BDS-Assignment3

Author: Jingyi Wu (jingyiw2)

Used Libraries:

Numpy

Pandas

Matplotlib

Regressors

Scipy

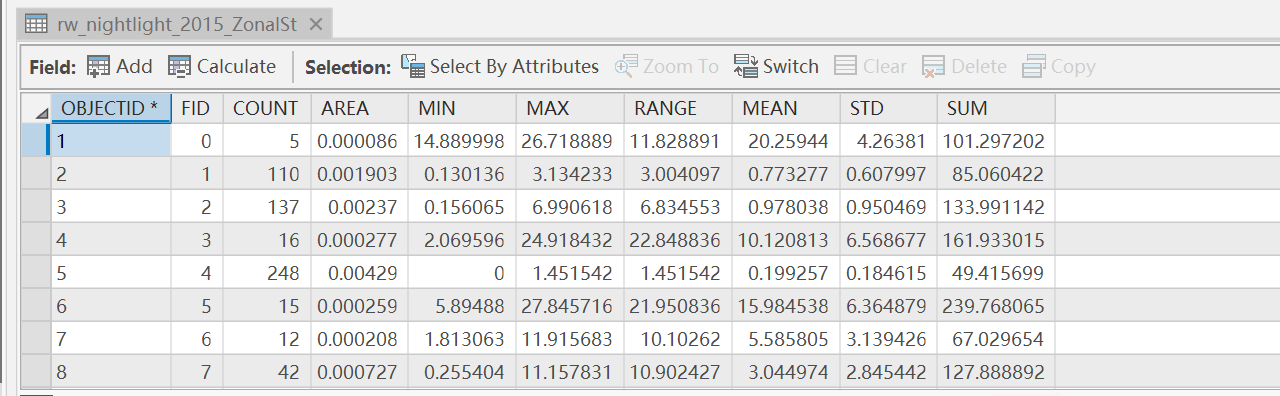
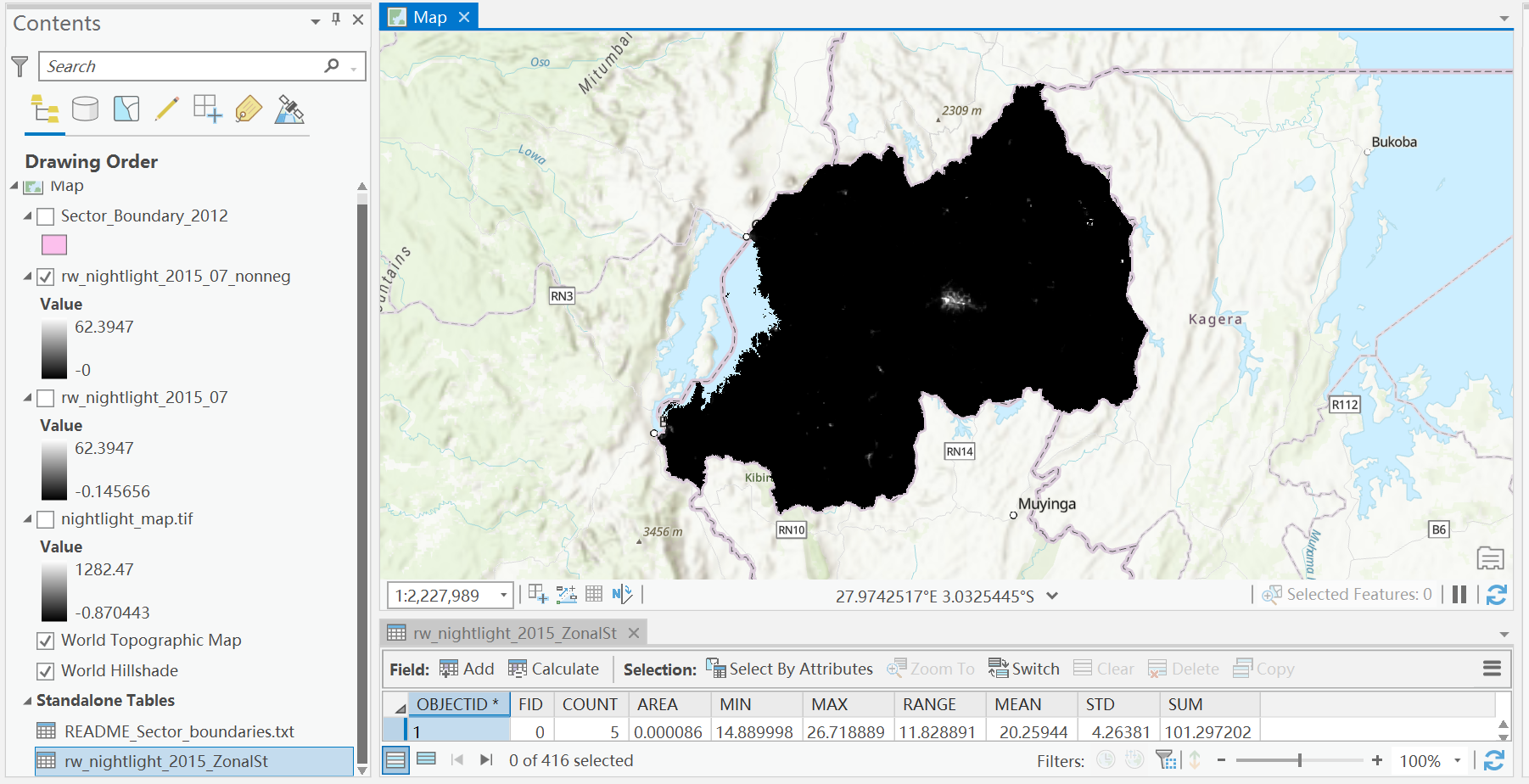
Statsmodel

