***The effect of the greenhouse simulation on the abundance of Megostigmata and the best set of predictors for predicting the abundance of Megostigmata*** *by Akshay Sashi Menon; Harish Harish; Mudit Bhargava; Rishabh Ajay Vishwakarma; Sathishkumar Ravichandran; Suraj Sankar Mandal; Tara Vivash.*

**Introduction**

The case study is done by team of seven members to do data analysis on Mesostigmata mites in the soil during a year, in different ecosystems. The examination depended on the information gathered in the San Jose segment Nuevo Mundo, situated in the Blanco River bowl of the Municipality of Calera. The prairie was chosen as a differentiating biological system and thinking about that cows are one of the fundamental monetary exercises of the region. The prairie comprised of the species Pennisetum clandestinum Hochst. ExChiov. (Poaceae) and Lolium sp. (Poaceae). The separation between the biological systems was roughly 500 m. An analysis was set up to reproduce the greenhouse effect and to record the reaction of edaphic Mesostigmata vermin during a year, in every environment. To reproduce the greenhouse effect, altered open chambers were built from past examinations (Marion et al., 1997; Aronson and Mcnulty 2009). In this case study we will be looking into a dataset collected over a period of year. During a time of one year, six samplings were completed in the principal seven day stretch of like clockwork: February, April, June, August, October and November 2010. Quickly before assortment, the dirt temperature was estimated at 10 cm profundity with a dirt thermometer. Soil tests were taken inside and outside each pyramid in all the periods. Each dirt example related to a chamber 5 cm in distance across and 15 cm profound, taken with a drill. Each example was set in a plastic pack and shipped in a fridge to the Laboratory of Ecology of Soils and Tropical Fungi, Bogotá, Colombia. In the wake of moving the soil examples to the research facility, everything was homogenized and 150g were taken for the extraction of the bugs. Expulsion of the fauna was finished by putting each example in a changed Berlese pipe, in which it remained roughly five days. The assortment jars of the pipes contained a 70% alcohol solution wherein the mites were put away. Here we will explore how the abundance of Mesostigmata mites is affected by the simulated greenhouse effect. We will also create predictive models to assess which predictors are important in the estimation of the abundance of these mites.

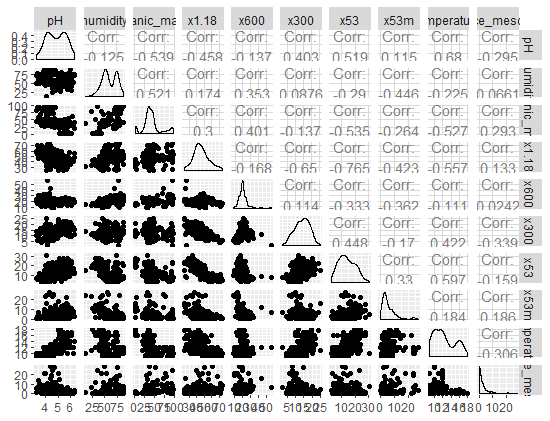
**Analysis**

*Data Cleaning*

At first, we examined to see which values were missing. We discovered one full row which had missing values for all sections including response, so we expelled this column. We also found that another three rows had missing values in the sp.Gamasina, sp.Uropodina and the response column, so we deleted these three rows as well. We then proceeded to omit columns which were not needed for the prediction, these included sp.Gamasina and sp.Uropodina and the ID. We had a look at the summary of our data and a table of abundance\_mesostigmata column. This indicated the observation 125 was a possible outlier as it had 91 for its response and could bias any predictions so we decided to also remove this observation. Also, while working on the data cleaning and manipulation part we had to work on choosing the predictors. During which we had to change a few of the things done in the pre-processing steps to make the predictions better.

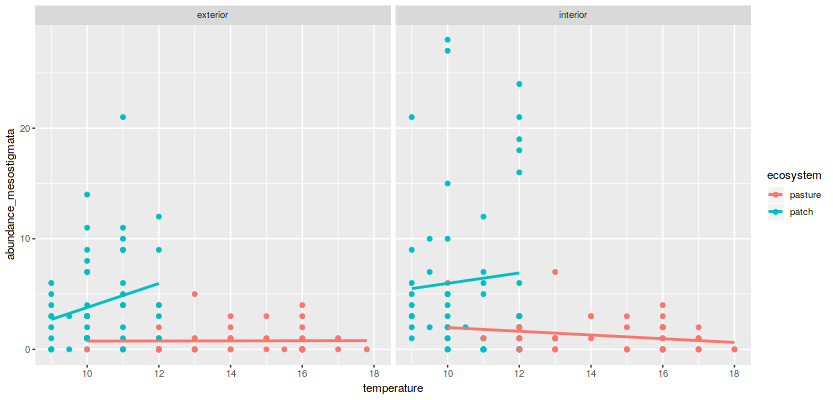
*Data Visualization*

We produced the following visualization which indicated what the best predictors would be and indicated collinearity amongst the predictors.

