Project Elective

Map Analytics

Nilay Kamat, Vineet Priyedarshi & Prince Butani

January - May 2024

Contents

Chapter 1

Curating A New Dataset

Objective

The primary goal of this week's work was to enhance the diversity of maps used for training and testing our machine learning (ML) model, specifically focusing on choropleth maps with discrete legends. The existing dataset primarily consisted of images from the MapQA dataset, and the need arose to validate the model's performance on a broader range of map types.

Activities Completed

1. **Map Selection:**
   - Went through different articles and research papers from credible sites to select choropleth maps which go along with our work and at the same time has some differences.
   - Conducted an extensive search on platforms such as ResearchGate and the official U.S. government website to ensure credibility.
2. **Dataset Curation:**
   - Curated a dataset of 20+ choropleth maps from diverse sources to evaluate the model's adaptability to different map types.
   - Some maps had differences in design, some had lower resolution, and some were not coloured not on the basis of states ,though it was inconsequential for us.
3. **Trustworthy Sources:**

   - *ResearchGate:* A reputable academic platform known for hosting peer-reviewed research. Maps from this source were considered trustworthy due to the rigorous vetting process of academic publications.

   - *U.S. Official Website:* Maps sourced from the official U.S. government website were deemed reliable, given the authority and accuracy associated with governmental publications.

4. **Challenges Faced:**

   - Limited Availability: Difficulty in finding a large number of relevant maps due to the specificity of the criteria (choropleth with discrete legends).

   - Skewed Distribution: Several potential sources were discarded due to a skewed distribution of legends, hindering the diversity required for robust model training.

