# CME291 Final Report

Yuan Gao, Ruohan Zhan

May 2019

**Abstract**

Geographic modeling on healthcare outcomes can facilitate policy making, which can contribute to more systematic and targeted care to be provided in developing country. However, high-resolution mapping of healthcare outcomes are not standard and currently available data are expensive and have limited resolutions in time and space. Recent research (Blumenstock et al. 2015, Jean et al. 2016) has shown the potential of machine learning as an alternate mapping approach which makes use of passively-collected high-resolution data such as satellite and cellphone call data records to infer a spatial distribution of socioeconomic variables. This project will extend previous work to obtain healthcare map in Nigeria, which has the potential to identify gaps in care and thus enable improved access to care and quality of the care available.

## 1 Introduction

As a developing country, Nigeria faces many socioeconomic issues. One significant problem is that the quality of Nigerian healthcare service is rather poor. The lack of healthcare resources affects the overall health condition of people in the country.

One important aspect of the healthcare system is maternal and perinatal care. Nigeria has a high maternal mortality rate at 560 per 100,000 live births by World Health Organization (WHO [2014]). As a comparison, the maternal mortality rate of the United States is only 28 per 100,000 birth (WHO [2014]). The main reasons behind the high mortality rate are insufficient prenatal care services and non-medical assistance during delivery. According to Bankole et al. [2009], in Nigeria, more that 40% of women do not receive any prenatal care from professional healthcare facilities before delivery.

Previous study by Jean et al. [2016] has successfully created maps to display the predicted economical conditions in Nigeria. In the scope of our project, we adapt the model used in the previous research combining health survey data, and extrapolate the model to map out the maternal and perinatal care status and vaccination coverage of entire Nigeria. For both aspects in healthcare system, there are different indicators that play important roles in monitoring the quality of healthcare system. Generating a map of each indicator can help us to find correlations between the indicator and geographic traits. Such maps can potentially be used to assist policymakers to improve their efficiency in allocating resources, further improving the overall healthcare quality of in the region.

## 2 Methodology

In this project, we utilize the social-economic information embedded in satellite data to predict healthcare related indices in Nigeria. Specifically, we adapt and extend the model developed in Jean's research Jean et al. [2016] which is mainly composed by the following three steps:

**Low-level feature extraction** which takes a pre-trained convolutional neural network model on ImageNet to obtain low-level image features that common to vision tasks, such as edges and corners;

**High-level feature learning** which trains a model to predict nightlight intensities correspond to daytime imagery and extract image features from daytime satellite images;

**Supervised learning** which is conducted using the extracted features from the second to last layer of the model and combine with survey data that contain important indicators for classification. In our case, we explore the topic of vaccination coverage and maternal/neonatal child health respectively.

## 3 Computation Resources Setup

In the first stage of the project, we focus on setting up the technologies required to run the model. The pre-trained CNN model was developed under Caffe deep learning framework and shall be run in the GPU mode. Thus, we asked for permission to use Sherlock, a High-Performance Computing cluster by the Stanford Research Computing Center. In the environment of Sherlock, the Caffe module requires to be run on a GPU cluster. Note that the waiting time for entering the GPU cluster varies depending on the resources and accounts (from minutes up to hours).

However, Sherlock only has Caffe2 module instead of Caffe module that the original model was developed in. We transferred the initial Caffe model into a Caffe2 model format to allow it to be run on Sherlock. In this step, we first needed to convert the layer definition of the original model to a more current version of Caffe with the embedded tool `upgrade_net_proto_text` to upgrade the proto; then, we used the `caffe_translator` in Caffe2 to upgrade the Caffe model.

The bulk of computation is managed by Slurm job scheduling system. After the data being prepared, we submit a Slurm job script to the scheduler to run our model.

## 4 Data

As we are familiar ourselves with Sherlock platform and set up the original model to work with our technologies, we also collected the data we need. There are two major parts of the data sources: daytime satellite images and healthcare survey. The daytime satellite images are collected from Google's Static Map API, while the healthcare survey data are obtained from the Demographic and Health Surveys (DHS) Program database.

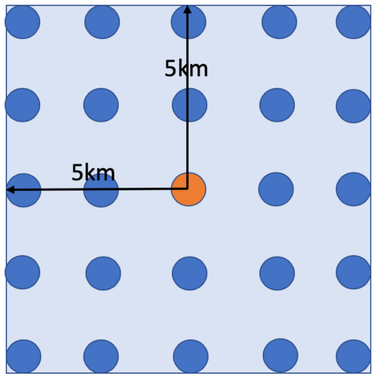
The health survey data we used is the Standard DHS data (in phase VI) of Nigeria collected in 2013, which is the latest Standard DHS in Nigeria we can request from the current database. The dataset contains multiple recode files, such as birth recode, children recode, individual recode etc. After examining the files, we find the following indicators most related to the goal of our project, and thus select them to conduct further analysis:

- Maternal care related data: months pregnant at first antenatal visit, number of antenatal visits during pregnancy, place of delivery, reason didn't deliver at health facility: too far/no transport.
- Vaccination related data: polio0, polio1, polio2, polio3, DPT1, DPT2, DPT3, BCG, measles, ever vaccination, health card.

Some data processing is also done here. For place of delivery, the original data file record this indicator in words, such as "home", "government hospital", "private hospital" etc. For the purpose of analysis, all places as health facilities are grouped together and assigned value 1, while home and other places are assigned value 0. Similarly, for reason didn't deliver at health facility, if the facility is indeed too far, we assign value to be 1, otherwise 0. The manual of DHS indicates a blank field meaning this question is not applicable to the interviewer, and a field with 99 or 999 meaning this is a missing data. In both cases, the data is filtered and not included in the experiment.

Besides, we collect two sets of satellite images, for training and testing purpose respectively:

- The first set is collected based on the geographic information in the DHS data file, which is used to fit the model. There are 904 clusters in the 2013 survey. Each cluster corresponds to one GPS location. The satellite images are extracted based on those GPS locations. Note that to protect the respondents' privacy, the GPS positions are randomly displaced up to 5 kilometers. To include the geographical attributes of the possible locations of the given cluster, we sample a 10 km $\times$ 10 km square area with the center to be the GPS location of each cluster. To reduce the computational cost, for each 100 square kilometers, we sampled 25 1 km $\times$ 1 km area, shown as Figure 1a. We download a 400 pixels $\times$ 400 pixels daytime satellite image that covers this 1 km $\times$ 1 km area through Google Static Map API for each of the 25 images in one cluster. Figure 1b is an example of a 400 $\times$ 400 pixels daytime image.




(a) How we sampled 25 images from each cluster.


(b) An example of the daytime satellite image collected.

- The second set of satellite images cover all regions in Nigeria, which will be used to draw a map of our prediction. Started from the northwest location of Nigeria, we collect one satellite image in each 20 by 20 km area. This image will then be fit into the model and predict health care index for that region.

# 5 Model pipeline

For a specific healthcare index, we use the following pipeline to fit the data and make prediction:

- We first download the pretained Caffe model of the second step from Jean's GitHub, which takes a satellite image as an input and output extracted features of dimension 4096;
- On the supervised learning stage, since our data is highly nonlinear and high dimensional, we use tree-based methods which are non-parametric and easy to deal with many features. Cross-validation is used to select the hyperparameters with R2 score as the evaluation metric.
- Once we have fitted our model, we can predict the value of health care indicator for a specific region, based on which we draw our health care map.

# 6 Experimental results

## 6.1 Fitted models

We randomly select 80% data as training data, and the rest 20% as testing data. Cross validation is used to choose hyper parameters. To evaluate our models, we use the following three metrics:

**R2** R2 score, which shows how our fitted values explain the variance of original true values.

**MSE** mean squared error of fitted values and true values.

**MAE** mean absolute error of fitted values and true values.

| | Training | | | Testing | | | Model |
|---|---|---|---|---|---|---|---|
| | R2 | MSE | MAE | R2 | MSE | MAE | |
| polio0 | 0.843 | 0.100 | 0.015 | 0.602 | 0.154 | 0.039 | XGBoost |
| polio1 | 0.569 | 0.115 | 0.021 | 0.207 | 0.145 | 0.039 | RandomForest |
| polio2 | 0.553 | 0.123 | 0.023 | 0.190 | 0.160 | 0.041 | RandomForest |
| polio3 | 0.443 | 0.134 | 0.027 | 0.017 | 0.170 | 0.044 | RandomForest |
| DPT1 | 0.889 | 0.089 | 0.012 | 0.591 | 0.165 | 0.046 | XGBoost |
| DPT2 | 0.982 | 0.034 | 0.002 | 0.591 | 0.167 | 0.045 | XGBoost |
| DPT3 | 0.771 | 0.121 | 0.022 | 0.556 | 0.173 | 0.044 | XGBoost |
| BCG | 0.837 | 0.108 | 0.019 | 0.635 | 0.153 | 0.043 | XGBoost |
| measles | 0.717 | 0.115 | 0.021 | 0.475 | 0.157 | 0.038 | XGBoost |
| Vaccination | 0.556 | 0.129 | 025 | 0.126 | 0.163 | 0.048 | RandomForest |
| health card | 0.916 | 0.078 | 0.010 | 0.588 | 0.167 | 0.049 | XGBoost |
| birth place | 0.883 | 0.097 | 0.015 | 0.613 | 0.182 | 0.050 | XGBoost |
| far | 0.544 | 0.111 | 0.021 | 0.052 | 0.158 | 0.050 | XGBoost |
| first visit | 0.503 | 0.431 | 0.295 | 0.090 | 0.561 | 0.556 | XGBoost |
| num visit | 0.886 | 1.305 | 2.659 | 0.660 | 2.351 | 8.567 | XGBoost |