Sklearn

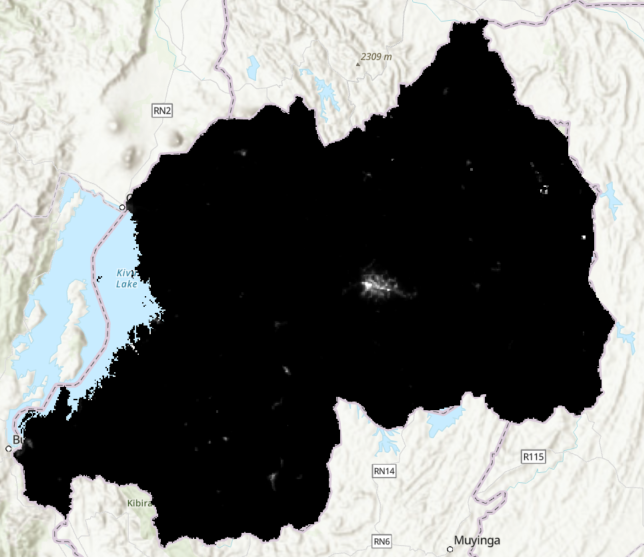
joblib

Q1-Q5:

Following the steps in the tutorials video, we can get nightlight distribution in Rwanda and night\_light sum for all areas. The screenshot of the project is as below.

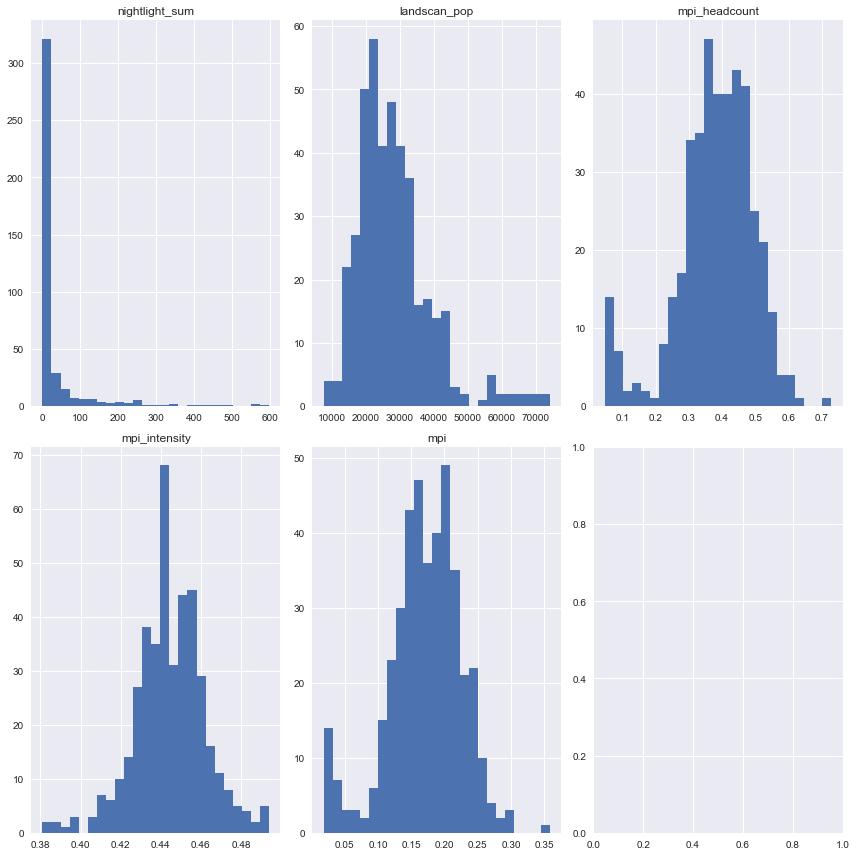


Compare the visualization of map and Google map, I highlighted the light areas with high nightlight values in google map and we can find out that Kigali has the highest nightlight value. Other areas include Ruhengeri (city and capital of Musanze District in the Northern Province), colline mpanga (a place in the province of Kigali), and Butare (a city in Southern province of Rwanda and the capital of Huye district). From the map visualization, the more prosperous the area is, the higher/denser the nightlight is.

