Extracting Predictive Models from Marked-Up Free-Text Documents at The Royal Botanic Gardens, Kew, London

Allan Tucker[1] and Don Kirkup[2]
allan.tucker@brunel.ac.uk

1 Department of Computer Science, Brunel University, UK
2 Royal Botanical Gardens at Kew, UK

Abstract. In this paper we explore the combination of text-mining, un-supervised and supervised learning to extract predictive models from a corpus of digitised historical floras. These documents deal with the nomenclature, geographical distribution, ecology and comparative morphology of the species of a region. Here we exploit the fact that portions of text in the floras are marked up as different types of trait and habitat. We infer models from these different texts that can predict different habitat-types based upon the traits of plant species. We also integrate plant taxonomy data in order to assist in the validation of our models. We have shown that by clustering text describing the habitat of different floras we can identify a number of important and distinct habitats that are associated with particular families of species along with statistical significance scores. We have also shown that by using these discovered habitat-types as labels for supervised learning we can predict them based upon a subset of traits, identified using wrapper feature selection.

1 Introduction

In the last two decades, there has been a surge in data related to biodiversity of plants through, for example, on-line publications, DNA-sequences, images and metadata of specimens. Much of the new data is characterised by its semi-structured, temporal, spatial and 'noisy' nature arising from disparate sources. Here, we focus on the use of textual data in floras. These are the traditional taxonomic research outputs from organisations such as the Royal Botanical Gardens at Kew, London, and deal with the nomenclature, geographical distribution, ecology and comparative morphology of the species of a region, explicitly linked to defined taxonomic concepts. We exploit the use of data mining (and in particular text mining) in combination with machine learning classifiers in order to build predictive models of habitat based upon plant traits.

Text mining has grown in popularity with the digitisation of historical texts and publication [12]. In particular, the use of text mining for bioinformatics data has led to a number of different approaches. For example, medline abstracts have been mined for association between genes, proteins and disease outcome

[11]. These can vary from simple statistical approaches to more complex *concept profiles* as developed in [8] where a measure of association between a pair of genes is calculated based not only on the co-occurrence of entities in the same document, but also on indirect relations, where genes are linked via a number of documents. An association matrix for gene-pairs can be generated, where each entry represents the strength of the relationship between genes, based on a database of scientific literature. Business Intelligence is another area where text mining has proved popular in relation to tweet messages and sentiment analysis [5]. In ecology, the use of text mining is a little less explored though there is a growing interest in the use of these approaches to extract knowledge [14].

There is a growing effort to taking a predictive approach to ecology [3] with the availability of larger and more diverse datasets. If we can build models that can predict biodiversity or species distribution, for example, then we will have greater confidence that the models capture important underlying characteristics. A related discipline, 'systems ecology', encourages a focus on holistic models of ecosystems [10]. This follows the success of similar approaches in molecular biological applications Many novel techniques developed from bioinformatics can be translated to the ecological domain [15]. Indeed, here we make use of a statistic that was previously developed for validating clusters in microarray data.

In this paper, we explore the use of text-mining where we exploit the fact that the flora that we analyse are marked-up to distinguish between descriptions of different plant traits and habitats. We cluster habitat texts and use a statistic (originally designed to validate clusters of genes from microarray experiments) to validate the discovered habitat-types against the plant taxonomy. We then exploit the trait texts to build probabilistic classifiers [6] for predicting the habitat clusters. The motivation for this research is to permit exploration of taxonomic and functional trait diversity. This will lead to better plant functional type classifications for input to vegetation models under differing climate change scenarios. From this, we can gain a better understanding of plant species distribution, vital for effective species and habitat conservation. In the next section we describe the general pipeline that we have developed to build these predictive models from the marked up text and plant taxonomy. We also describe the probabilistic models and statistics that we use to assess our results. In the results section we document the results from the different stages of the pipeline with insights from plant ecology before concluding.

2 Methods

2.1 Data

The Flora of Tropical East Africa (FTEA) is one of the largest regional tropical Floras ever completed, covering 12,500 wild plant species from Uganda, Kenya and Tanzania. Together with Flora Zambesiaca and Floras Somalia, these floras cover equatorial, tropical and subtropical biomes of [16] and major phytochoria of [17]. Virtually all the main vegetation types are represented. These floras have been digitised to create the EFLORAS database - a unique data source

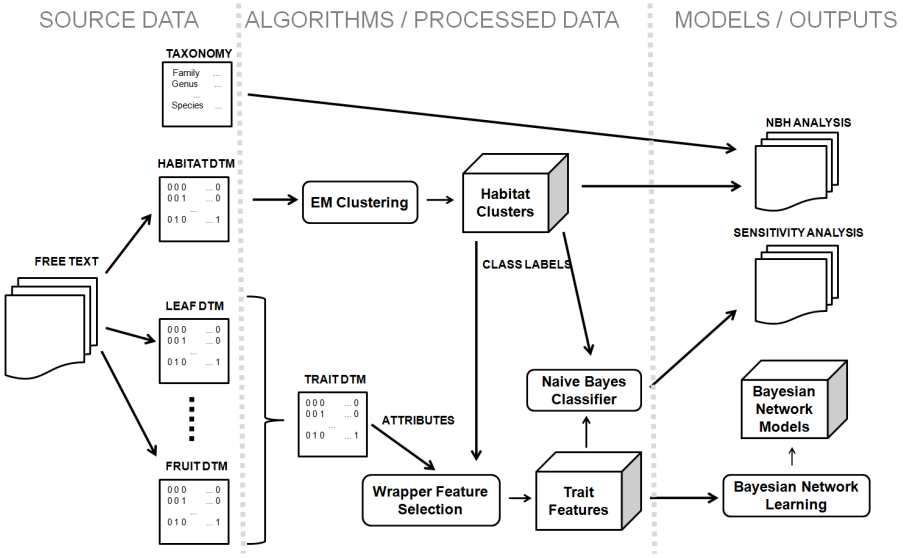


Fig. 1. Pipeline for Converting Free Marked-Up Flora Text into Predictive Models of Ecosystems

of tropical plant species distribution, ecology and morphology, together with historical data on plant collectors [9] in the EFLORAS corpus. Each document represents a *taxon* (in this case a species) which is identified by a unique ID and contains a digitised paragraph, tagged as to whether it describes a number of different characteristics: *habitat, habit, leaf, fruit, or seed* (flower features were not available for this study). In total there are 8252 documents (i.e. species), containing each paragraph. Standard text mining procedures were employed for each paragraph type in order to remove stop-words, white spaces, punctuation and numbers, and to stem all necessary words [4]. This results in an n (terms) by m (documents) matrix for each paragraph, where cells contain the number of times a term has appeared in the corresponding document. Terms can include anything such as 'bilobate', 'golden' and 'elongated' reflecting traits but also other terms such as 'beautifully' and 'actually' reflecting a particular author's writing style.

2.2 Experiments

We exploit the text concerning plant traits to predict habitat. Therefore, the matrices for all types of trait are combined into a single document term matrix, the *trait matrix*, for all types except habitat. Clearly, the combined trait matrix and the *habitat matrix* are sparse and any terms that appear in less than 10% of all documents are removed from both. This reduces the size of the trait matrix to 759 terms, and the habitat matrix to 106. We use the *tm* package in *R* for all of this processing [4]. Having processed the text into two matrices: one that

represents the plant traits and the other that represents the habitat characteristics, we exploit probabilistic clustering to the habitat data in order to identify different types of habitat. We found that a simple Expectation Maximisation approach to clustering [1] identified meaningful clusters without the need to supply the number of clusters. We exploit plant taxonomy information to validate these clusters. This contains details of the plant family, genus, and species for each flora document. We make use of a statistic previously developed for assessing clustering in microarray data against known gene functional information [13]. This *NBH* statistic is used to score the significance of each plant family being associated with a particular habitat based upon the number of times a plant family is associated with it, and the number of times the family is associated with others. This probability score is based on the hypothesis that, if a given habitat, i, of size s_i, contains x documents from a defined family of size k_j, then the chance of this occurring randomly follows a binomial distribution and is defined by:

$$pr(\text{observing } x \text{ docs from family } j) = (\tfrac{k_j}{x})p^x q^{k_j - x}$$

$$\begin{gathered}\text{where } p = s_i/n, \\ q = 1 - p\end{gathered}$$

As in [13] we use the normal approximation to the binomial to calculate the probability where:

$$\begin{gathered}z = (x - \mu)/\sigma, \\ \mu = k_j p, \\ \sigma = k_j pq\end{gathered}$$

This cluster probability score is used to identify statistically significant families allocated to each habitat (at the 1% level).

