Image and attribute based identification of Protea species using machine learning techniques

Peter Thompson

Supervisor: Dr Willie Brink

Stellenbosch University, Applied Mathematics Division

Introduction

South Africa is very rich in plant species with roughly 24,000 taxa, of which 80% are endemic.

- Cape Floristic Region (mainly fynbos) contains 9,000 of 24,000 taxa in a 6% area
- Genus Protea is archetype of fynbos
- 80 Protea species in fynbos
- How to identify them?

Figure 1: *Protea magnifica*

Data

Protea Atlas Project (PAP)

- Ran for 10 years and headed by Dr Tony Rebelo
- 150,000 species records at 62,000 localities
- Includes location, elevation, flowering times, numbers etc.

iNaturalist

- Natural continuation of PAP
- Amateur botanists upload pictures of species, with added metadata, i.e. location, flowering etc.

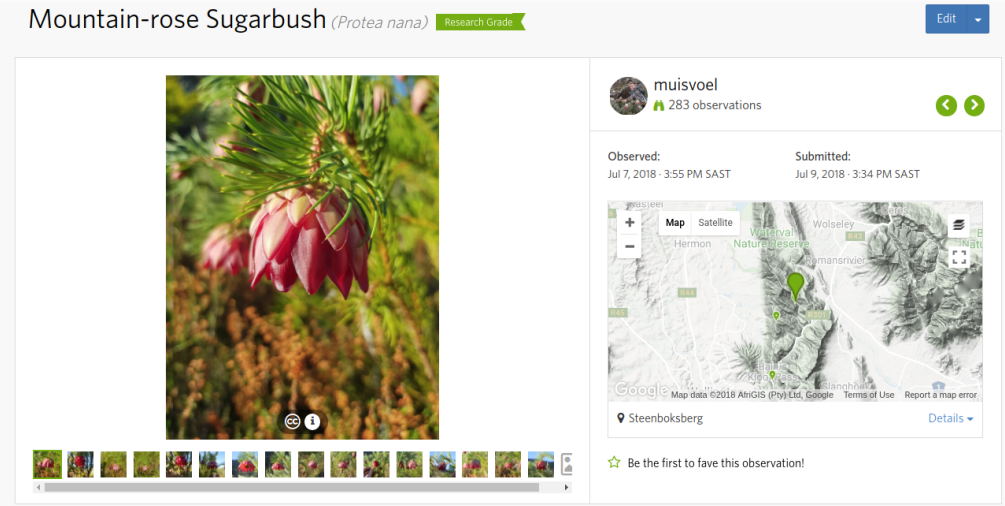
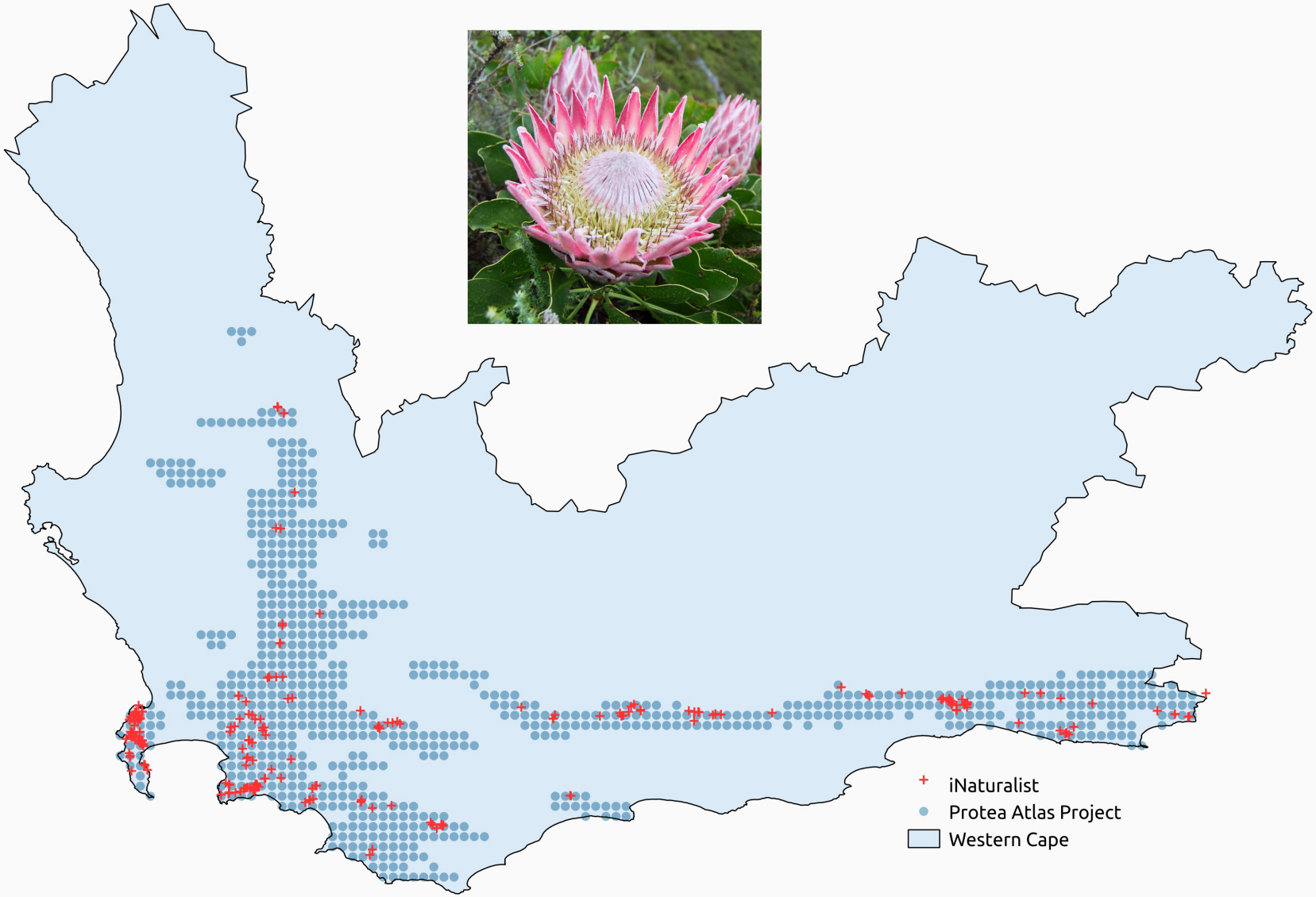
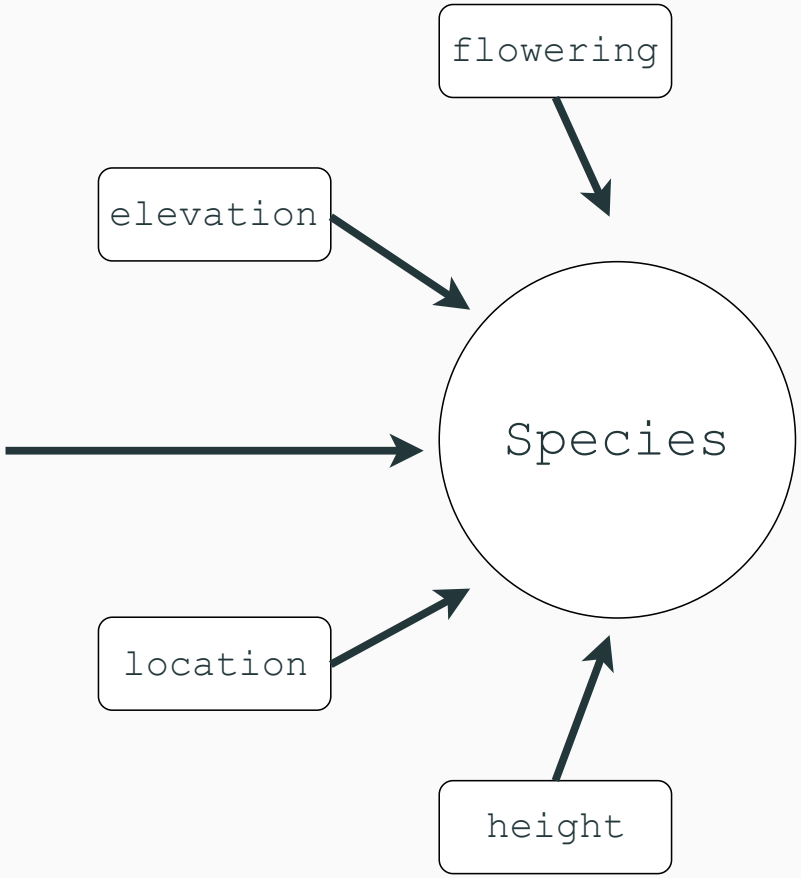


Figure 2: iNaturalist observation of *Protea nana*

Distribution of Protea cynaroides in the Western Cape



Problem Statement



Current Approach

$$P(\text{protea}_i|\text{loc}, \text{ele}, \text{image}, \ldots)$$

Current setup

- Naive Bayes
- 20% accuracy which jumps to 80% when considering top-5
- Two CNNs built on Inception
- First CNN classifies 8 most observed species (72% accuracy)
- Second CNN looks at the rest

Difficulties

- Small dataset with large tail
- 3,500 observations with 2,400 flower head photos
- 50% of data in 7 species
- Intraspecies variation often larger than interspecies variation
- Large dataset bias for common species

Future Work

Ideas

- Incorporate visual aspect
- Consider dependencies between attributes
- Incorporate more attributes
- Generative approach to image classification (e.g. VAEs), linking with the attributes in a PGM

Figure 3: *Protea rupicola* high up on the Kammanassie