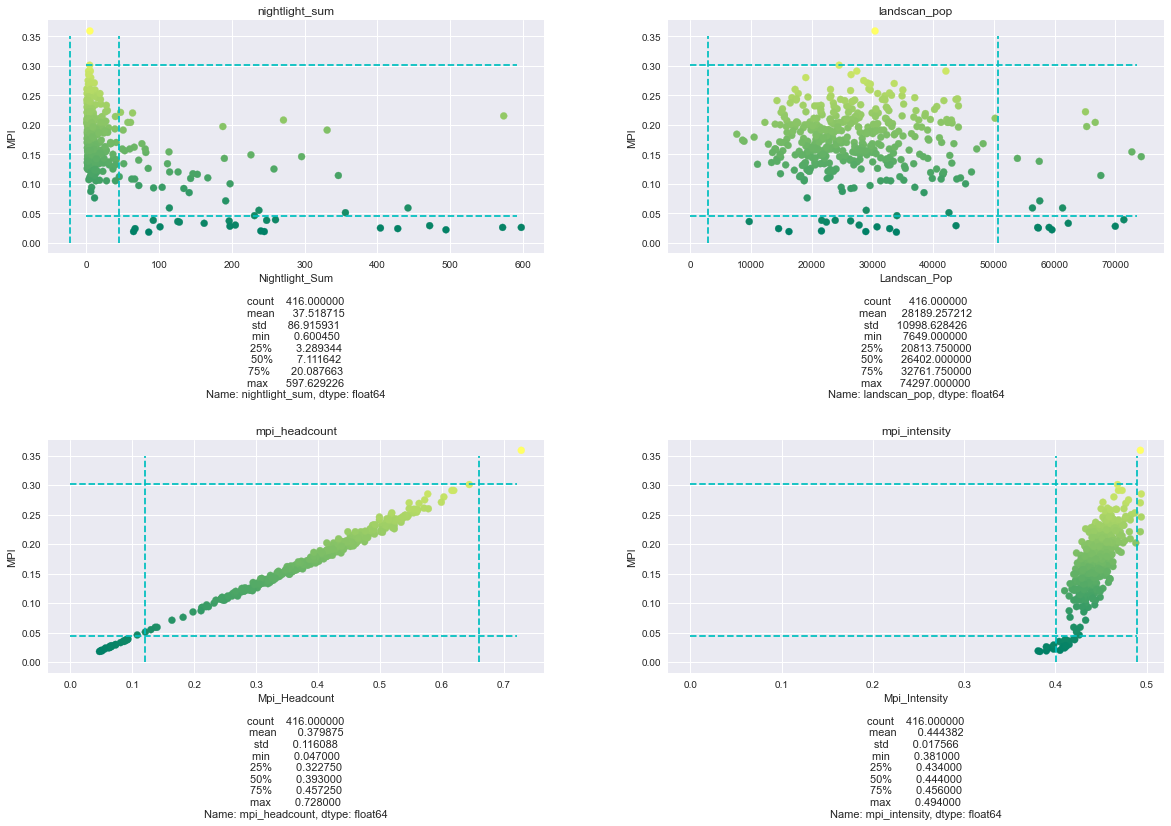


Q6:

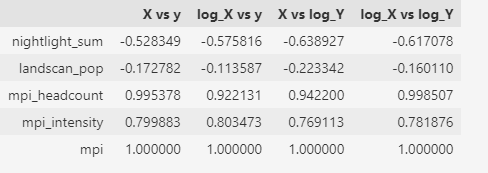
1. Use read\_excel function to load the data file.
2. The histograms of all features are as below. Features except nightlight\_sum can be considered as normally distributed. Nightlight\_sum is right skewed.



1. Scatterplots of all features to MPI are plotted as below.



1. We can see that not all features are linear correlated to MPI. MPI shows a decreasing trend with nightlight increasing. No apparent relationship appears between landscan\_pop and MPI. Other two features, mpi headcount and mpi\_intensity shows linear increasing relationship with MPI.
2. To find out significant outliers, I added upper bound threshold and lower bound threshold dotted lines to the scatter plot. The upper bound is 0.75Quantile+1.5IQR and the lower bound for each feature is 0.25Quantile-1.5IQR. Points outside of these ranges are potential outliers. Maybe these point should be removed when building the model.
3. Correlations for each feature with MPI:

we can see that

for nightlight\_sum, it has strongest correlation with log of MPI.

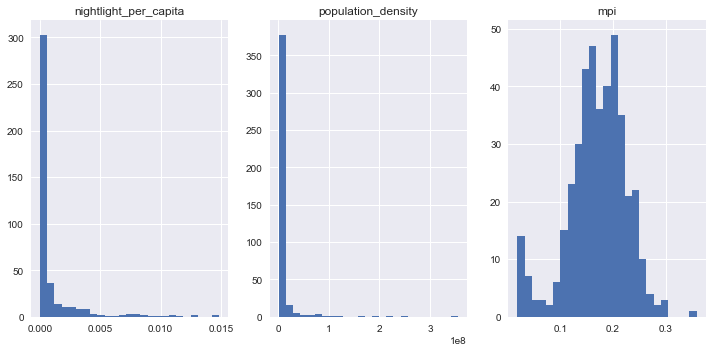
For landscan\_pop, log of it has strongest correlation with log of mpi.

Log of mpi\_headcount has strongest correlation with log of mpi. Log of mpi\_intensity has strongest correlation with mpi.

Q7:

After creating two new features, we plot the histograms for these new features as below.

b. From the charts below, we can see that only mpi is normally distributed and other two features are right skewed.



c. new features’ correlations with MPI are as below.

Night\_light\_per\_capita is strongly correlated with log of MPI.

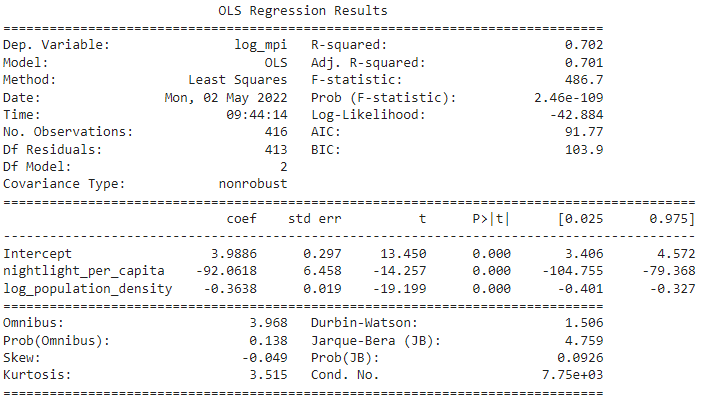
Log of population\_density is strongly correlated with log of MPI.

