Aim: to include the presence of pests in weather based disease prediction models for cotton and rice crops.

Challenges or limitations to implement this:

there are a variety of pests in the Indian context whose distributions aren’t the same throughout the country, therefore a pest that is acting as a vector for a particular disease in orissa may not be present at all in tamil nadu, even if the disease is occurring in both places. What this means is that each pest plays the role of a vector in differing magnitudes at different locations. Therefore a single all encompassing model for the entire country would require data pertaining to all kinds of pests in all the regions.

Try to re-explain:

Lets say we have data for disease X, and pests a,b,c in MH, and pests a,d,e in AP. If

Challenges with the data (Data pre-prcoessing or cleaning or parsing or filtering or selection steps ):

1. There were severe variations in the list of pests measured spacially as well as temporally. What this means is that its possible that pests a,b,c were recorded in MH while a,d,e were recorded in AP. But how does this matter? This meant that we couldn’t leverage the entire dataset to build an all encompassing model that would function for the entire country, because, that would skew the model towards declaring A as the only factor that drives disease X, since there will be missing data for all other pests. This necessitated us to build location specific models for each disease.
2. And even within the same location, all the pests weren’t measured across all the years where data was recorded, so all the years of data couldn’t be utilized as including these years would decrease the prominence of certain pests which don’t have measurements in certain years. So it was necessary to curate the data to a certain number of years where most pests in that region are recorded.

For this reason, to conduct statistical modeling using pests, it is necessary that we have data related to a particular set of pests in a certain region across many years.

In this regard, the database was parsed for identifying those subsets that fit the bill.

However, a downside to developing models with such complexity, is that the real life implementation of such models will not be possible unless farmers are ready to invest the money to accurately measure pest numbers in their farms. There is also the case where models prepared using data pertaining to a certain set of pests will not be accommodative of other pest species which may turn up all of a sudden. A third downside, with pest based disease models is it should be capable of forewarning the occurrence of disease without waiting until the pest numbers shoot up heavily by which time the disease would have propagated considerably.

So, the aim was modified to develop a ML model that incorporated the presence of pests but is also easily implementable, flexible enough to accommodate other pests and at the same time capable enough of predicting a sharp rise in disease occurrence two weeks in advance.

we need a dataset that has recordings of different types of pests,

one, each pest plays the role of a vector in differing magnitudes,

Methodology:

1. Description of Framework : Categorical input from farmers for a set number of pests based on location, and allowance for newer pests, categorical disease prediction output that can also forewarn.
2. Data curation : it was necessary to curate the data to come up with subsets for particular locations where a particular set of pests were recorded across a number of years such that there was sufficient data for training and testing. Any extra pests recorded during this period were first ensured that they had the same units, and then clubbed together into an others column to accommodate for real life scenarios.
3. Discretizing Pest values to ensure real-life practicality of the models : each kind of pest has a baseline population level that is present most times of the year, the assumption we make is that the farmers will be able to easily identify when a pest population is above the baseline levels just by utilizing light traps in a customary manner without resorting to meticulous measurements. This ensures that the farmers are not at pressure to identify the pest presence with extreme accuracy. A leeway is given for those pest levels in the transition zone that are randomly allocated either Yes or No on 50-50 basis. We shall first divide values into 4 quartiles with the lowest quartile having the most number of points or maximum number of weeks. The 25-50% quartile is treated as transition zone. And the top two zones are treated as Yes.
4. Splitting Training and Testing data : Because our models takes as input the weather variables as well as pest presences which are by themselves cyclical in nature, splitting in random manner can result in the models being trained more on a particular period of the year. To avoid this, we randomly selected x% of the years for which data is available, and then passed the entire year’s data for training the model, and the remaining years are used for testing. This ensured that the training sets are based on the same distributions of weather patterns that we encounter in a particular year.
5. Deep Learning Modeling and Validation – We will be utilizing whichever is the best and describing if we are making any changes. Compare the results with a purely weather based system, weather and pest but with actual pest numbers. The former doesn’t include pests at all. So does including pests into our models raise the accuracy? Also, the latter requires strict monitoring of pest numbers, what is the accuracy we lose out when compared with latter?
6. Forewarning System – Training on disease outputs 2 weeks later to develop a system that can predict disease onset 2 weeks in advance. Testing out various weeks precedence and comparing how the results vary, and which is the sweet spot for predicting. 1-4 weeks. Repeat the comparative analysis of Step 5. Make forewarning models of the two other systems also and find out which one performs best, and try to see if we can make any changes at all?
7. Measure how allowance for flexibility allows for improvement or decrease of accuracy? Talk about others column and how they accommodate for the model being flexible.

The numbers and results

Cotton for Bacteria Leaf Blight at Lam –

Main pests – Thrips, Whitefly, Jassid

Years – 92, 94, 99, 01

Other pests – American Bollworm in 92,94,99; Aphids in 94,01; Spodoptera and Pink Bollworm in 99.

So we need to find 50 and 70 percentiles for Thrips, Whitefly and Jassid in Lam. (Same thresholds for both ALB and BLB)

|  |  |  |
| --- | --- | --- |
| Pest Name | 50% | 70% |
| Thrips | 0 | 0 |
| Whitefly | 0 | 0 |
| Jassid | 0 | 0.6 |

Rice for Leaf Blast at Rajendranagar

Main pests – BPH, Gall midge, GLH, YSB

Years – 93,94,95,96,98,99,00.

Other pests – None.

|  |  |  |
| --- | --- | --- |
| Pest Name | 50 | 70 |
| BPH | 0 | 100 |
| Gall midge | 0 | 1 |
| GLH | 5 | 275 |
| YSB | 12 | 50 |