Table 1: Tree-based methods fitted results

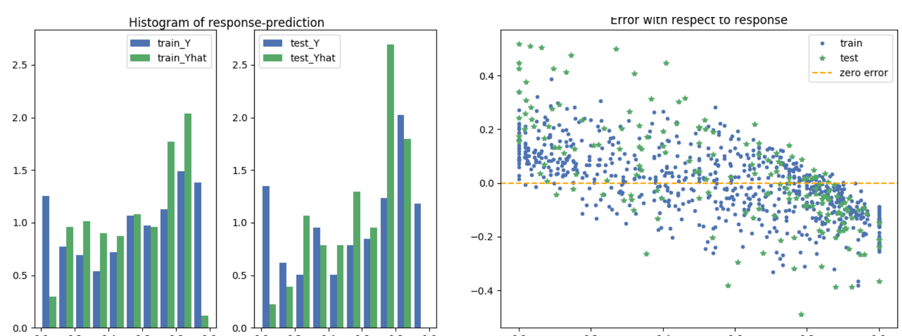



Figure 2: Polio0

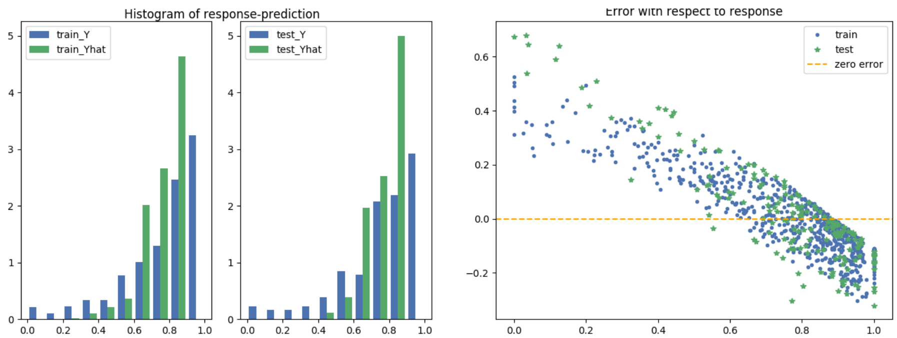



Figure 3: Polio1

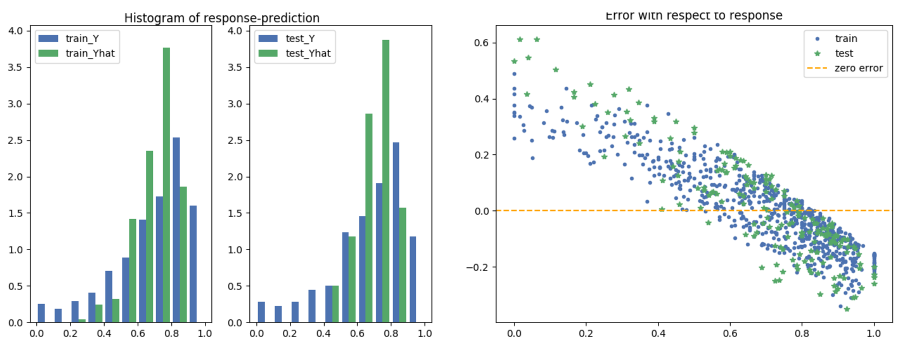



Figure 4: Polio2

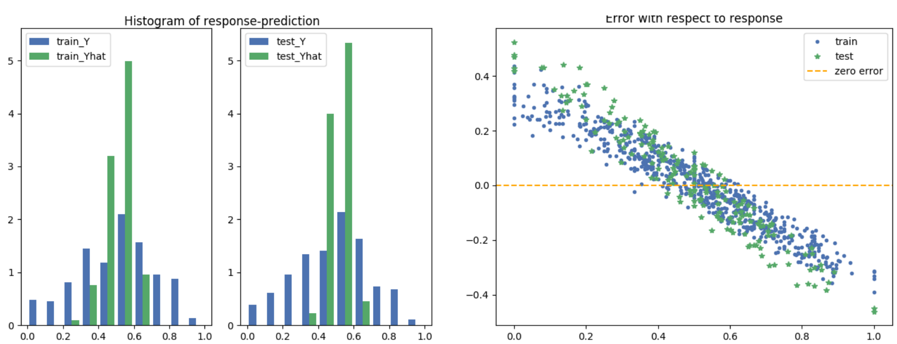



Figure 5: Polio3

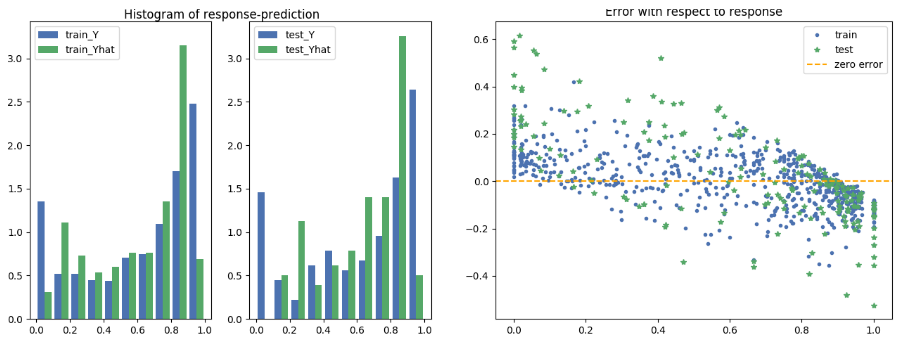



Figure 6: DPT1

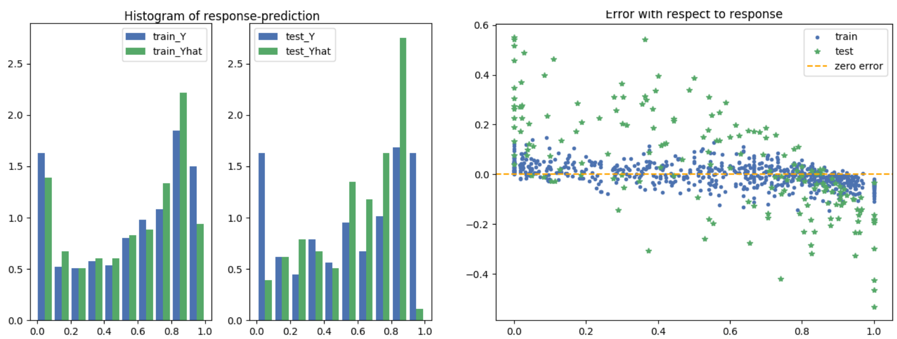



Figure 7: DPT2

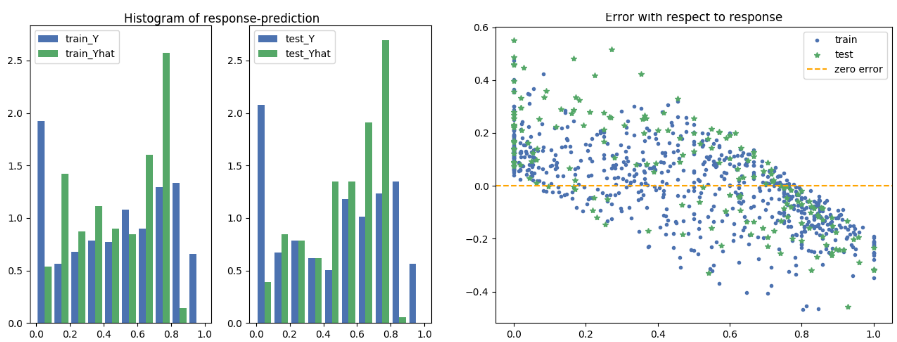



Figure 8: DPT3