Images

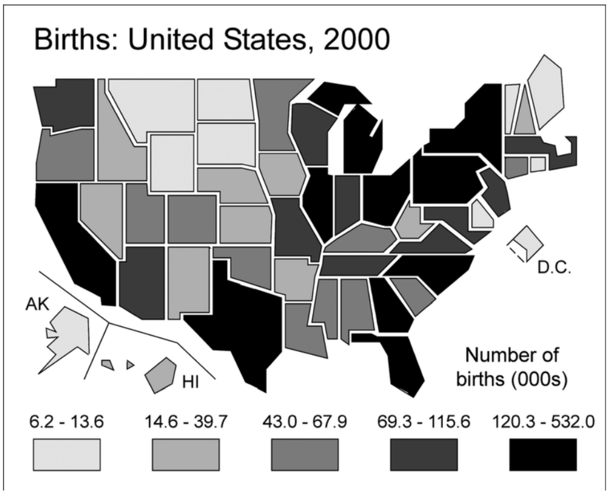


Figure 1.1: More geometric and discrete than mapQA maps

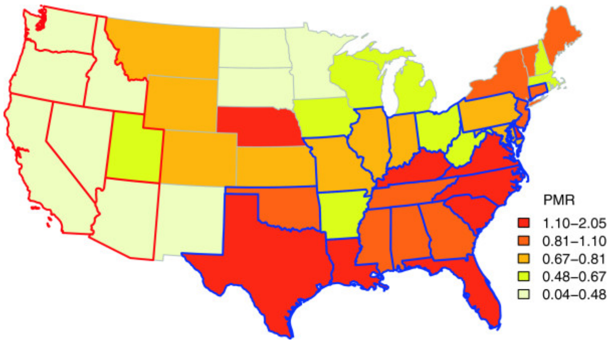


Figure 1.2: Distinct outlines around the states

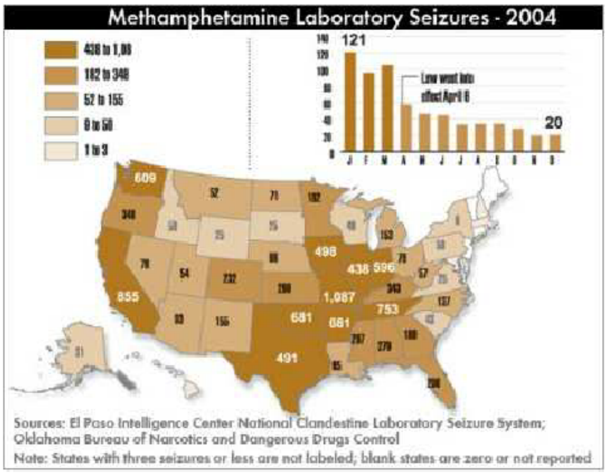

Methamphetamine Laboratory Seizures - 2004
438 to 1,087
162 to 348
52 to 155
8 to 50
1 to 3
121
20
Law went into effect April 6
609
348
855
491
498
438
596
1,087
681
681
753
Sources: El Paso Intelligence Center National Clandestine Laboratory Seizure System; Oklahoma Bureau of Narcotics and Dangerous Drugs Control
Note: States with three seizures or less are not labeled; blank states are zero or not reported

Figure 1.3: Very low resolution of the map and the legends

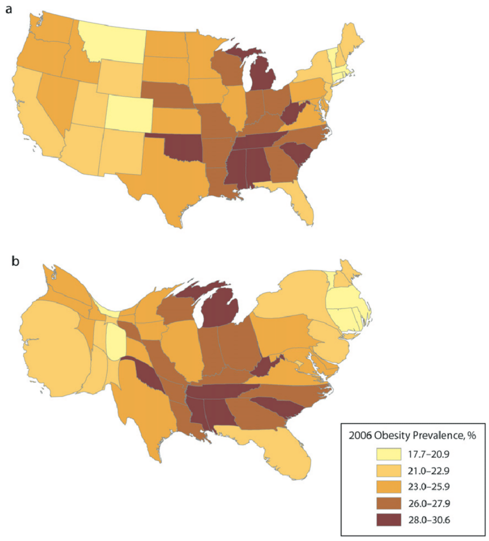
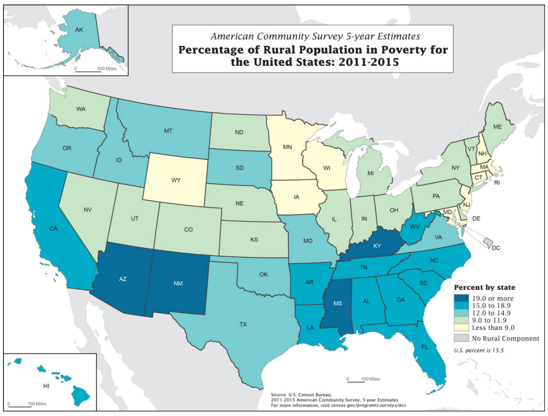


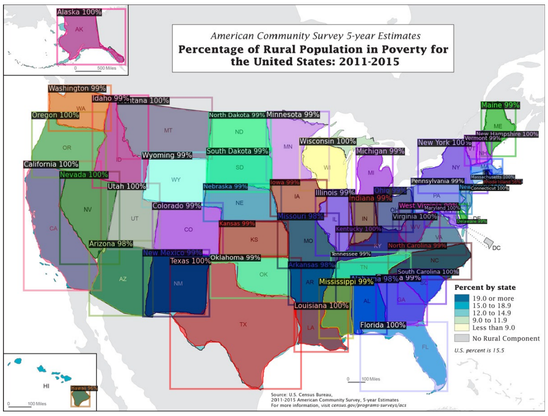
Figure 1.4

1.1 Outcome

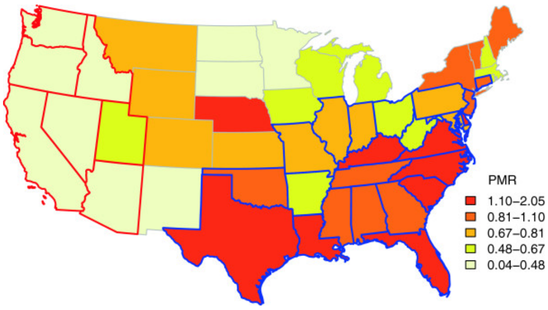
On running them through segmentation model, the results were good but not great. While some maps did show accurate output, the more complicated and different ones showed wrong labels. Below are some maps and their corresponding output on running through the segmentation model.