The cluster labels identified through the clustering are then used to identify predictive features in the trait matrix using a wrapper feature selection approach [7] to explore combinations of predictive terms. We use the Naïve Bayes Classifier [6] as the classifier for the wrapper as this was found to be the most predictive. Whilst we expect there to be interesting interactions between terms, it appears that the simplicity of the Naïve Bayes is suitable to classifying a large number of habitats by minimising parameters. What is more, the flexibility of Bayesian classifiers allow us to use different nodes as predictors so we can use the resultant models to predict both neighbouring plant traits as well as habitat type.

The Naïve Bayes classifier makes the simplifying assumption that each feature is independent of each other given the class. This corresponds to the efficient factorization

$$p(x|c) = \textstyle\prod_{i=1}^{n} p(x_i|c)$$

Assuming uniform priors, a Bayesian estimate of $p(x_i|c)$ is given by

$$\hat{p}(x_{ip}|c) = \tfrac{1+n(x_{ip}|c)}{s+n(c)}$$

where s is the number of discretized states of the gene variable X_i, $n(x_{ip}|c)$ is the number of cases in the dataset where X_i takes on its pth unique state within the samples from class c, and $n(c) = \sum_{p=1}^{s} n(x_{ip}|c)$ is the total number of samples from class c. From $\hat{p}(x|c)$, an estimate of $p(c|x)$ is calculated using Bayes rule and the resulting classification rule assigns the sample x to the class associated to the highest estimated probability.

Having identified the relevant features to predict habitat type, we explore how predictive these features are using a Naïve Bayes classifier under a 10-fold cross-validation regime. Finally, we explore the interactions between features within each habitat type by carrying out 'what if' experiments on a sample of habitats and build Bayesian network structures from data associated with each habitat type to see if any traits / network-of-traits are highlighted for that habitat in particular. The general pipeline is illustrated in Figure 1.

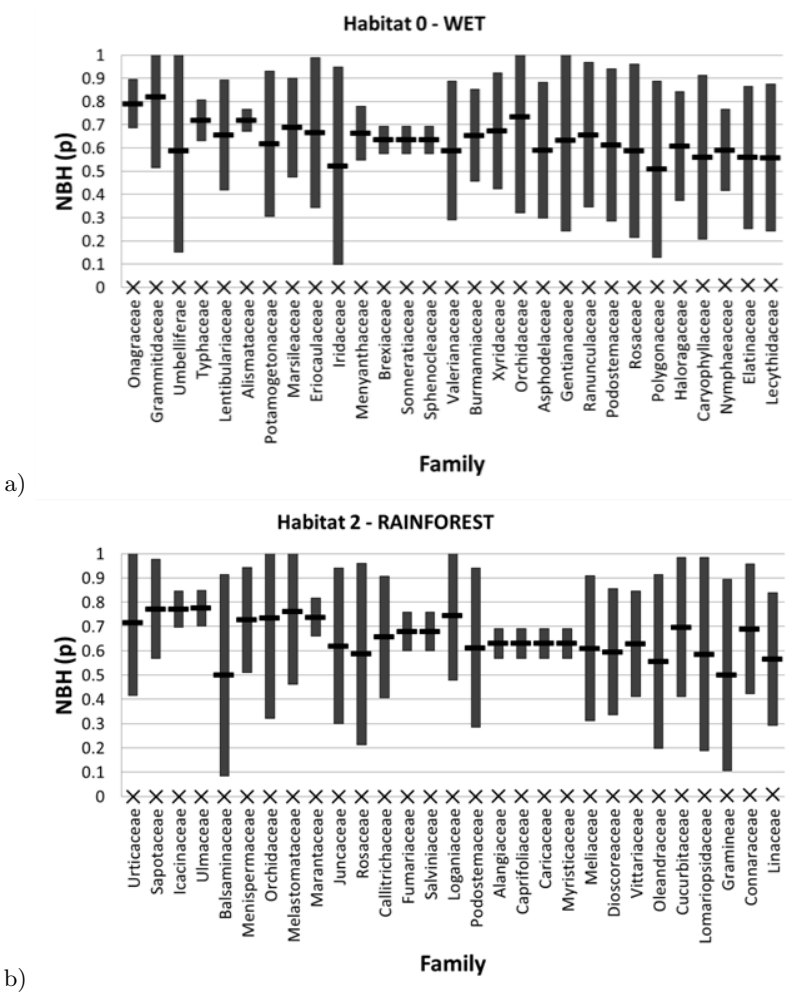
3 Results

3.1 Discovering Habitat Clusters

Having clustered the data into 9 different habitats based upon the document term matrices generated from the habitat corpus, the individual term frequencies were calculated and explored in the context of habitats that they likely represent. The following descriptions could be elicited from experts based upon the terms associated with each habitat cluster.**Habitat 0** - appears to reflect vegetation in wet places that are largely, but not exclusively, upland (For the remainder of the paper we refer to this habitat type as WETLANDS - *WET* when abbreviated). **Habitat 1** reflects a mixture of woody and herbaceous vegetation in drier conditions, including deciduous types (DECIDUOUS BUSHLAND - *BUSHLAND*). **Habitat 2** clearly reflects lowland and upland, wetter forest types (RAINFOREST). **Habitat 3** contains a variety of upland vegetation (MONTANE). **Habitat 4** appears to represent disturbed vegetation and cultivation (DISTURBED) **Habitat 5** contains vegetation in open sites, and margins including cultivation, similar to 4 and not readily separable (OPEN/DISTURBED - *OPEN*). **Habitat 6** is large and contains a combination of open woody and herbaceous vegetation in wetter areas, including evergreen types (WOODLAND + WOODED GRASSLAND - *WOODED*). **Habitat 7** is a mixture of drier lowland forest, scrub and evergreen bush (FOREST + SCRUB + BUSH - *SCRUB*). **Habitat 8** contains mixed habitats including rainforest and dry vegetation (FOREST + SCRUB + BUSH - *SCRUB2*).

The distribution of documents to habitat varied dramatically. In general, the three larger clusters (habitats 1, 6 and 8) were less specific and generally mixed different habitat-types. For this reason, these were omitted from the feature selection and classification analysis though further work will involve exploring finer grain clusters to split these into more detail. The identified habitats were validated by using the *NBH* statistic [13]. The distribution of all plant families occurring in the texts over each specific habitat were explored. Families with an

NBH statistic with p values at less than the 1% level were selected and compared to the distribution over the other habitats in order to highlight the specific association between that family and the discovered habitat (shown in Figure 2). These results highlighted some expected families of plants based upon their



a)

b)

Fig. 2. Significant Plant Families for 2 Selected Habitats Using the NBH Statistic (denoted with an '×') Compared to the Distribution over all Other Habitats (Denoted by Error Bars)

habitat types: Habitat 0 (WET) is clearly dominated by families of aquatic and marshplants (figure 2a). Habitat 1 (BUSHLAND) contains Burseraceae, Leguminosae, Capparidaceae which are dominant dry bushland components. Portu-

lacaceae is also characteristic of this type of habitat. Habitat 2 (RAINFOREST) contains herbs and understory shrub/treelet families are well represented (including ferns), followed by tree families (figure 2b). Habitat 3 (MONTANE) contains montane ferns. Habitat 4 (DISTURBED) are mostly families with weedy species which make sense for disturbed regions (figure 2c). Habitat 5 (OPEN) includes a mixture of herbaceous and woody families. Habitat 6 (WOODLAND) families are a mixture and this is not surprsing considering the large mixed habitats that were identified earlier. It is intriguing as to why the mistletoe families are so prominent (Loranthaceae, Viscaceae, Santalaceae). Habitat 7 (SCRUB) families are scrub component families and Habitat 8 (SCRUB2) fits with forest herbs shrubs and trees. For all identified families the p-value compared to the distributions of other families and habitats illustrate that they are well separated and significant.

3.2 Plant Trait Feature Selection and Classification

We now turn to the plant trait documents. We wish to use these to predict habitat type. A wrapper feature selection procedure was carried out on the plant traits to identify combinations of traits that characterise the different habitat clusters. A greedy search scored with classification accuracy was used to identify the features. Figure 3 illustrates the identified features and how the expected frequencies of these terms vary for each habitat type. For example, the term *FRUIT_exsert* representing the term 'exsert' in the text describing 'fruit' is identified as relevant and, as can be seen here, has a much higher expected frequency in scrub habitats compared to others. Features marked with an asterisk '*' were those that were expected to be good at discriminating between the habitats.

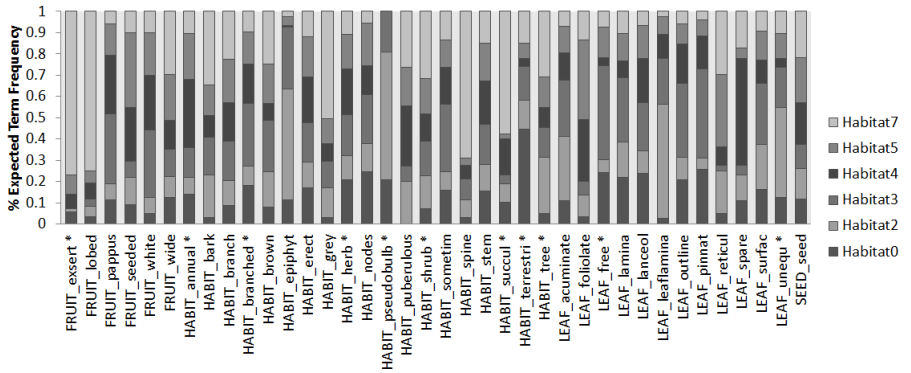


Fig. 3. Identified Features (using Naïve Bayes Wrapper) - Expected Frequencies for each Habitat

-	H0	H2	H3	H4	H5	H7	AUC
Habitat0:	0.06	0.05	0.01	0.03	0.01	0.01	0.70
Habitat2:	0.05	0.15	0.01	0.03	0.01	0.03	0.71
Habitat3:	0.02	0.02	0.02	0.02	0.00	0.01	0.69
Habitat4:	0.03	0.03	0.01	0.08	0.02	0.01	0.70
Habitat5:	0.02	0.03	0.01	0.05	0.02	0.02	0.64
Habitat7:	0.02	0.03	0.00	0.03	0.01	0.07	0.75
Wtd Avg.:	-	-	-	-	-	-	0.70

Table 1. Classification Results (Confusion Matrix and AUC) of Habitats Given Features as Percentages (10 fold Naïve Bayes Classifier)

The results of applying 10-fold cross-validation to predicting the habitat type with Naïve Bayes is shown in Table 1. The predictive accuracy varied depending on the habitat with Habitat 7 (SCRUB) being the most accurately predicted. The table shows the distribution of Areas Under the ROC curves for each habitat. The confusion matrix indicated the typical misclassifications involved mistakenly classifying Habitats 0 (WET) and 2 (RAINFOREST), and 4 (DISTURBED) and 5 (OPEN) which makes sense as vegetation could easily overlap between these.

3.3 'What if?' Experiments

Having identified both habitat type and plant traits relevant to predicting habitat type, we explore the interaction discovered between the different features. This allows us to explore combinations of terms as well as their relationship to different habitats. Figure 4 illustrates the expected frequencies as inferred from

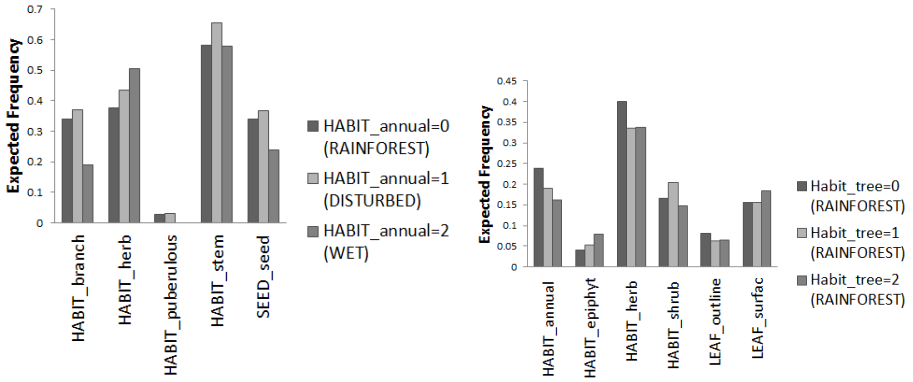


Fig. 4. 'What if' Experiments Illustrating Distributions of Key Traits Using Different Observations on *HABITAT_annual* (left) and *HABITAT_tree* (right)

the predictive model for some selected plant traits. The three bars in Figure 4a

represent expected frequencies for the other traits when HABIT_annual is set to 0, 1 and 2 respectively. Terms in brackets illustrate the most probable habitat given the observation.

For the scenario in Figure 4a where *HABIT_annual*=0 is observed, the most likely habitat is RAINFOREST, and the highest expected values are for *HABIT_herb*, *HABIT_stem* and *SEED_seed*. This is somewhat counterintuitive as the features *herb, stem* and *seed* are associated with annuals, but here annual has a frequency of zero. For *HABIT_annual*=1, the most likely habitat is DISTURBED, with highest conditional expected frequencies for *HABIT_stem, HABIT_herb* and *HABIT_branch*. For *HABIT_annual*=2, the most likely habitat WET, with highest conditional expected frequencies for *HABIT_stem, HABIT_herb* and *HABIT_branch*, while *HABIT_puberulous* is 0. This scenario makes sense: As the frequency of *HABIT_annual* is increased, so too are the probabilities of observing the 'annual related' features (stem, herb, branch etc.). There are comparatively few annuals in rainforests but as expected they are a major element of disturbed habitats. In addition, very high numbers of annuals appear to be associated with the wet habitat and these plants apparently are never puberulous (shortly hairy). Aquatic plants are frequently glabrous, that is, without hairs.

Unlike *HABIT_annual*, if we observe *HABIT_tree* as either 0,1 or 2 (see Figure 4b), the most likely habitat is always RAINFOREST. This could be because the habitat is more species diverse. However, the intermediate scenario *HABIT_tree*=1 gives the highest probability for habitat 2. This is because RAINFOREST is a rich habitat which contains both tree and non-tree species. It could be that *HABIT_tree*=2 precludes non-tree species typical of RAINFOREST.

3.4 Networks of Traits

For the final piece of analysis, we explored learning network structures for different habitat-types by splitting the data accordingly and learning networks using the K2 algorithm of [2]. Some sample networks are documented in Figure 5 (detail). Some interesting characteristics emerge when focussing on the 'hub' nodes - those that have higher degree of connectivity. For example, in Habitat 0 (WETLANDS) there are two clear hubs: *HABIT_shrub* and *LEAF_free*. The former links to features of woody plants (expected to be mostly absent from typical habitat 0 plants). The latter contains many aspects of leaf descriptions (lamina, outline, lanceolate, pinnate, surface) but there are some connections which are not immediately clear (connections from *HABIT_terrestrial, FRUIT_wide* and *FRUIT_pappus*). The term *LEAF_free* may cover several different situations eg. free stipules, free petiole (all parts of the leaf). In Habitat 2 (RAINFOREST) there is a *HABIT_epiphyte* hub (containing terms *epiphyte* and *pseudobulb*) which could be linked to orchids (an epiphyte is a plant that grows on trees such as orchids). Also *epiphyte* and *tree* are linked which could be related to plants specifically growing on trees. In general, many of the hubs make sense in terms of why they may be connected (often descriptive terms that are related to similar parts of a plant). There are also some interesting relationships that appear to be specific to their habitat such as orchids in rainforests.

