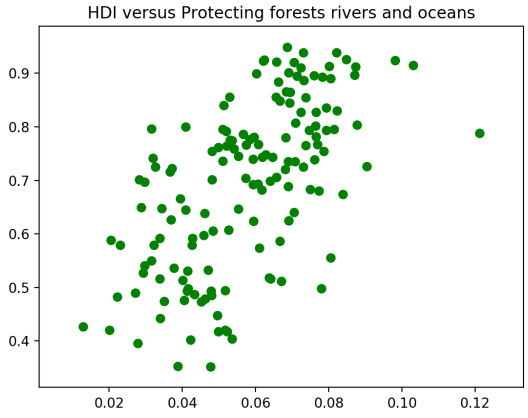
3b

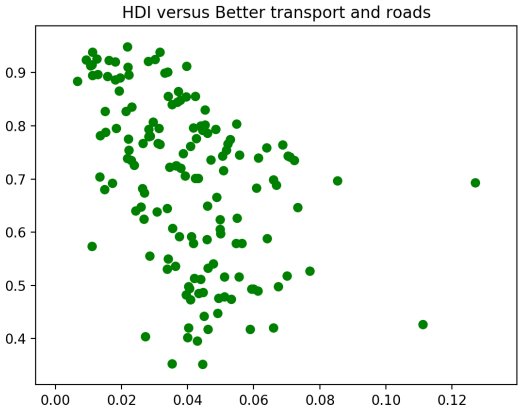
Most positively correlated: Protecting forests rivers and oceans 0.613458756271

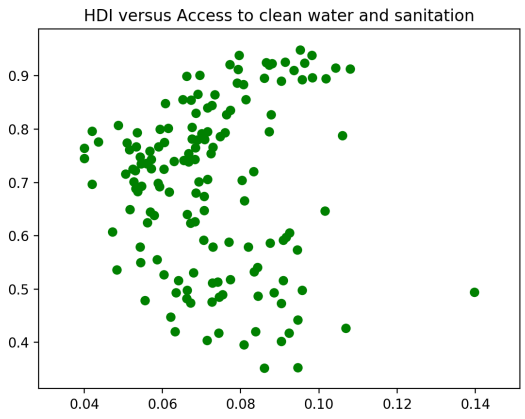
Most negatively correlated: Better transport and roads -0.439633638622

Least correlated: Access to clean water and sanitation: -0.018169084456

3c

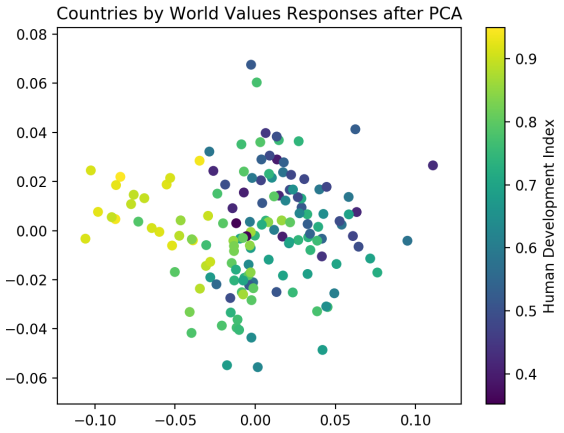






We can observe positive correlation, negative correlation and almost no correlation in the three figures, respectively.

3d


Countries by World Values Responses after PCA
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.10
−0.05
0.00
0.05
0.10
0.9
0.8
0.7
0.6
0.5
0.4
Human Development Index

3e

Ridge Regression

RMSE: 0.12303337350607803

Coefficients
[[0.80823467 -0.74985758 -0.17800015 -1.28408103 -0.66293176 -0.82203172
   0.73733884 -0.92891581 -0.82049672 0.39614952 2.0708291 -0.06718981
   0.48310656 0.72671425 0.42921192 -0.13808023]]

3f

Lasso Regression

RMSE: 0.12602242808947528

Coefficients
[0.1590192 -0.72844929 -0.   -0.85945074 -0.66274144 -0.02556703
  0.33904781 -0.29897158 -0.   0.   3.48536375 0. 0.
  0.87057995 0.32897045 -0.   ]

3g

Lasso weights have 6 zeros, while Ridge have none.

3h

Average the outputs of its k nearest nerghbors. Can also be weighted by inverse distance.