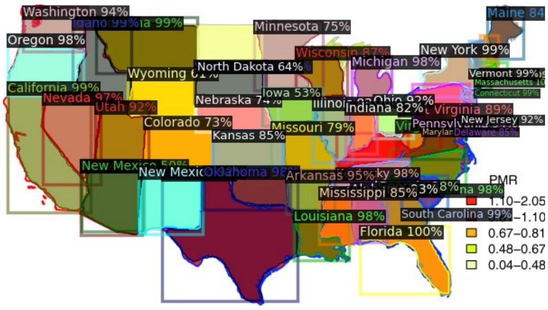
(a) Before segmentation



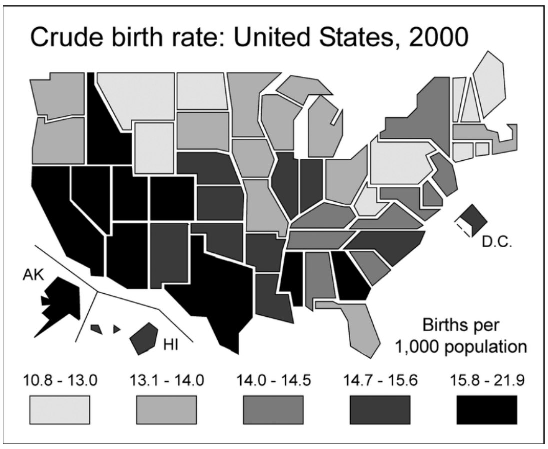
(b) After segmentation



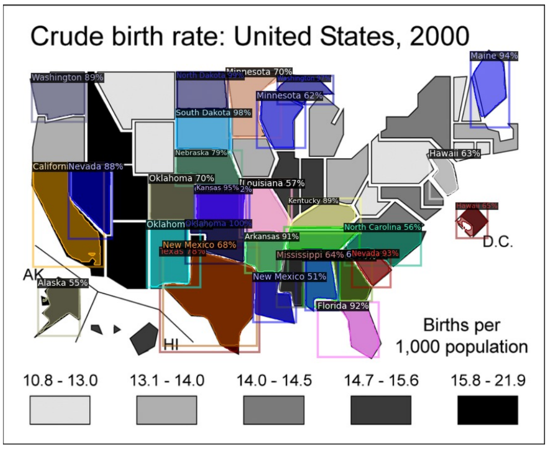
(c) Before segmentation



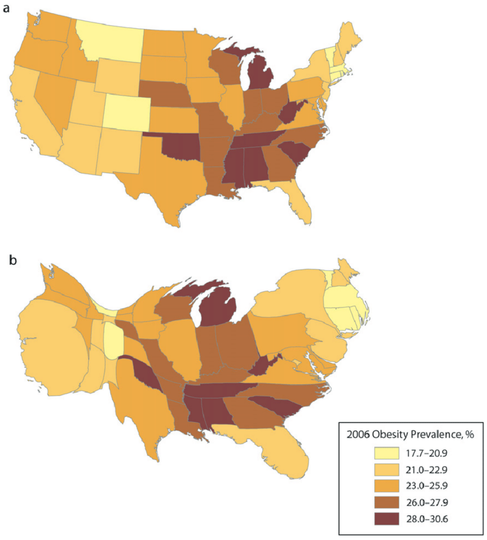
(d) After segmentation



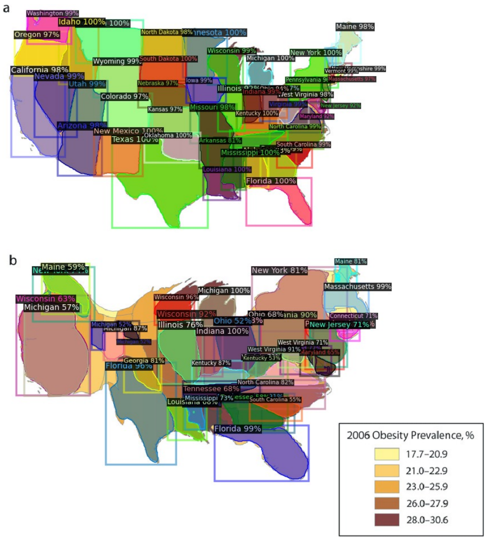
(e) Before segmentation



(f) After segmentation



(g) Before segmentation



(h) After segmentation

Chapter 2

OCR for numerical data

2.1 Objective

The aim of the project is to retrieve the numerical data present in each map image and evaluate it for the countries in the USA. The extraction of the numerical data is being done with the help of OCR techniques.

2.2 KerasOCR and current issue

The initial attempt at obtaining the numerical data values from the map was made with the KerasOCR model. We coded a script to recognise text in the map image and form bounding boxes around them. This script can be found here : KerasOCR. The numerical data detected in the Discrete Legends was found to be inaccurate. The KerasOCR model failed to accurately retrieve the numerical data and would often get confused in 0s and o's, ”-” symbols and so on.

2.3 Plan for this semester

We plan to try out another OCR model, ”Tesseract OCR”, for the numerical data extraction. We also have to obtain the distance in number of pixels that we need to move to the left hand side in order to gain information on the colour corresponding to that numerical text.

Working with Discrete Legends:

1. The numerical data must be retrieved more accurately. Currently with KerasOCR model,

the model isn't able to extract the data accurately.

2. There are several other ML models pertaining to OCR that can be tried on the map images to see which works best - Tesseract OCR, Google Cloud Vision, etc.

3. Once the data is extracted accurately, the colour associated with the discrete values can be obtained via observing the pixels located to the left of the data. The code for this will also need to be implemented using the bounding boxes of the recognised numerical data.

Working with Colour Bar:

1. In colour bar, the value is based on the position of the colour in the colour bar. To evaluate the value, the methodology would be to obtain the colour of the country from the map and locate it in the colour bar.

2. We can obtain the range of the values that can be achieved by using OCR and then find the length of the colour bar.

3. After the colour is located, we can then use the length of the colour bar and the position of the colour we have obtained to exact information on the value.

4. How can the colour be located? This has to be thought about, maybe a better understanding of choropleth map image generation and colour theory will help out.

Integration with Segmentation

1. Upon extracting the numerical data from the map image in images containing the colour bar or the discrete legends, the colours corresponding to the data need to be matched with the colours present in the map, i.e. the colours pertaining to each country.

2. The colour of each country can be stored as a dictionary with ”State : Colour” as the elements.

3. The task is then to match the colours with that of the numerical data. This will be a simple task in the case of images with Discrete Legends since the data value and the colour is already ready.

4. In the case of images with colour bar, there will be a need to identify the colour's location in the colour bar. Again, this needs to be pondered upon and a better understanding of choropleth map image generation and colour theory might help out.

Chapter 3

Segmentation and Visualization of Map Images

3.1 Segmentation

We will focus on segmentation of the map images into 50 states. Using detectron2 model, we identified the states and segmented them. **By mid semester**, we will perform segmentation of the images, associate the extracted data, be it discrete or continuous value, to the respective state it belongs.

3.2 Visualization

By end semester, we will create end to end pipeline and demonstrate the use different kinds of bars, plots and graphs to generate visualizations with the data extracted from map images, for whole country or for few states. We will also try to generate an automated data story from the visualizations created and get insights as to what the visualizations mean.