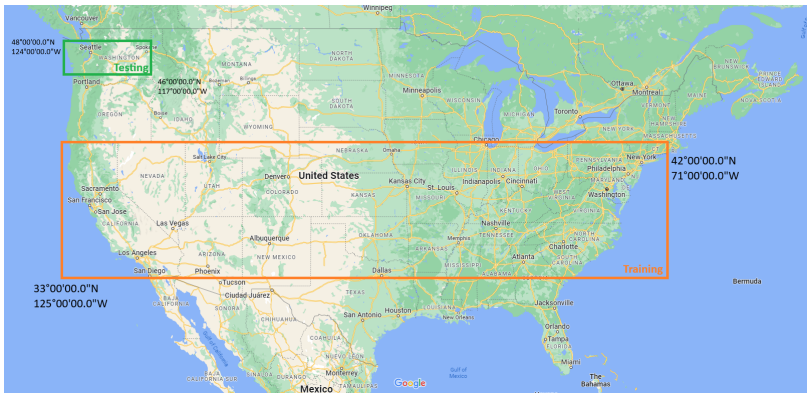
OpenCellID cell tower data

- Link: https://www.opencellid.org
- Profile and clean by: Yuchen Zhao (yz8759)
- DataSet Description: OpenCellID is the world's largest collaborative community project that collects GPS positions of cell towers. The OpenCellID project was primarily created to serve as a data source for GSM localisation. As of October, 2017, the database contained almost 36 million unique GSM Cell IDs. More than 75,000 contributors have already registered with OpenCellID, contributing millions of new measurements every day in average to the OpenCellID database.
- The data schema includes **CID (GSM Cell ID), Latitude, Longitude, Radio Type (like LTE, UMTS)**, etc

USGS National Elevation Dataset

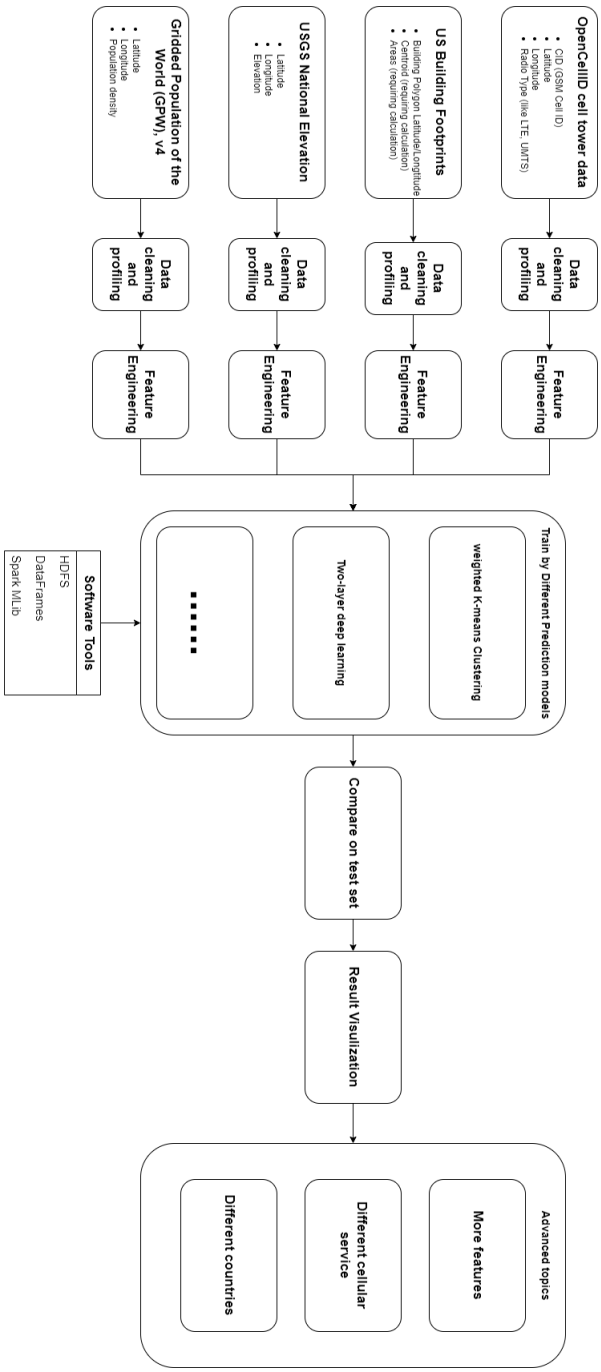
- Link: USGS National Elevation Dataset
- Profile and clean by: Alex Wu (tw2883)
- DataSet Description: Data is 1/3 Arc Second (˜9.7 meter) in GeoTIFF format. We will use only the data from a single year 2023, and the resulting file size is 168.6MB. The information we need includes **Longitude, Latitude, Elevation**.
- GeoTrellis library in Scala will be utilized for geospatial analysis, which performs operations like rasterization, filtering, or any specific geospatial computations. Additionally, Spark SQL will be used to perform exploratory data analysis on the processed geospatial data.

3 Data range



- Training dataset and Testing dataset
  - Training dataset: All data points in the orange box
  - Testing dataset: All data points in the green box
- Why choosing Washington State as the test dataset?
  - It has both a metropolitan area and sparsely populated areas.
  - It has complicated terrains.

4 Initial Design Diagrams



5 Proposed Methodology

5.1 Data cleaning and Profiling

- Transform data formats
- Remove useless columns
- Remove duplicated and useless records

5.2 Feature Engineering

- Interpolation and sampling
- Scaling and normalization
- Generating new features

5.3 Prediction

- Weighted K-means Clustering
- Two-layer deep learning model