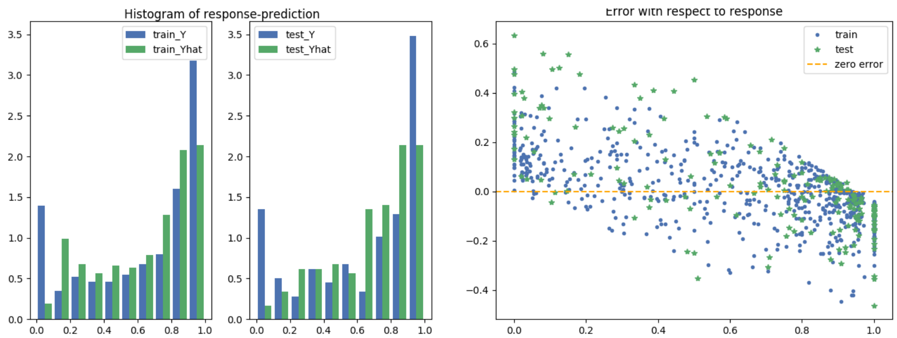



Figure 9: BCG

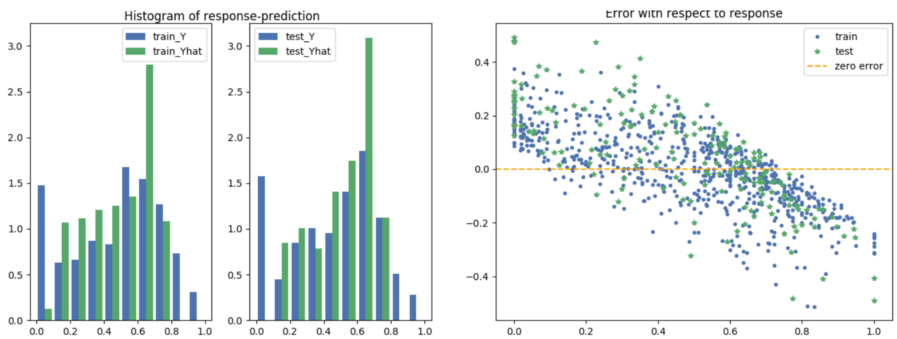



Figure 10: Measles

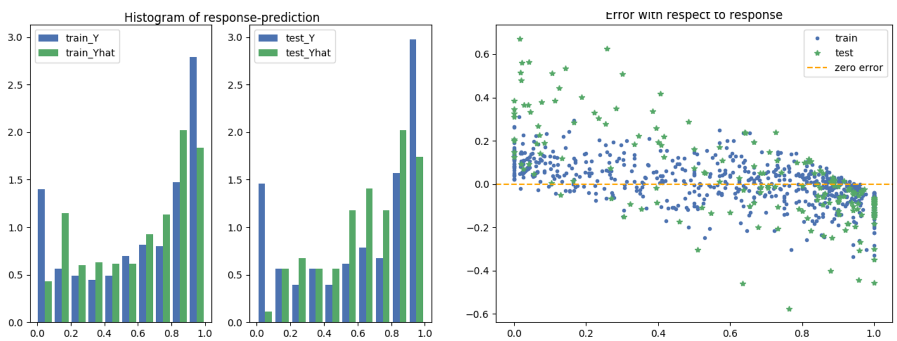



Figure 11: Health card

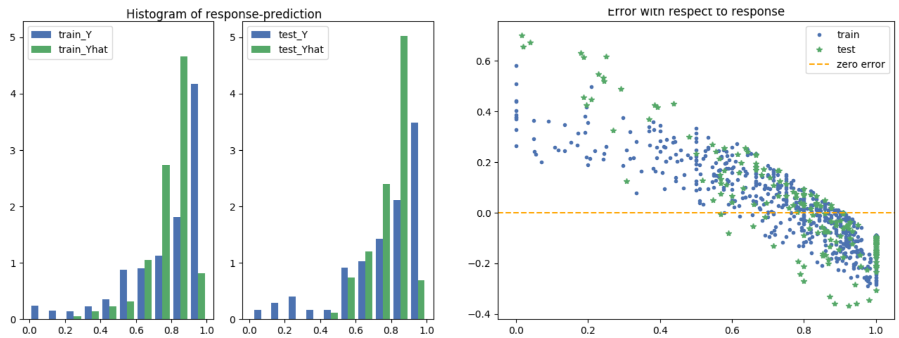



Figure 12: Vaccination

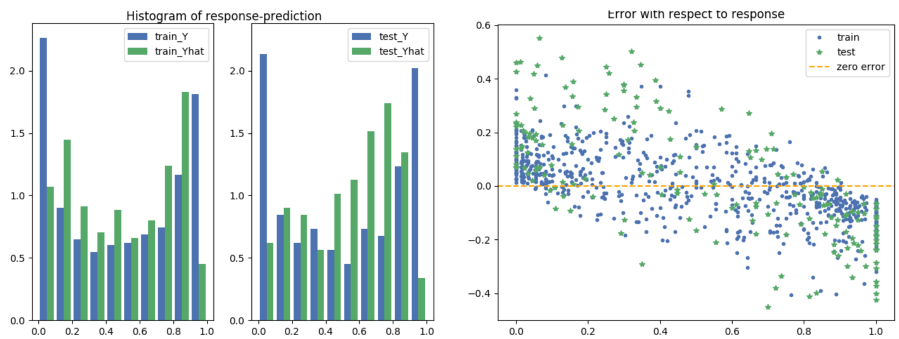



Figure 13: Birth Place

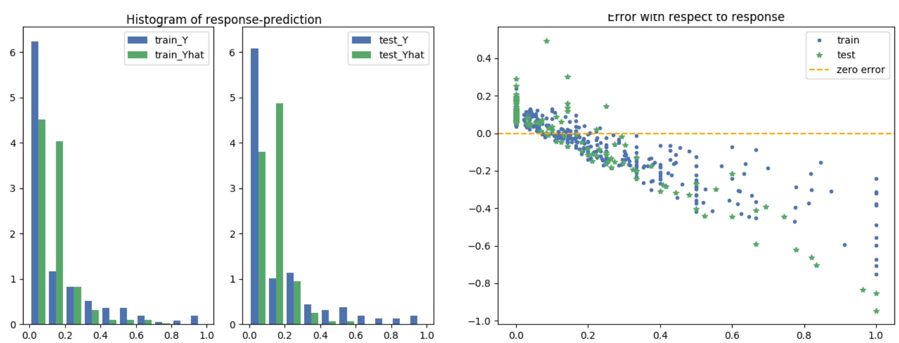



Figure 14: far

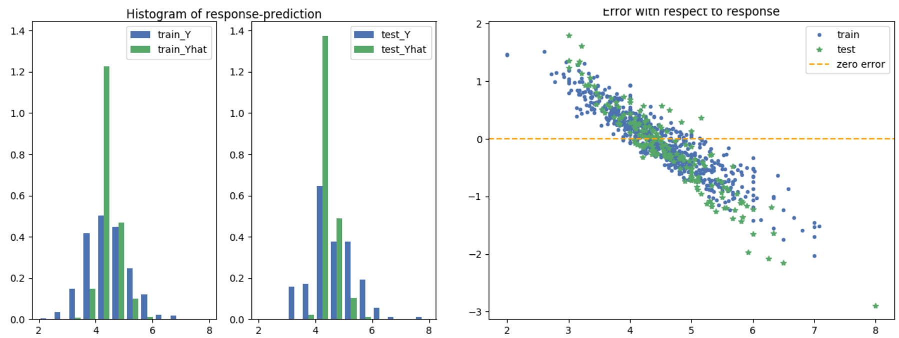



Figure 15: first visit

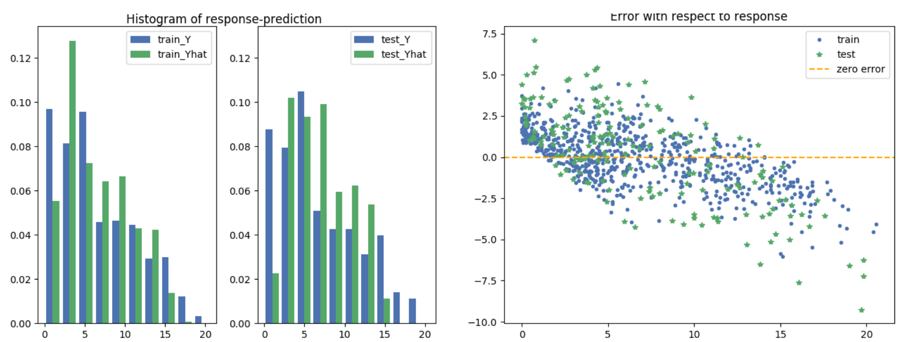