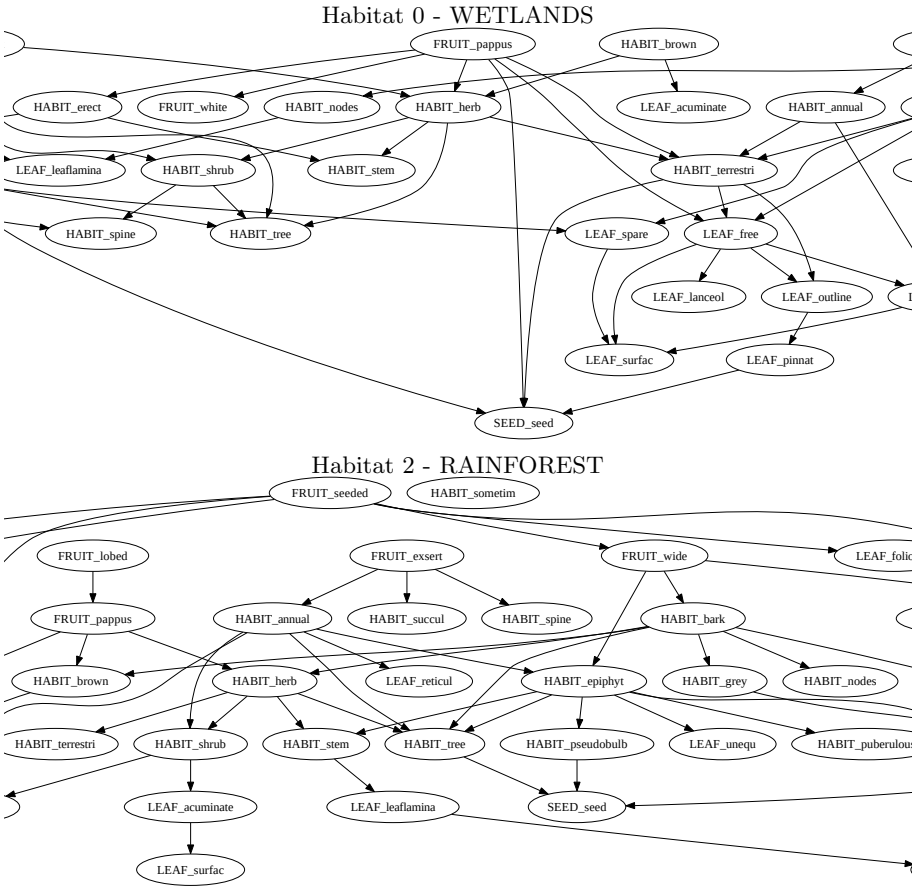


Fig. 5. Portion of Networks Learnt for 2 Sample Habitats with a Focus on Hub Nodes

4 Conclusions

In this paper we have explored a pipeline for converting text documents at the Royal Botanical gardens at Kew, London describing different plant families into models that can predict habitat type and neighbouring plant characteristics, based upon plant traits. The pipeline identifies distinct habitat types and integrates taxonomy data in order to highlight significant plant families within those habitats by exploiting a statistic previously developed for bioinformatics applications. A combination of wrapper feature selection and naive Bayes classification is exploited to identify the discriminative features and build models that can predict both neighbouring plant traits and habitat type. Future work, will involve exploring other predictive capabilities between the text and other data such as the taxonomy. For example, we will explore how well our models

can predict families and species directly rather than via the habitat. We will also explore other ways to quantify the value of the 'what if' results.

The paper documents the start of a larger project that explores the hypothesis that a comprehensive understanding of neighbouring species and what a plant looks like will indicate where it grows. Our tools will enable predictions about individual species and their functions in ecosystems of other regions. This will be facilitated through identifying factors (including taxonomic and environmental) that influence biodiversity and stability of ecosystems, vital for effective species and habitat conservation.

References

1. J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report TR-97-021, ICSI*, 1997.
2. G.F. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, (9):309–347, 1992.
3. M.R. Evans, K.J. Norris, and T.G. Benton. Introduction: Predictive ecology: systems approaches. *Philosophical Transactions of the Royal Society: Part B*, 367(1586):163–169, 2012.
4. Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, March 2008.
5. Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.
6. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, (29):131–163, 1997.
7. I. Inza, P. Larrañaga, R. Blanco, and A.J. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine*, (31):91–103, 2004.
8. R. Jelier, Martijn J. Schuemie, Antoine Veldhoven, Lambert C.J. Dorssers, Guido Jenster, and Jan A. Kors. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biology*, 9(6):R96, 2008.
9. D. Kirkup, P. Malcolm, G. Christian, and A. Paton. Towards a digital african flora. *Taxon*, 54(2), 2005.
10. Drew Purves, Jorn Scharlemann, Mike Harfoot, Tim Newbold, Derek P. Tittensor, Jon Hutton, and Stephen Emmott. Ecosystems: Time to model all life on earth. *Nature*, (493):295–297, 2013.
11. E. Steele, A. Tucker, and M.J. Schuemie. Literature-based priors for gene regulatory networks. *Bioinformatics*, 25(14):1768–74, 2009.
12. D.R. Swanson. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc.*, (78):29–37, 1990.
13. S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam. Consensus clustering and functional interpre-tation of gene-expression data. *Genome Biology*, 5(11):R94, 2004.
14. Javier Tamames and Victor de Lorenzo. Envmine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, 11(294. doi:10.1186/1471-2105-11-294), 2010.

15. A. Tucker and D. Duplisea. Bioinformatics tools in predictive ecology: Applications to fisheries. *Philosophical Transactions of the Royal Society: Part B*, 356(1586):279–290, 2012.
16. H. Walter. In *Vegetation of the Earth and Ecological Systems of the Geo-biosphere*. Springer-Verlag, 1979.
17. F. White. In *The Vegetation of Africa A descriptive memoir to accompany the Unesco/AETFAT/UNSO vegetation map of Africa*. UNESCO, 1983.