**TABLE OF CONTENTS**

1. **Introduction & Literature survey**
   * **Remote sensing for invasive species detection**
   * **Existing datasets (satellite imagery)**
   * **Chromolaena odorata in Nepal**
   * **Climate change effects on biodiversity**
2. **Problem formulation and initial results**
   * **Unsupervised classification & label propagation**
   * **Results from Landsat + Field survey results**
3. **Future work ideas**
   * **Temporal dimension and projection (forecasts)**
   * **Incorporating other data (temperature, human activity)**
4. **Possible directions using machine learning techniques**
   * **Cycle consistent GANs/ convolutional LSTMs**
   * **Transfer Learning**
   * **Theory guided ML (model of diffusion, species competition)**
   * **Abduction loop (for informing field surveys)**

# Introduction

# Literature Survey

# Problem formulation and initial results

**Methodology overview:** For our initial we adopted an unsupervised learning approach with label propagation and tested cluster memberships of known reference locations. Before describing the initial results, we begin with a high level description of the clustering problem.

Given a collection of points and number of clusters , a clustering algorithm produces a partition where are disjoint non-empty subsets that cover set . (. K-means clustering is a well-known clustering algorithm that begins with randomly initialized cluster centers and iteratively recomputes cluster centers until convergence, or max iterations is reached. In order to recompute cluster centers in each iteration, each point in the collection is assigned to its closest cluster center according to some distance function . The algorithm can be written down as follows:

* Initialize cluster centers by randomly choosing any K examples.
* Assign each to its closest cluster center to produce the partition.
* Recompute cluster centers based on

In addition to the given collection of points , let be a set of observation points for which the ground truth is known. In a binary classification problem, where each input point is to be assigned one of two labels (‘true’ or ‘false’), let be the set of points whose label is known to be ‘true’ (positive training examples). The idea of label propagation with unsupervised learning, is to run the clustering algorithm and identify the cluster(s) which contain the points in , and assign the label ‘true’ to all other points in those cluster(s). Such a method can be evaluated through cross-validation by separating into different sets of and . The robustness of this approach can also be tested with respect to the number of clusters .

**Initial results:** Landsat 8 satellite imagery data was obtained from USGS EarthExplorer (<https://earthexplorer.usgs.gov>). These satellites carry the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) consisting of 9 spectral bands with a spatial resolution of 30 meters for Bands 1 to 7 and 9 (Barsi et al. 2014). For our initial study, we collected images spanning 05/08/2018 – 05/17/2018. Quantum GIS (<https://www.qgis.org>) was used to create an input map from bands 2-7 of the satellite imagery data. These were respectively Blue, Green, Red, Near Infrared (NIR), Shortwave Infrared (SWIR) 1 and SWIR 2.

In our initial study Two field visits were conducted by the team in Nepal. The first one (June 5-7, 2018) was conducted in Chitwan District for the purpose of identifying an initial field site. Four different sites (Barandabhar, Rampur, Ganganagar, Ghaila Ghari) were visited for different species.

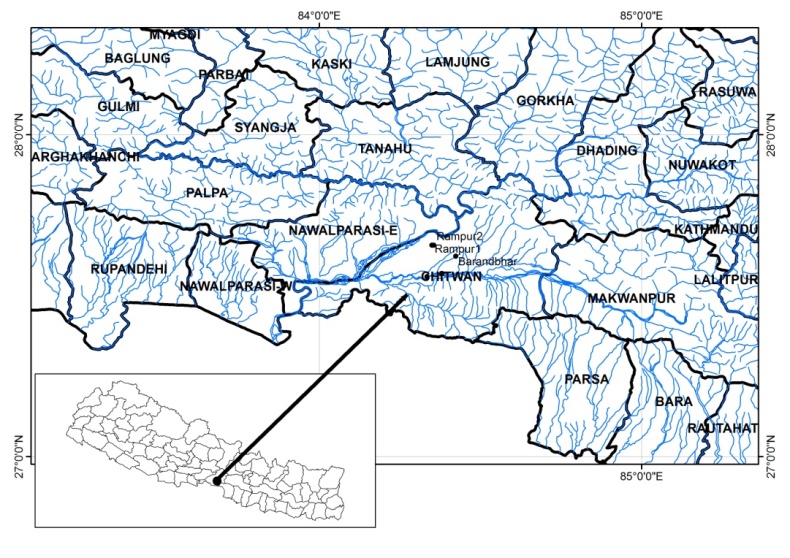
 

Fig X: Field Survey Sites in Chitwan District

Specifically, Chromolaena odorata was identified as the dominant species in three different locations as listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| Location | Longitude | Latitude | Elevation (masl) |
| Barandbhar | 84.42294 | 27.62182 | 215 |
| Rampur1 | 84.34933 | 27.65634 | 164 |
| Rampur2 | 84.3549 | 27.65648 | 167 |

From the satellite imagery pertaining to the Chitwan region, we constructed clusters with semi-automatic classification with chosen bands (2-7 from LandSat 8).