Figure 16: visit number

## 6.2 Healthcare maps

In this step, based on the models fitted from the supervised learning, we draw a health care map by predicting the index at particular geographic locations. Specifically, we make grids of Nigeria and choose one point per $20 \times 20km^2$. After we predict the health care index for each point, we use the average indices for each province and make the maps as follows:

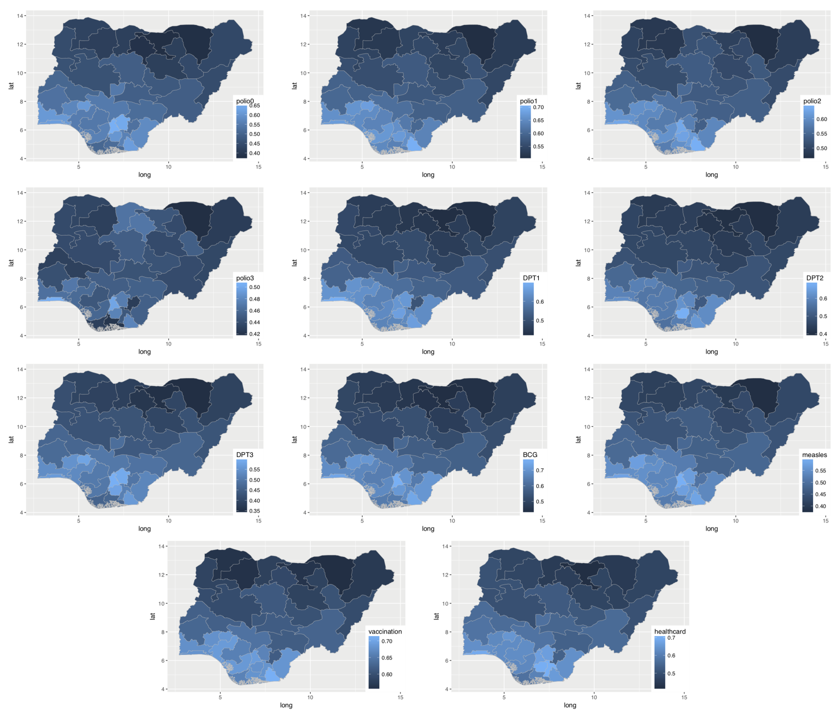
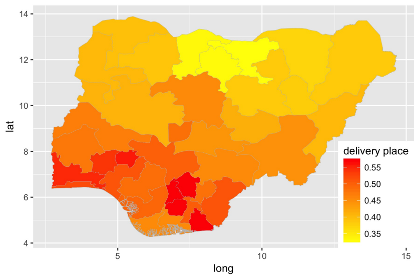
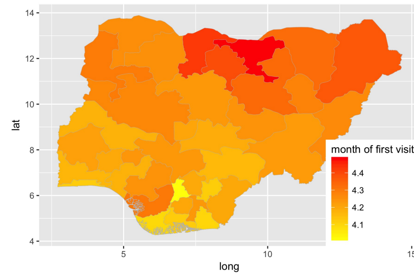



Figure 17: Health care maps of vaccination coverage indicators.

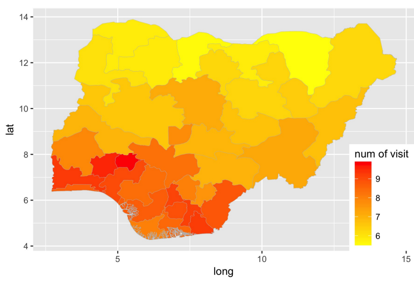
It can be seen from the map that the vaccination has more coverage in the southern Nigeria compared to the northern Nigeria. Furthermore, we compare our maps with IHME DPT1 and DPT3 vaccination coverage in 2016 in Nigeria below. We can see that despite the year difference, our predicted maps have very similar patterns with the IHME maps.