3i

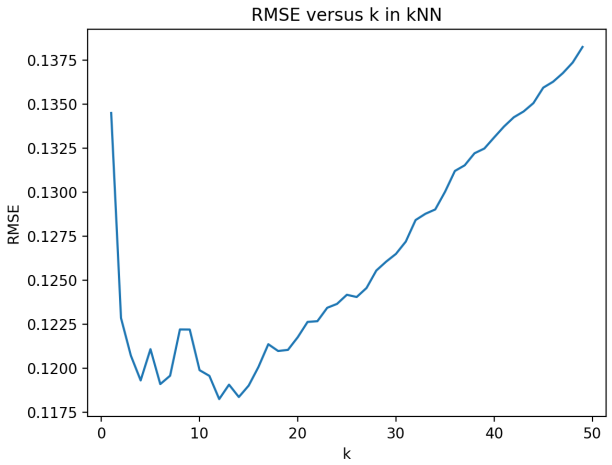
```
The closest countries are:
45      United States
90            Ireland
61     United Kingdom
37            Belgium
108           Finland
69              Malta
132           Austria
110            France

The distances are:
0.          0.01765082  0.02157067  0.03010195
0.0302092   0.03349605
   0.03583628  0.0389719
```

3j

k Nearest Neighbors Regression

RMSE: 0.11824589460776892

Best k: 12



3k

When k is large, we sampled more before making a decision, this will lead to small variance. So large k, small variance, small k, large variance.

When k is large, we tend to include more samples from other classes, this will bring in bias. So large k, large bias, small k, small bias.

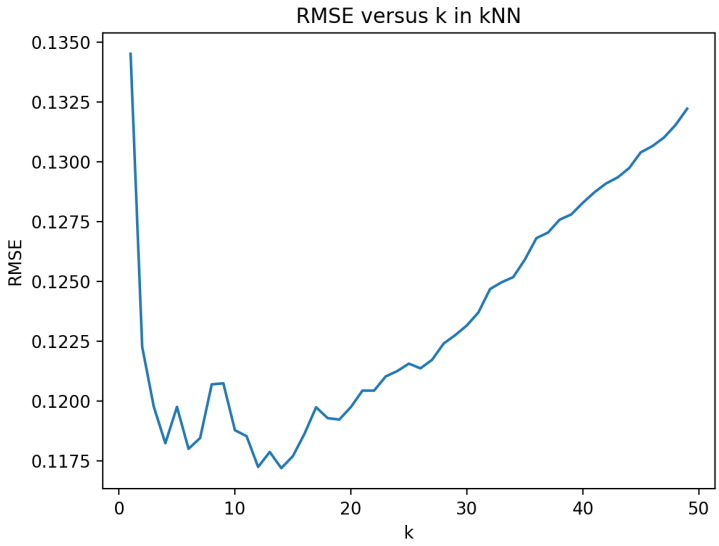
So we will have a sweet spot where the sum of bias and variance is the smallest.

3l

k Nearest Neighbors Regression: inverse distance weighted

RMSE: 0.1171925270311745

Best k: 14



3m

```
45      United States
90             Ireland
61     United Kingdom
108            Finland
37             Belgium
69               Malta
110             France
132            Austria


 [[ 0.          1.1895638   1.37459492  1.81818296
1.94193658  2.28988047
   2.29063682  2.39411947]]
```

Compare to unscaled result, Finland and Belgium exchange positions, and France and Austria exchanged positions.

