Datathon Challenge: Map With AI

For this challenge, you will be solving a few problems related to map data. Knowing where things are is a vital part of Facebook's mission of connecting the world. We will be working with geographic data in Tanzania.

The data you will work with was generated by a variety of machine learning algorithms at Facebook. All of the data come from a region approximately 30 km x 30 km in the vicinity of Dodoma, the capital of Tanzania. (You might be surprised to see that even around the capital, many roads have not been mapped!) They include:

- Population density data – Facebook has generated detailed maps of the population density of the world. The maps were made by using neural networks to find buildings in satellite imagery and then distributing the population of each country from census data among 30m x 30m patches based on the building density. Details of this project and the publicly available data may be found in the links below
    - https://dataforgood.fb.com/tools/population-density-maps
    - https://data.humdata.org/dataset/highresolutionpopulationdensitymaps
- ML generated roads – the MapWithAI initiative at Facebook aims at improving OpenStreetMap (OSM), the so-called "Wikipedia of Maps". Current maps are quite good in the industrialized world, but in the developing world and rural areas maps remain incomplete, and are constantly changing. MapWithAI aims to help map the world by using neural networks to find roads that can be added to OpenStreetMap.
    - ML predictions as raster images. There are 64 images with 8192x8192 resolution.
    - The individual files are labeled by Bing quadkeys. Quadkeys are used to tile the world represented in the Mercator projection.
    - We have converted the raster images to csv files with rows labeled by the corresponding pixels in the image. The pixel intensity ("val" in the csv files) ranges from 0 to 255 and is proportional to the estimated probability of the indicated pixel being a road. For your convenience, we have also converted the pixel positions to latitude and longitude at the northwest corner of the pixel. To reduce the size of the csv files we have truncated the data to val >= 5, removing the lowest probability pixels.
    - In the csv files, the pixel positions are indicated with the columns 'pixel_i' and 'pixel_j' which correspond to representing the image as an array image_array[i, j] where (i, j) label the rows and columns of the array in the standard way. So i increases going from North to South and j increases going from West to East.
    - Each pixel approximately corresponds to an area of 0.5m x 0.5m. We'll assume that horizontal and vertical distances in lat/long units are the same in physical geometry (and very close to the same in pixel units.)
    - There are various standard packages that might (or might not) help you handle this data. An example for Python is mercantile.
- A third dataset to investigate if you have time is the existing set of OpenStreetMap roads for the same area. This is given to you in the form of an XML document. You will need to parse this XML to extract the roads.
- For the purposes of this challenge, please work in units of latitude-longitude. They don't correspond precisely to 2-d geometric distance, but because we are working in a small area near the Equator, it is close enough. For reference, 1e-5 in lat-long units is approximately 1 meter.

The challenge has two parts.

1. By combining the ML-generated roads and the population density map, estimate which people and how many people are reachable by the ML roads. Precisely, you should compute:
    a. For every point in the population density dataset tz_popdens_sample.csv, find the distance to the nearest road pixel **with val > 75** (the limit is imposed to be reasonably confident in the prediction) in latitude-longitude units.

b. Please report your answer as a CSV file with three columns: ‘latitude’, ‘longitude’, and ‘distance’.

c. How many people in this area live within 0.001 lat/lon (about 100m) of an ML road?

d. Extra challenge (if you have time): How does the population reach of the ML road network compare with the OSM roads?

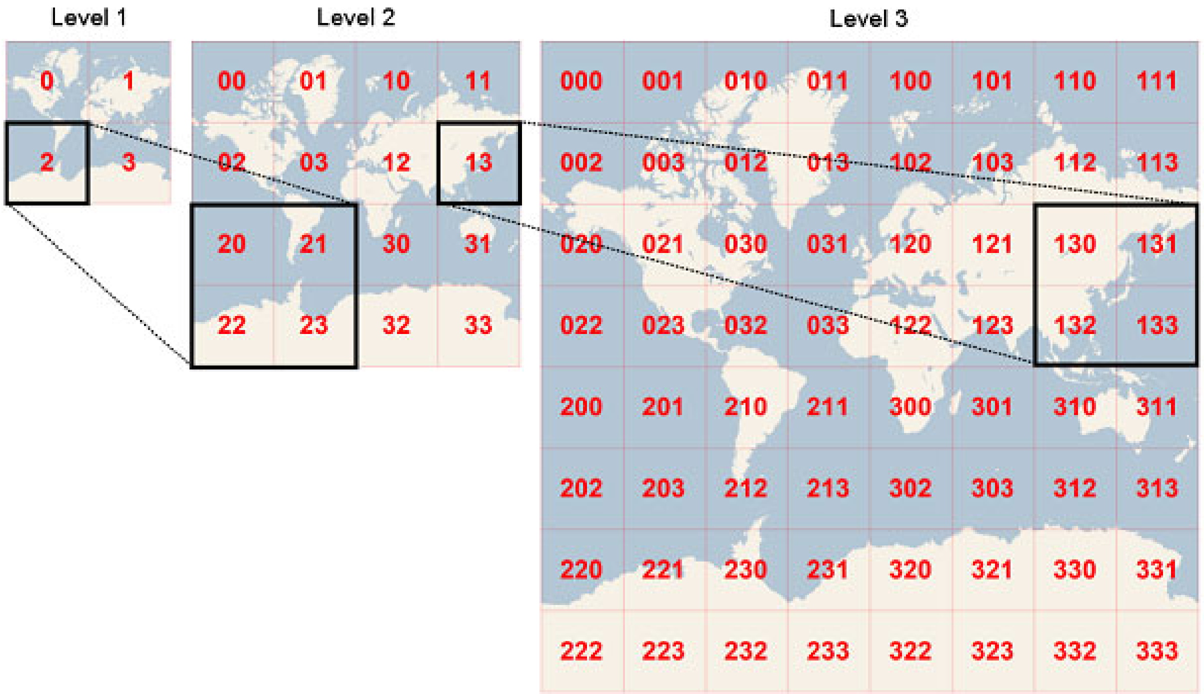
2. Suppose that we need to use the ML roads to find paths between pairs of points.

    a. Given a set of pairs of points in the file routing_challenge_pairs.csv, estimate the distance (again, in lat/long units) along the estimated road network (that is, tracing along pixels of the ML predictions with positive road probability.)

    b. Please report your answer as a CSV file with five columns: ‘latitude_src’, ‘longitude_src’, ‘latitude_dst’, ‘longitude’, and ‘distance’.

    c. We will score your answers as fully correct if they are within 5% of our reference solution, so intelligent approximations are strongly encouraged.

Bing quadkeys:



Sample of ML raster predictions: