Project Proposal

October 26, 2017

Yassine Kadiri, Zsolt Pajor-Gyulai, Santiago Novoa, Manuel Serrano Rebuelta

Abstract

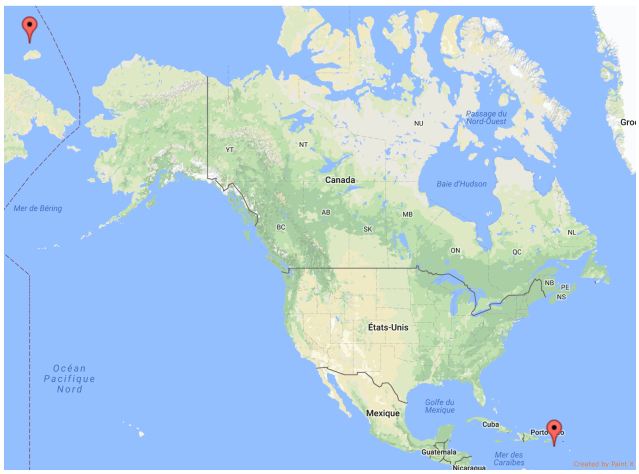
In this project, we attempt to estimate the population of a certain geographic area based on satellite imagery. We interpret this task as a classification problem where the classes are given by appropriately chosen population intervals. The idea comes from a paper[a] where the authors develop a similar model.

[a] Caleb Robinson, Fred Hohman, Bistra Dilkina, A Deep Learning Approach for Population Estimation from Satellite Imagery, 2017

The Data

For this project we will use two main data sources:

(1) We obtain labels on the target variable (population) from the US census data. This data is stored and downloaded as an ASCII file with some features and a two dimensional array representing the continental United States. Each entry in that array stands for the population density of the corresponding region. We have been able to read this data and convert it into a convenient pandas DataFrame. We also relabeled the columns and rows so that each column represents the latitude of the western sides and the rows represent the longitude of the northern sides of the squares in the grid. The grid covers the rectangle defined by the following two locations on the map:



As we are interested in estimating the population in the United States, we will need to perform some preprocessing to remove areas corresponding to Canada, Mexico and other countries involved.

(2) The second data source is satellite imagery captured by Landsat 7, which serves as the raw input for our neural network. This data is available on several websites but we will extract it through the Google Earth Engine API. The images will be chosen to align perfectly with (a perhaps coarsened) grid defined by the census data. We will randomly sort these images into training and validation sets.

The images we consider are approximately $5\text{km} \times 5\text{km}$. To clarify what this means, suppose that 1° in latitude and 2° in longitude are respectively the best approximations to 5km in real distance. Then we will choose data from the population grid and satellite images so that they cover exactly 1° in latitude and 2° in longitude.

The Problem

As mentioned above, we will approach this task as a classification problem. Based on the area captured on a single image, we will determine the optimal number of classes and their precise definition.

Using the census data as labels, we will then use a pre-trained convolutional neural network (Vgg16) to obtain a classifier. Once the CNN is trained, we will evaluate it on our holdout set of satellite images. The precise metric to measure the performance is yet to be determined but we will use tools studied in this course. In this phase, we will determine the optimal number of epochs and the value of other relevant hyperparameters. We expect this to be a difficult step that will require a considerable amount of time and effort.

Scope of the project

Initially, the scope of the project is restricted to the continental United States. However, if the continental US turns out to be a rather ambitious task, we would consider limiting our scope to neighboring states such as Colorado and Ohio which happen to be almost perfectly rectangular (In fact, Colorado is exactly rectangular). We expect this to be a simplified setting as in this case we would study states that have the same architectural and urban patterns and therefore train the CNN quite efficiently.

Application of the project

It is crucial problem for every government to accurately estimate the geographical distribution of their countries population. For example, local governments and cities often receive funding based on the population of their jurisdiction. In the absence of further data, the government has limited access to this information in between two census years. Intricate mathematical models are frequently used to predict the evolution of the population, however, there is often little opportunity to validate these predictions. Our project might give an answer to this problem (if it performs well) by allowing us to predict the population of a given geographical area in an inexpensive way.

Another use of this project would be related to intelligence agencies allowing them to assess a countrys population even though, as in the case of North Korea, some countries keep this information secret or often falsify it.