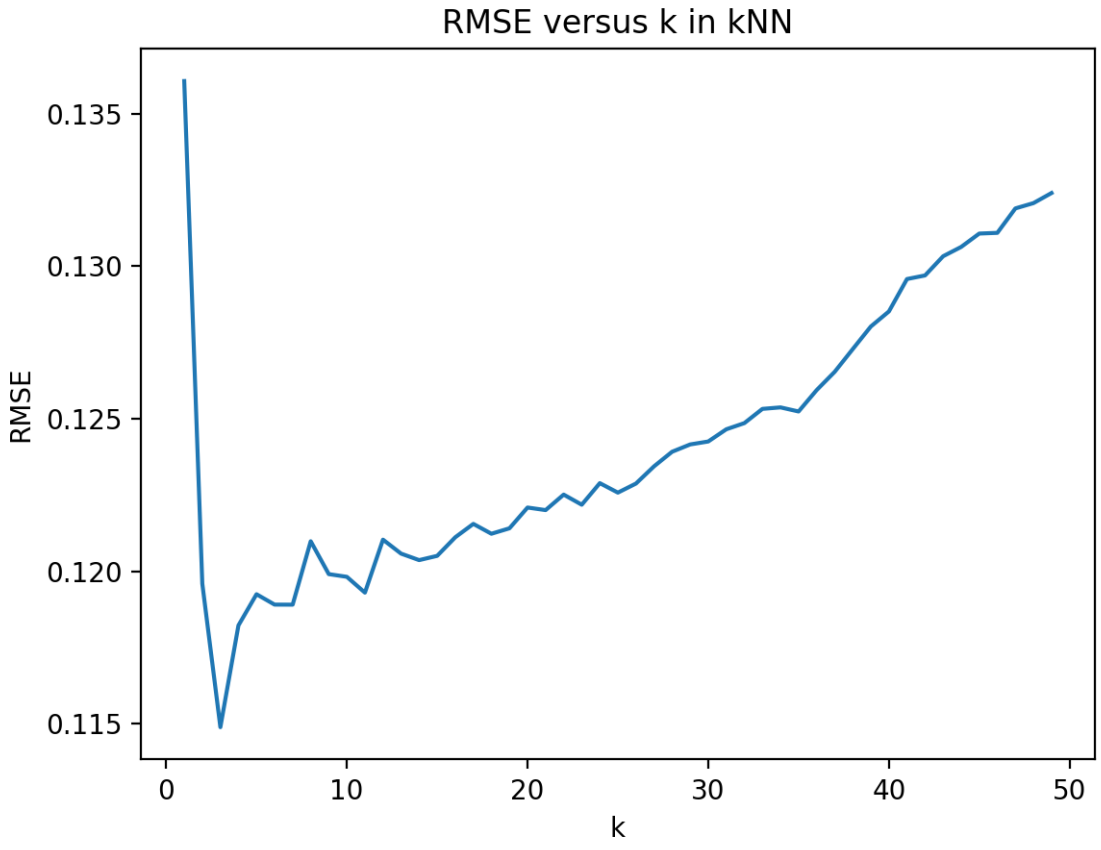
3n

k Nearest Neighbors Regression: with weights, with scaling

RMSE: 0.11488547357936414

Best k: 3

Grid search should be performed again every time we changed the setting.



3p

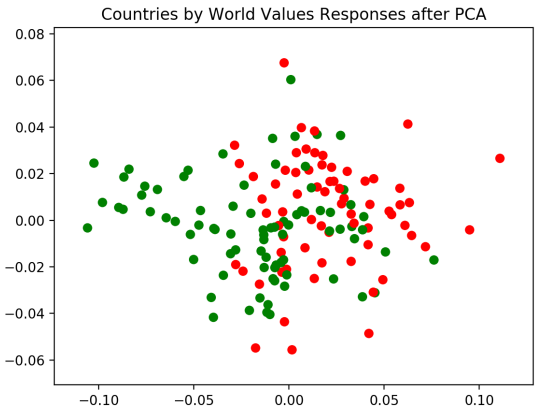
Test predictions:

[[0.5109997]
 [0.61570753]
 [0.68331113]
 [0.67852229]
 [0.7718753]
 [0.60419024]
 [0.65159897]
 [0.67945579]
 [0.73105346]
 [0.8053592]
 [0.67938174]
 [0.66520881]
 [0.7793257]
 [0.70584762]
 [0.66568196]
 [0.56691101]
 [0.69252235]
 [0.72706267]
 [0.70498736]
 [0.77513888]
 [0.63384268]
 [0.65089538]
 [0.39709827]
 [0.73442026]
 [0.7583716]
 [0.71157266]
 [0.44080564]
 [0.90460366]
 [0.62818537]
 [0.65208228]
 [0.6630599]
 [0.86539214]
 [0.72870152]
 [0.90989302]
 [0.924362]
 [0.8432591]
 [0.49853809]
 [0.80810577]]

3q

At least we are guaranteed a 1/k accuracy. Since the largest class should includes more than N/k samples.

3r



3s

A linear boundary cannot easily separate the two classes.

3t

SVM Classification(kernel='linear', C=48.0)

Accuracy: 0.75

3u

SVM Classification(kernel=’linear’, PCA_k=8, C=1.0)

Accuracy: 0.77027027027 (improve a little)

3v

SVM Classification(kernel=’rbf’, C=73, gama=0.097)

Accuracy: 0.743243243243

3w

k Nearest Neighbors Classification(weights='distance', k=4)

Accuracy: 0.763513513514

k Nearest Neighbors Classification(weights='distance', k=4, scaled)

Accuracy: 0.77027027027(scaling helped a little)

3x

3y

If we are given the target locations and the distance from each target for each sensor, then for each sensor, find its k nearest neighbors, and do a weighted average of the locations of the k neighbors.

The problem with this is that, if all targets are on one side, and all sensors are on the other side, then the prediction by kNN will be that all sensors are amid the targets, which is not correct.

3z

I learned that kNN can also be used for regression, and in some cases works better than ridge and lasso.

For SVM classification, ‘rbf’ is not always better than ‘linear’.

Scaling and standardization is important.