Q8:

Using the strongest correlation result above, we build model between log of MPI and log of population\_density and Night\_light\_per\_capita to build three models.

**<Backward stepwise>**

Both features are significant in the model. The p-values for them are as below.



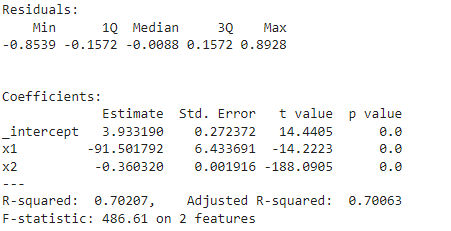
p-values here are all 0.000 and below 0.05, so all features are significant under the significance level α=0.05.

R-squared here is 0.7021 and Adjusted R-squared is 0.70067 for reference.

The p-value for this model here is 0.9999999999999776, not significant under the level α=0.05. Note what we use for testing the significance of the whole model is t-test ind. The result indicates the mean value of predicted log of MPI does not equal to mean of actual log of MPI.

**<Ridge regression>**

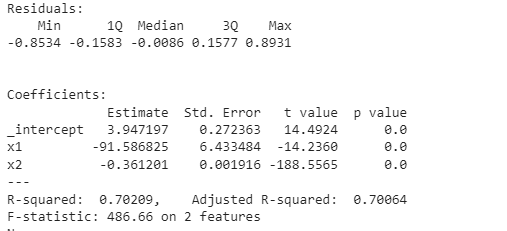
We use RidgeCV to find the best alpha and base our model on that alpha to find out the p-values for each feature.

Under the significance level α=0.05, both features are significant (p-values are all 0.0).

p-value for the whole model is 0.9999999999999774, not significant under the level α=0.05.

**<Elastic Net>**

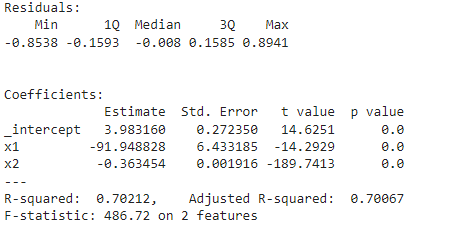
We tried two parameters here for elastic net and select the one with best R2 values.

p-values for both features here are all 0.0 and indicate features are significant under the level α=0.05.

p-value for the whole model here is 0.9999999999999831, indicating the model is insignificant under the α=0.05.

Q9:

Try **lasso regression** with log of MPI and two features and the result is as in the picture.



p-values for two features are all 0.0, below 0.05, indicating two features are significant under the level α=0.05.

