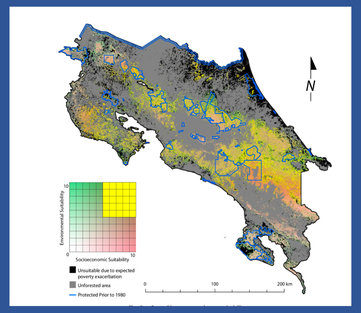
Deforestation in Costa Rica : Classification and Prediction

Polina Koroleva, Xinrui Zhong, Junyao Gu

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

BACKGROUND

Deforestation is a major threat to biodiversity and ecosystems in Costa Rica. The government has made substantial investments in its protected area systems, as well as in forest monitoring and data collection. The collected data can be used to predict the areas of higher deforestation risk. Assigning limited conservation resources to these areas can help to reduce costs and time of the process.



Source: Ferraro et. Al (2011)

OBJECTIVE

- Find the missing labels
- Build prediction models with different classifiers
- Examine dynamics of deforestation through the years
- Find the best predictors
- Compare the outcomes of the analysis using entire dataset and dimension-decreased dataset
- Forecast the deforestation

Original Dataset

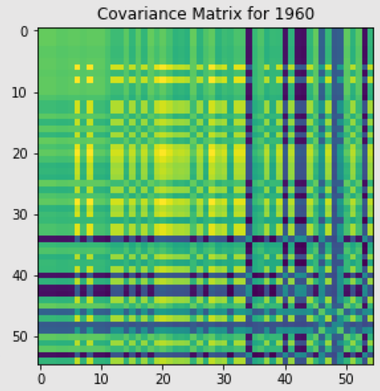
- The original data set contains 47,107 forest units characterized with 98 variables
- The shape of the cleaned dataset is: 39,778 * 105
- Outcome variable: forested (1) and deforested (0)
- The features (variables) used for prediction are numerical
- Prediction: 3 subsets for 1960, 1986, 1997

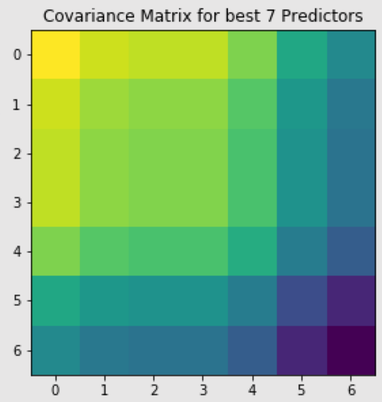
Label	Deforested (0)	Forested (1)
1960	16487	23291
1986	22189	17589
1997	21100	18678

Methods

1. Data preprocess:
   a) Normalize
   b) Remove NaN / missing values
   c) Create 3 subsets for different years
   d) Reduce dimension with PCA
   e) Spectral Clustering
2. Classification within each subset
3. Prediction model
   a) Predict 1986 deforestation using 1960 data
   b) Test model with 1986 data and 1997 labels
   c) Predict 2020

Prediction Error Rate	1986	1997	1986 with top 7 indexes
GNB	36.22%	33.95%	29.99%
QDA	28.48%	37.01%	29.96%
LDA	24.40%	25.69%	27.13%
KNN	20.71%	21.58%	22.92%
DT	24.29%	21.57%	25.08%
RF	19.98%	17.66%	21.87%





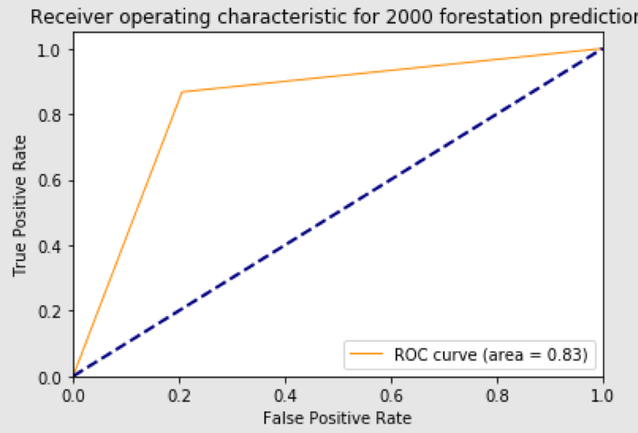
Analysis

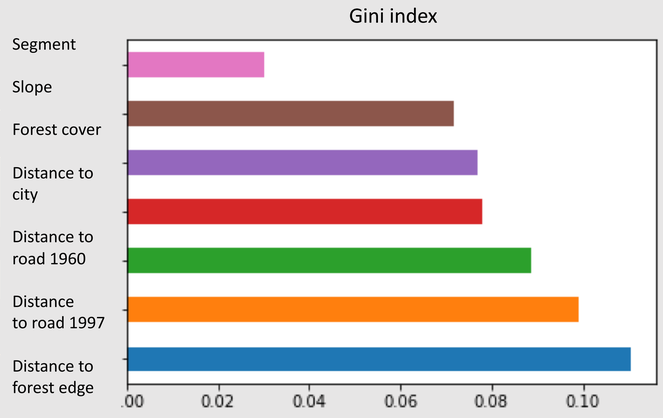
Confusion Matrix of 1986 Predictions (top 7 indexes)

	Actual Labels	
Predicted Labels	0	1
0	**6233**	**983**
1	**1888**	**4023**

Confusion Matrix of 1997 Predictions

	Actual Labels	
Predicted Labels	0	1
0	**18995**	**2105**
1	**2922**	**13756**





Results

Classification average accuracy: **93.7%**
Random Forest gives the best prediction accuracy.
Prediction of 1986: **81.02%**
Prediction of 1986 using the 7 best predictors: **79.13%**
Prediction of 1997 using the same model: **82.34%**

Future Development

1. Cost benefit analysis
2. Deep Learning techniques
3. Fill in missing data by imputing
4. Create different classifiers by training with different number of variables to deal with missing data

Conclusions

Classification model helps to label missing data.
Prediction model accuracy is satisfactory and can be used for future predictions. The variables amount can be reduced without significant harm to accuracy of prediction. The model will help to decrease the cost of the data collection and analysis. It can be used to target the areas of higher deforestation risk more effectively. Our forecast for 2020 is that 22372 of the units will be deforested (0) and 17407 covered by forest (1), suggesting there will be less forested area in Costa Rica.

Acknowledgements

Data is provided by Dr. Paul Ferraro, Environmental Health & Engineering Department , JHU and Dr. Merlin Hanauer, Department of Economics, State University of Sonoma
Ferraro et al. (2011) Conditions associated with protected area success in conservation and poverty reduction, *PNAS*,10.1073/13913–13918

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com