Rampur2

Rampur1

Barandbhar

Fig X: Result from Barandbhar training set



Rampur2

Rampur1

Barandbhar

Fig X: Result from Rampur training set

We used the Semi-automatic classification plugin (SCP) in GIS to provide an input set of coordinates and observe the resulting clusters. The above figures show the results of SCP with Rampur and Barandbhar coordinates respectively given as ground truth. We note that for each training set, the test reference points fall in the same cluster, validating the identified clusters. The points colored yellow and orange respectively in Fig X and X, are the initial guesses of locations where C. odorata may be found. Given the sparsity of reference points, these mappings carry a high level of uncertainty. Also note that between Barandbhar training set (1 reference point) and Rampur training set (2 reference points), the number of points assigned to the labeled cluster diminish, hinting that the identified clusters may be more general rather than species specific. In the following section, we describe some machine learning techniques that may be helpful, and sketch out a few potential lines of research investigation.

# Future work ideas

# Possible directions using machine learning techniques

## Artificial Neural Networks

## Transfer Learning

Machine learning (ML) algorithms typically perform well when the quantity and quality of training samples is high. However, coming up with data points is hard for the problem at hand; identifying sites infested by the target species is resource intensive. Also, seasonality, cloud cover and other factors can add to the challenges by restricting the number of usable satellite. A natural question to ask is if we can leverage a learner meant for a different but related problem. This learner might have been used to identify the presence of some other plant or to identify under a different class of inputs. Under these circumstances transfer learning (TL) is a design methodology in ML that enables one to transfer the knowledge learnt in one task (referred to as source task) to be transferred to other tasks (target tasks) (Pan, Yang, & others, 2010).

TL capitalizes on the fact that ML algorithms typically take a layered approach for learning (particularly deep learners). For images, it is observed that the initial layers extract low level features such as edges, changes in color, or gradient in temperature, while the middle layers might perform complex tasks of finding the correlation among the feature from different modalities with the higher layers performing task-specific functions such as defining classification rules or determining parameters of the prediction model. Although the distribution of the data might differ, the feature extraction process and the data modelling aspect of the layers would generally remain the same for the source and target. In cases where the target lacks labels but the tasks remain the same as the source, one could learn a mapping between the target and the source which then enables one to employ the labels available in the source domain. TL is also helpful when we consider tasks with insufficient training data. During the training phase, one could employ pre-trained feature extraction layers and augment the correlation capturing capability learned in the source domain to the available training data and then fine tune the task-specific learning functionality. The use of pre-trained layers allows one to invest precious training data for learning higher level functionalities. Utilizing the feature extraction functionality implies that all the training data can be utilized only for task-specific layers.

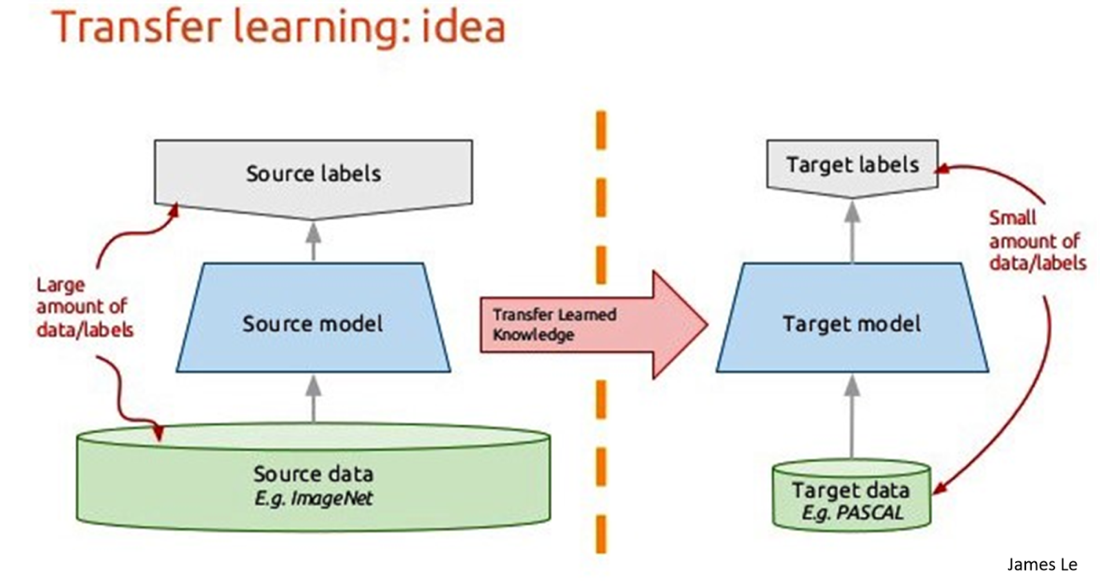


Figure 1. A schematic of transfer learning. Courtesy: William Vorhies (https://www.datasciencecentral.com).