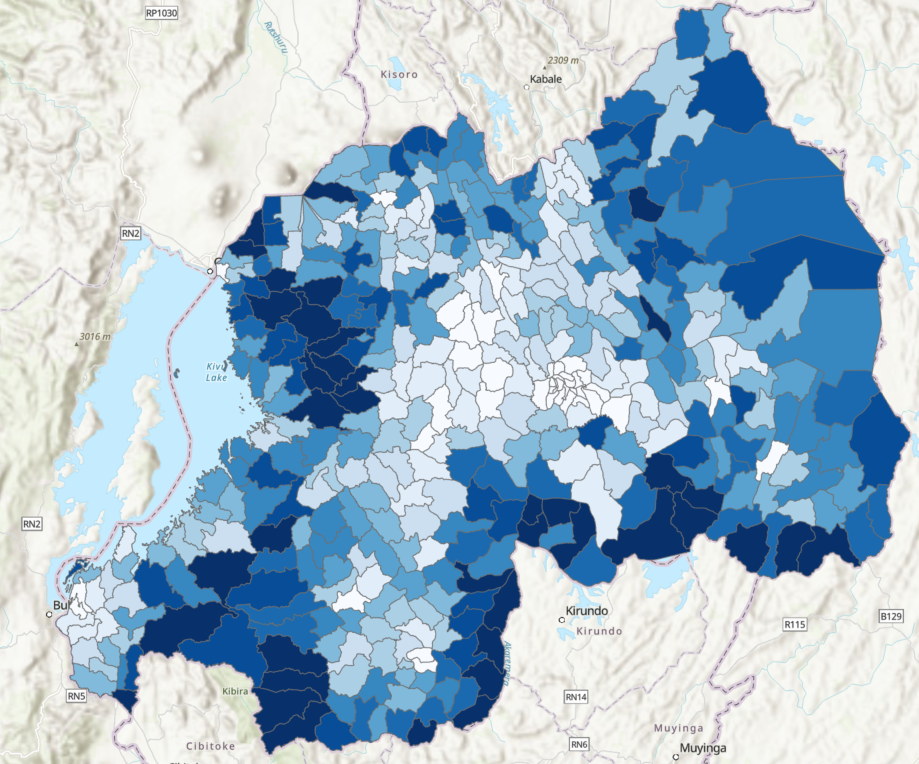
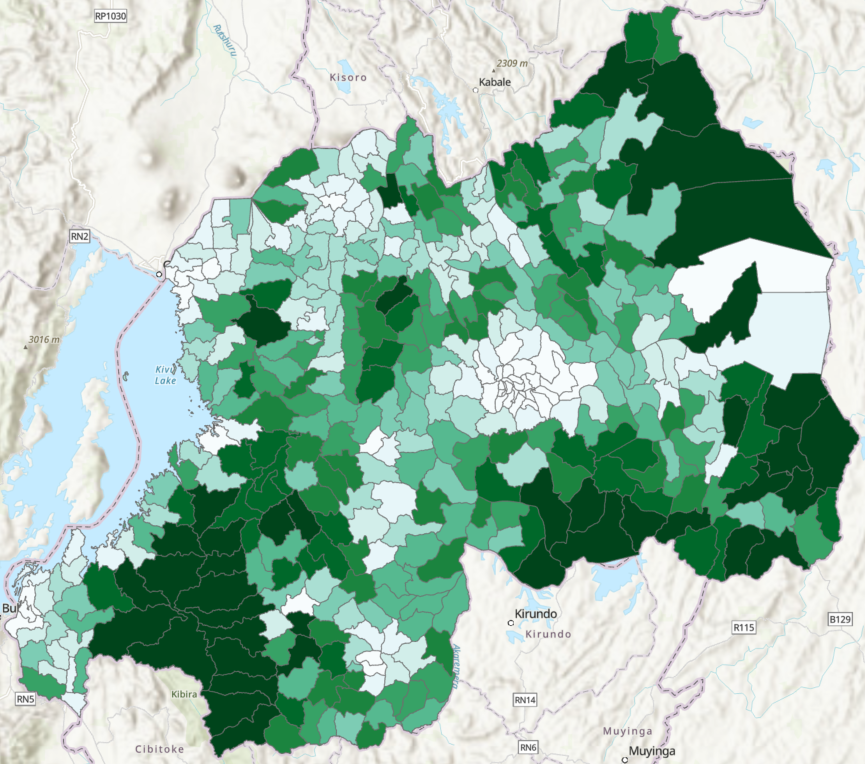
The correlation between predicted values and actual log of MPI is 0.83792, indicating the predictions are highly correlated with actual values. The model has a good predictability.

R-squared value here is 0.70212, and compared this value with the previous R2, we can see that lasso is slightly better than other models. The lasso model’s predictability is good.

Q10:

In ArcGis, visualize the actual log of MPI and estimated log of MPI and the graphs are as below.

Actual (left), Estimated(right):

Compare the two maps, we can see they are somewhat similar but yet different in some sectors. In general, our estimation get the MPI pattern correctly. The general pattern here is central part with low MPI and some sectors highlighted in red circle have a quite high MPI. Kigali in both maps display rather low MPI value.

But some sectors in the central part of Rwanda have a higher estimated MPI. Some sectors, for example those highlighted in yellow, actually have a quite low MPI but are estimated quite high. Also, some sectors which have in fact have high MPI are estimated low.