APPLYING BIG DATA TECHNOLOGY TO REMOTE SENSING FOR SPECIES IDENTIFICATION

By

MORTEZA SHAHRIARI NIA

A DISSERTATION PROPOSAL PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHYLOSOPHY

UNIVERSITY OF FLORIDA

2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Proposal Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Phylosophy

APPLYING BIG DATA TECHNOLOGY TO REMOTE SENSING FOR SPECIES IDENTIFICATION

By

Morteza Shahriari Nia

April 2015

Chair: Dr. Daisy Zhe Wang
Major: Electronics and Computer Engineering

Ecological sciences benefit from the huge diversity of plant species which play an important role in large scale ecological aspects such as global warming, land cover change, CO^2 emission, invasive species, fire hazard, and etc. State-of-the-art species classification techniques utilize remote sensing data such as hyperspectral and LiDAR, however this task involves plenty of field data collection which is both highly time consuming, costly and can only be accomplished by ecological experts. Among thousands of the most commomly found plant species there is huge similarities between them from a remote sensing point of view which makes the task of species classification very daunting; therefore we see a whole body of literature specifically dedicated to this issue which is yet far from real world scenarios with thousands of possible species. While this is an indicator of the importance and complexity of the issue, little has been done to tackle the problem from a computational point of view harnessing the power of "big data". Periodic airborne campaigns can generate terrabytes of data on vast swaths of land. To tackle these problems we propose to use probabilistic knowledge bases and deep learning both of which work best when there is lots and lots of data. Probabilistic knowledge base captures ecological expert knowledge in terms of probabilistic rules, which will be maped to remote sensing data and used to infer new facts and therefore enhance species classification accuracy. Deep learning on the other hand as a semi-supervised algorithm will benefit

from the vast amounts of data available and capture intrinsic features of data through its layered architecture and thus help in reducing the amount of labeled data required.

CHAPTER 1
INTRODUCTION

Understanding the dynamics of ecological structures is very important in determining how climate, land cover , fire hazards, and biodiversity evolve. Precision study of plant species is of high environmental and economical impacts which is only possible through geo-mapping the distribution of plant species abundances at ecological scale. Large scale study of ecological domains has been made possible through spaceborne or airborne campaigns utilizing remote sensing technologies such as *(multi/hyper)-spectral* and *LiDAR*. In this project we focus on airborne hyperspectral and LiDAR data. Each campaign covering tens of acres of land can generate terra-bytes of data depending on measurement resolution (large volume). On the other hand, apart from state-of-the-art machine learning algorithms, there is a great wealth of expert ecological knowledge covering a whole variety of domains (along with their in-ground associated data) that can be used to enhance species mapping that is not being used and is left for ecological scientists for manual interpretation (data variety). Furthermore, data is being generated at faster pace day after day as technology becomes more afordable. After satellite sensors, airborne sensors came into place and now as airborne is still costly there is a surge of interest towards more affordable drone campaigns (Zhou et al., 2009). So we are facing data being generated at unprecedented rates (data velocity). The final aspect is verasity: imperfect sensors, non-standardized measurements, atmosphere impacts (clouds, humidity, aerosols) and et cetera all create uncertainities that need to be accounted for. Velocity, verasity, volume, and variety are the four V's that indicate ecology is stepping into the realm of "big data"(Hampton et al., 2013; Soranno and Schimel, 2014).

1.1 Remote Sensing

From an ecological point of view, there are two types of remote sensing approaches: active and passive. *Passive* remote sensing uses sunlight as the source of energy and sensors captures the intensity of light being reflected from earth's surface. Light intensity measurements happens at various wavelengths; if a few (usually 3 to 10) relatively broad wavelength bands are captured it is called multi-spectral. If light intensity at dozens to hundreds of narrow band signals are collected it is called hyperspectral. *Active* remote sensing on the other hand uses laser light emission as its source of energy and captures the intensity of returned signals. LiDAR is a popular active remote sensing technology. Below we explain each in more details:

1.1.1 Hyperspectral

Spectrometers measure the amount of light reflected from surface materials: An optical dispersing element (like a prism) refracts the received light into its constituent spectrums and the energy in each band range is measured by a separate detector. Bands can be as narrow as 0.01 m over a wide wavelength range of typically 0.4 to 2.5 m. Figure 1-1 shows the basic components of an imaging spectrometer.

Raw sensor readings (digital number) can be affected by light source conditions, sensor, atmosphere, and surface material. Raw data which is a unit-less light intensity measure is then calibrated into radiance which has a physically meaningful unit through applying a gain and offset to the pixel values. It essentially means how much light the instrument "sees" from the object being observed. Some reference materials like a pure

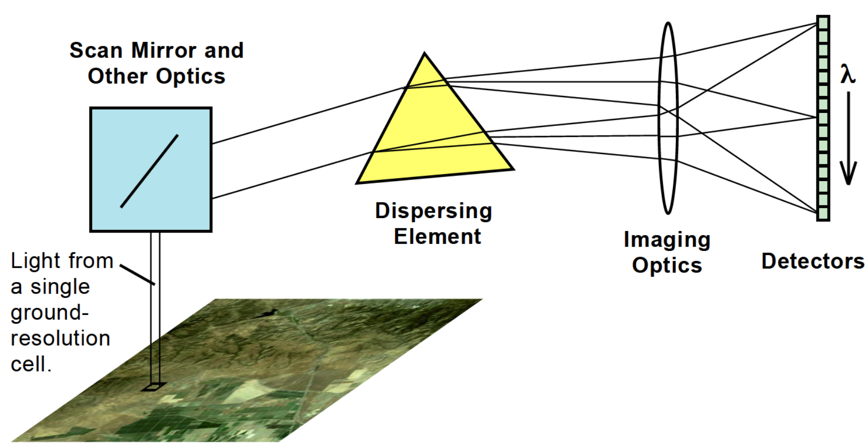

Figure 1-1. Schematic diagram of the basic elements of an imaging spectrometer where λ is the wavelength (Smith, 2006).

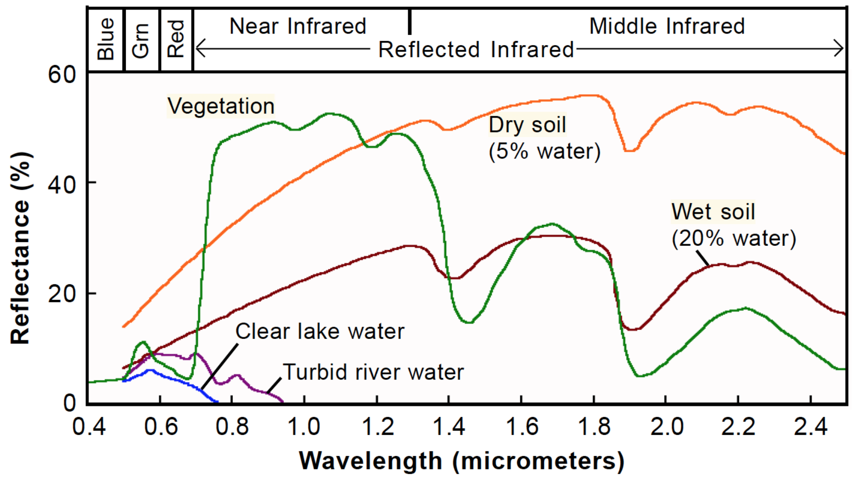

Figure 1-2. Some reflectance examples as how reflectance of different material show different absroption features at differnet bands. (Smith, 2006).

white or pure black sheets can be used in this process. After adjustments for sensor, atmospheric, and terrain effects are applied, pixel reflectance value is calculated which is the proportion of the radiation striking a surface to the radiation reflected off of it. Reflectance demonstrates light abroption features of the surface material and can be compared with field or laboratory reflectance spectra in order to recognize and map surface materials such as particular types of vegetation or diagnostic minerals associated with ore deposits. Reflectance varies with wavelength for most materials because energy at certain wavelengths is scattered or absorbed to different degrees (Smith, 2006). In this project we deal with reflectance values and refer the reader to (Varshney and Arora, 2004) for more details on how to compute reflectance values.

Figure 1-2 shows some example materials when observed through an spectrometer. In vegetations, chlorophyll and some leaf pigments show high absorptions in blue and red ranges and not so much in green; therefore our eyes see vegetation as green. We can see this as a small peak in green compared to other visible wavelength range. From red to near infrared there is a sharp rise known as *red edge* up to a value of about 50% for some plants. High values in the near-infrared region is mainly due to internal cellular structure of leaves which differs significantly across species but can also be different in a single specie due to plant stress. High reflectance in near-infrared can interact with other leaves in the canopy and therefore its sensor readings can be dependent on canopy structure as well. Beyond 1.3 m reflectance decreases with increasing wavelength, except for two pronounced water absorption bands near 1.4 and 1.9 m. At the end of the growing season leaves lose water and chlorophyll. Near infrared reflectance decreases and red reflectance increases, creating the familiar yellow, brown, and red leaf colors of autumn (Smith, 2006).

1.1.2 LiDAR

1.2 Data Variety

As of

1.2.1 Markov Logic Network

Markov logic network is a probabilistic logic that applies the concept of Markov network to first order logic. For inference, instead of usning intractable algorithms of prolog or lisp, it uses MCMC sampling.

1.3 Proposed Work

1.4 Proposal Structure

CHAPTER 2
SPECIES CLASSIFICATION

2.1 Species Classification using SVM

2.2 MESMA, SPICE, PCOMMEND

2.3 LiDAR Stack and Field Statistics

CHAPTER 3
INFORMATION EXTRACTION FROM TEXT

CHAPTER 4
BIG DATA TECHNIQUES

In this chapter we introduce the tools that we will be using to accomplish the proposal.

4.1 Markov Logic Networks

as first order logic software systems become intractable with not so large set of rules and without even probabilities mln adds probabilities and uses mcmc to tackle scalability.

4.2 Deep Learning

use large amounts of data at hand.

CHAPTER 5
PROPOSED WORK

In this chapter we demonstrate how we use mln and deep learning for species classification.

REFERENCES

Hampton, Stephanie E, Strasser, Carly A, Tewksbury, Joshua J, Gram, Wendy K, Budden, Amber E, Batcheller, Archer L, Duke, Clifford S, and Porter, John H. "Big data and the future of ecology." *Frontiers in Ecology and the Environment* 11 (2013).3: 156–162.

Smith, Randall B. "Introduction to hyperspectral imaging." *Microimages. Retrieved on June* 30 (2006): 2008.

Soranno, Patricia A and Schimel, David S. "Macrosystems ecology: big data, big ecology." *Frontiers in Ecology and the Environment* 12 (2014).1: 3–3.

Varshney, Pramod K and Arora, Manoj K. *Advanced image processing techniques for remotely sensed hyperspectral data.* Springer Science & Business Media, 2004.

Zhou, Guoquing, Ambrosia, Vince, Gasiewski, Albin J, and Bland, Geoff. "Foreword to the special issue on Unmanned Airborne Vehicle (UAV) sensing systems for earth observations." *Geoscience and Remote Sensing, IEEE Transactions on* 47 (2009).3: 687–689.