LAB #8: LAND USE, LAND COVER ANALYSIS USING SATELLITE IMAGES

CS 109A, STAT 121A, AC 209A: Data Science

Fall 2016

Harvard University

ANNOUNCEMENTS

Helpful things to keep in mind for homework, project, midterm and beyond this class:

1. **There is no "right" answer:** every model is flawed
2. **There is no "perfect" data:** all data is incomplete and noisy
3. **Nothing is certain:** there are *only* heuristics and choices, not rules (not for when to regularize, not for which variables to select, not for choosing hyper-parameters or priors, not for which loss function to minimize)
4. **You are only and always accountable to the data and the task:** a good model is one that include full disclosure of choices you made, good rationale for making them and honest discussion of their potential drawbacks.

LAND USE AND LAND COVER ANALYSIS

Land cover or land use analysis is the task of broadly classifying features on land and how land is used.

This analysis includes:

- Identifying types and locations of natural features of the land
- Identifying types and locations of man-made features of the land
- Quantifying the amount of resources on land and intensity of use
- Identifying patterns, cases and scales of change

LAND USE AND LAND COVER ANALYSIS

Land cover or land use analysis is the task of broadly classifying features on land and how land is used.

What's it good for?

- Identifying inefficiency of use
- Identifying inequity in distribution of resources
- Anticipating future needs or conflicts

This analysis is import in informing policy decision of regulatory or planning bodies and for stakeholders with special interests (environmental preservation, economic development etc).

LAND USE AND LAND COVER ANALYSIS

Land cover or land use analysis is the task of broadly classifying features on land and how land is used.

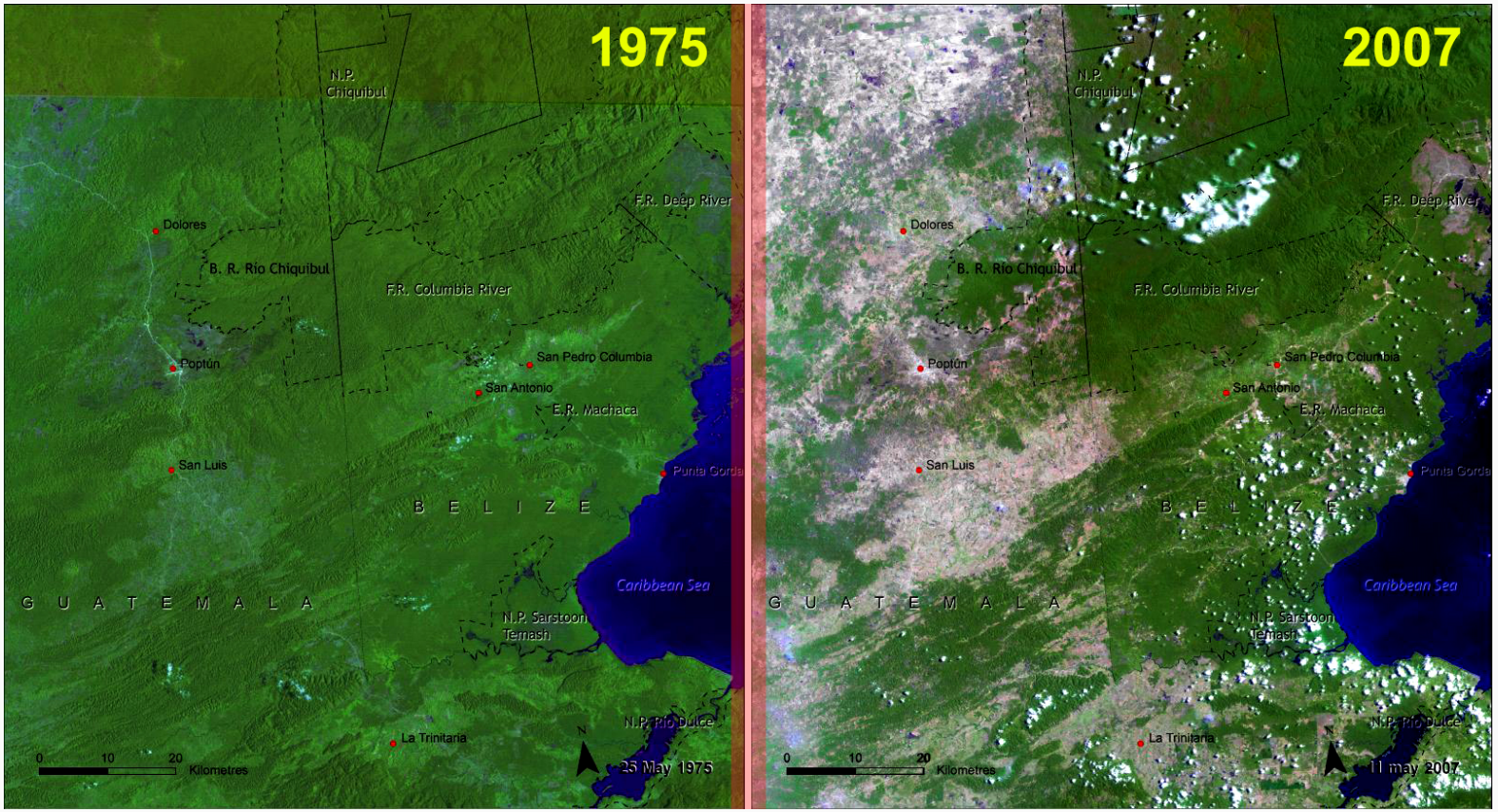
How is it done?

- Aerial photographs or satellite images are interpreted (cross referencing with census and other types of data)
- Follow-up ”windshield surveys” or field studies are conducted for accessible areas

For large areas or areas for which we have many images, we’d like to interpret the features in these images **automatically**.

LAND USE AND LAND COVER ANALYSIS

What changed in the land use/cover? What's driving the change? What are the implications of these changes?



CLASSIFICATION FOR LAND USE, LAND COVER ANALYSIS

Question: What does this have to do with classification?

Given a satellite image and a small sets of locations that have been labeled (as forest, water or farmland etc) by hand. The goal is to use the labeled locations to train a model that will predict the label of a new location.

CLASSIFICATION FOR LAND USE, LAND COVER ANALYSIS

You'll be working with a simplified land user/cover analysis task: classify locations in satellite images as **vegetation** or **non-vegetation**.

You have a set of data from satellite aerial images. The first two columns contain the normalized latitude and longitude values. The last column indicates whether or not the location contains vegetation, with 1 indicating the presence of vegetation and 0 indicating otherwise.

Question: Which method of classification should you use and **why**?

Question: What's the final word? Which model is better for identifying regions with vegetation in satellite images?

Question: What's the final word? Which model is better for identifying regions with vegetation in satellite images?

But wait! Before you answer this question, think about what we mean by "better" in the context of this problem? What are the challenges of this task (identifying regions with vegetation in aerial images) and which model is more equipped to handle these challenges?

Question: Which splitting criterion should we use? Why?

1. classification error
2. Gini index
3. Cross entropy

Let's consider a simple example: 100 points in a satellite image, of which 51 are class 1 and 49 are class 0. Choose one these splits:

- Region 1: (11, 37), region 2: (40, 12)
- Region 1: (25, 48), region 2: (26, 1)

Which split is intuitively "better"? Which one will be picked by each of the split criteria?