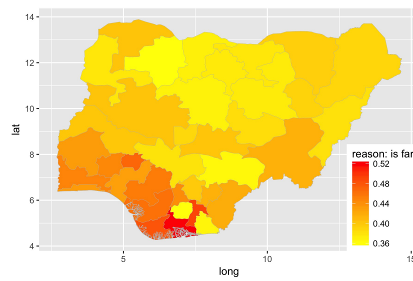

(a) Place of delivery




(b) Months pregnant at first antenatal visit

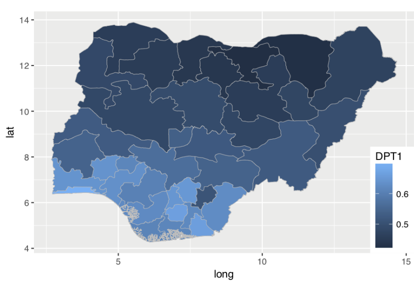



(c) Number of antenatal visits during pregnancy

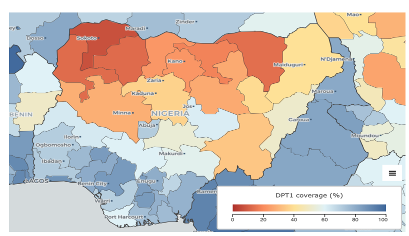



(d) Reason didn't deliver at health facility: too far/no transport

Figure 18: Health care maps of maternal care indicators.

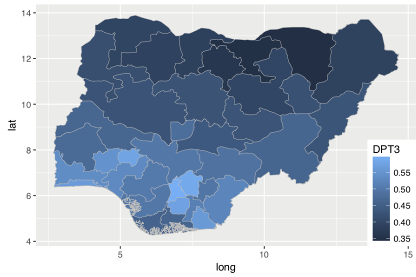



(a) our model predicted map in 2013

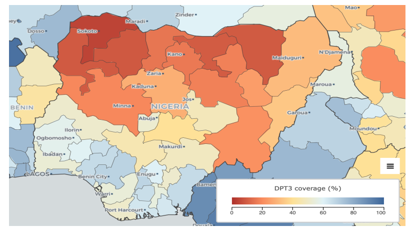



(b) IHME interpolated map in 2016

Figure 19: DPT1 coverage map comparison

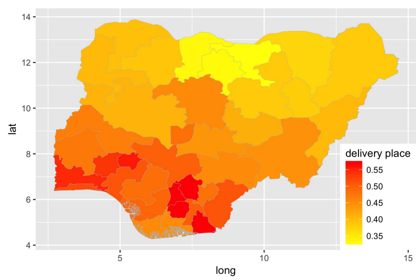



(a) our model predicted map in 2013

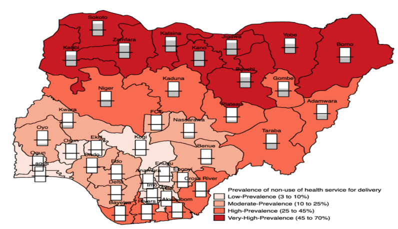



(b) IHME interpolated map in 2016

Figure 20: DPT3 coverage map comparison




(a) our model predicted map for place of delivery in 2013




(b) Percentile map showing the prevalence of non-use of health service for delivery and level of illiteracy among mothers in 2013