## Combining data-driven models with theory-based models

Increasing availability of high-quality data and high-performance computing methods, have fueled the development of techniques such as deep learning. However, largely, these techniques are agnostic to underlying theory or understanding of the phenomenon under study. Mechanistic models, which are developed on the basis of fundamental of years of experience are not taken into account. It is natural that researches have eventually realized that the potential of combining these two paradigms of computational models for understanding complex phenomena. In this regard, two bodies of work have emerged in recent years, though the general concept has been around for a long: Theory-guided machine learning and the abduction loop. The integrated system has the following advantages: (i) Greater **consistency** as it respects known or accepted physical relationships between variables; (ii) **generalizability** allowing the user to make educated guesses under scarcity of data; and (iii) ability to evaluate **counterfactual** scenarios. For our study, potentially, these concepts can help create robust models under data inadequacy, guide data exploration, avoid overfitting, and help evaluate various intervention strategies.

### Theory-guided machine learning

Models based on scientific knowledge represent relationships between variables, for example, temperature and diapause. These are empirically proven, deduced from first principles or widely accepted by domain experts. Unlike these “theory-based” models, data-driven models use a set of training examples to learn a model that can automatically extract relationships between the input and output variables. Combining theory-based models with current data-driven models enables overcome many shortcomings of either methods. This has been demonstrated in a number of works concerning physical processes in domains such as hydrology and computational chemistry, and has been summarized in a survey (Karpatne et al., 2017). The survey describes the advantages and disadvantages of both modeling paradigms. More importantly, it clearly describes methods (with use cases) to fuse theory-based models with data-driven methods.

### Abduction loop (for informing field surveys)

Abduction is an inference approach that uses data and observations to identify plausible (preferably, best) explanations for phenomena (Peirce, Hartshorne, Weiss, & Burks, 1960). One-step abduction first generates data through experiments or observations. Then, data analysis consists of searching for patterns and generalizing these into phenomena, which is an inductive step. These results are used to formulate hypotheses based on theories whose purpose is to explain the data. Hypotheses may exist (from a previous loop) or may be proposed in this step, and can be removed (e.g., via falsification). Multiple candidate theories may be posed for a given phenomenon. Multi-agent models and mechanistic models of plant growth and response to ecological factors are candidate models at this stage. The best explanation, or hypothesis/theory appraisal, is the process of identifying the best explanation for the phenomenon; this includes hypothesis falsification. This stage can potentially identify gaps in data and guide surveys. Finally, the last step in an iteration is to determine what to do next, in terms of designing new experiments. The iterative process may terminate for any number of reasons; e.g., a best  
explanation has been found.

**References**

1. Modelling the distribution of the invasive shrub species Chromolaena odorata in Nepal and its impact on the native flora: <https://www.nwo.nl/en/research-and-results/research-projects/i/30/20030.html>
2. C.M. Joshi(2006): Mapping cryptic invaders and invasibility of tropical forest ecosystems: Chromolaena odorata in Nepal. , Enschede May 22, 2006
3. C.M. Joshi, J. de Leeuw, I.C. van Duren(2004): Remote sensing and gis applications for mapping and spatial modelling of invasive species *International Archives of the Photogrammetry, Remote Sensing and Spatial Informational Sciences Conference Istanbul 2004* pp. 669 - 677
4. C.M. Joshi, J. de Leeuw, A.K. Skidmore, I.C. van Duren, H. Oosten(2006): Remotely sensed estimation of forest canopy density: a comparison of the performance of four methods *International Journal of Geoinformatics and Earth Observation* pp. 84 - 95 ISSN: 0303-2434.
5. Barsi, J.A.; Lee, K.; Kvaran, G.; Markham, B.L.; Pedelty, J.A. The Spectral Response of the Landsat-8 Operational Land Imager. *Remote Sens.***2014**, *6*, 10232-10251. [doi:10.3390/rs61010232](http://www.mdpi.com/2072-4292/6/10/10232)
6. Congedo Luca (2016). Semi-Automatic Classiﬁcation Plugin Documentation. DOI: <http://dx.doi.org/10.13140/RG.2.2.29474.02242/1>
7. Chromolaena odorata observations on iNaturalist: <https://www.inaturalist.org/taxa/199400-Chromolaena-odorata>
8. Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., … Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. <https://doi.org/10.1109/TKDE.2017.2720168>
9. Pan, S. J., Yang, Q., & others. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.
10. Peirce, C. S., Hartshorne, C., Weiss, P., & Burks, A. W. (1960). *Collected Papers of Charles Sanders Peirce: Edited by Charles Hartshorne and Paul Weiss: Principles of philosophy and Elements of logic* (Vol. 1). Harvard University Press.
11. Agnihotri P, Husain T, Shirke PA, Sidhu OP, Singh H, Dixit V, Khuroo AA, Amla DV, Nautiyal CS. Climate change-driven shifts in elevation and ecophysiological traits of Himalayan plants during the past century. Current Science (00113891). 2017 Feb 10;112(3).
12. Bhattacharjee A, Anadón J, Lohman D, Doleck T, Lakhankar T, Shrestha B, Thapa P, Devkota D, Tiwari S, Jha A, Siwakoti M. The Impact of Climate Change on Biodiversity in Nepal: Current Knowledge, Lacunae, and Opportunities. Climate. 2017 Dec;5(4):80.