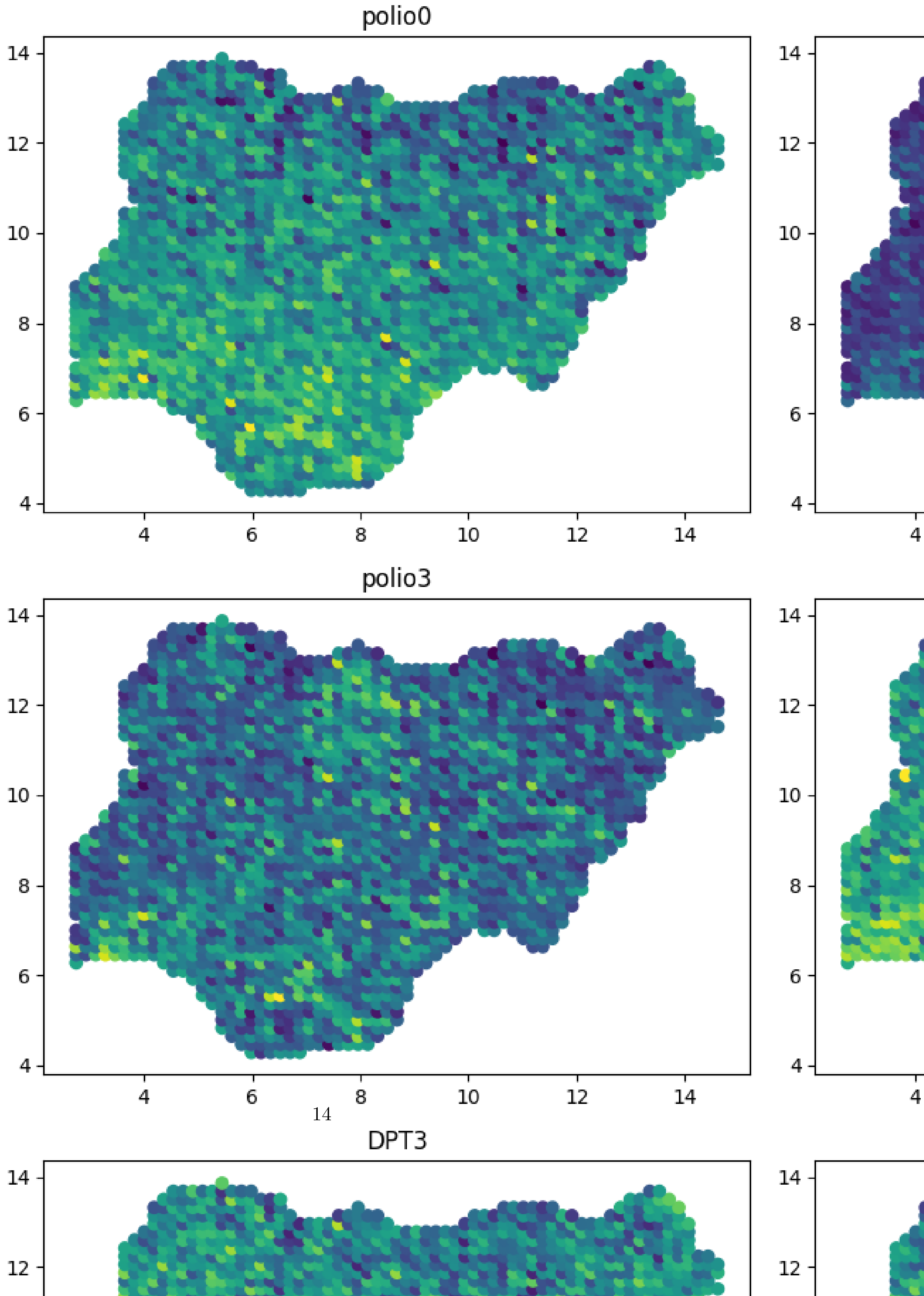
Figure 21: Non-use of health service for delivery map comparison

# 7 Discussion and Conclusion

We extend the model from Jean et al. [2016] to predict health care indices from satellite images, based on which, we make a health map of Nigeria. The model has good predictions of polio0, DPT1, DPT2, DPT3, BCG, measles, health card birth place and number of visits, where the R2 score are above 0.45. However, the model can not achieve good results for polio1-3, vaccination, far and first visit. This is because the tree based models are optimized by minimizing the mean square errors, and we can see that when the reponse in within 0-1, they can achieve comparable MSE for different indices. However, polio1-3, vaccination, far and first visit have smaller variance as we observed in the histogram, and

$$R2 = 1 - \frac{MSE}{Var(y)}, \tag{1}$$

so the R2 score for those healthcare indices are not satisfying. We also compare our maps in 2013 to IHME maps in 2016 for DPT1 and DPT3. It can be seen


polio0
polio3
DPT3

that these maps are consistent with each other. Our map for place of delivery map (Figure 21 is also consistent with the map from a previous study done by AdedokunEmail et al..

One of the future lines could be comparing our health care maps with our social economic maps such as poverty and income to see if the pattern could match with each other. Another interesting direction would be draw the maps for different years, and associate the change of patterns through the years with policies carried out during the time.

# 8 Personalized reflection

## 8.1 Yuan

I think this is a great project. It is a valuable opportunity to work with professionals in the institution and to contribute to a task that could be beneficial to other human being. There was a learning curve when we first set up the all the resources and technologies for computation. It was not an easy process and took us a much longer time than we expected. But everything went back on track after we figuring that out. I appreciate the support of our mentor, who have provided great guidance along the way and also brought help from his peer researcher to answer our questions.

## 8.2 Ruohan

Overall this is a good experience. We learned the whole pipeline, to transfer the satellite data into health care indices. It's a good opportunity to apply what we learn in our coursework to the real things outside. Also our mentor is very supportive, and I much appreciate his help and commitment. I would say the toughest part was when we were figuring out the computation, but it also forced us to learn how to use GPU through Slurm on Sherlock, so at least we learned something, which is good.

# References


World health statistics 2014. 2014.

AdedokunEmail, Uthman Sulaimon, and Olalekan. Women who have not utilized health service for delivery in nigeria: who are they and where do they live?

Akinrinola Bankole, Gilda Sedgh, Friday Okonofua, Collins Imarhiagbe, Rubina Hussain, and Deirdre Wulf. Barriers to safe motherhood in nigeria. 2009.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.