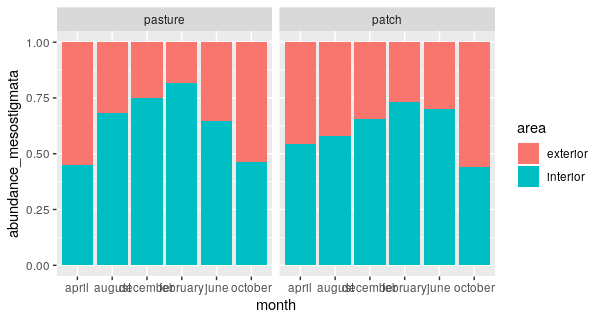


*The Greenhouse Effects*

We produced the following visualization to compare the variation in abundance of Mesostigmata found at different temperatures, external versus internal, for pasture and forest land. From this we see a higher percentage of Mesostigmata were found inside the pyramid rather than outside the pyramid indicating that the greenhouse simulation affected the abundance of Mesostigmata. The difference appears to be more significant for the patch ecosystem.



We produced the following visualization to compare the variation in abundance of Mesostigmata found during each month, external versus internal, for pasture and forest land. From this we see a higher percentage of Mesostigmata were found inside the pyramid rather than outside the pyramid indicating that the greenhouse simulation affected the abundance of Mesostigmata.



*Predictions*

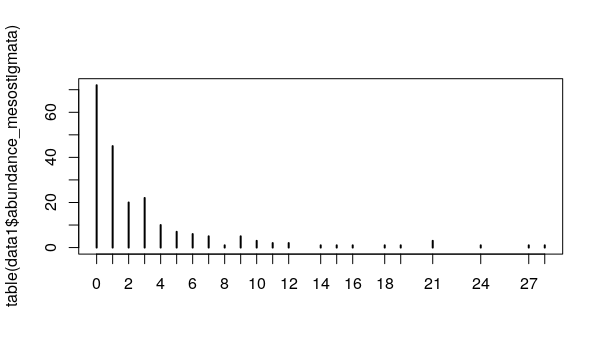
We use the AIC of the base model as multiple predictors did not seem to be as important. We then used the Chi-squared test to check for overdispersion in the model. By using Quasi Poisson Regression, we found a very large number of zeroes in the data suggesting that there may be zero inflation in the data.

The Model predicts 24% observations to be 0 but about 34% of observations are 0. This is evidence that there may be a zero inflation in the data. A zero inflated model of negative binomial distribution gives the best AIC values.

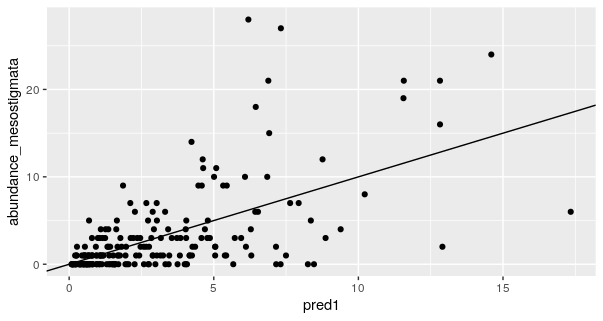
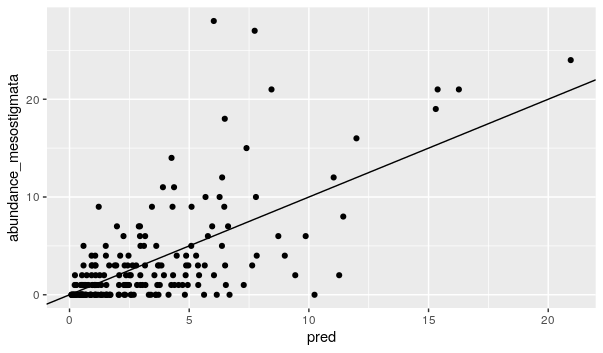
Thus, comparing the Mean Absolute Errors from both models does not show much of a difference. Since the dataset is small it isn't possible to perform a formal testing on a test set. So, we decided to choose the best model based on a combination of AIC values and Mean Absolute Errors.

**Result**

We found the best set of predictors to be area, pyramid, month, humidity, organic matter, x600, x300 and temperature. To improve the model, we used these predictors to build a zero-inflation model. This plot further supports the evidence for zero inflation as there are quite a few more zero values.



We found the zero inflated negative binomial to be the best prediction based on the AIC value of 881 compared to 1007. However, on comparing the Mean Absolute Errors from both models, it indicated that the Poisson regression gives a mean squared error value of 2.2913 which is better than the negative binomial model which gives a mean squared error value of 2.4228. The graphs below further indicated that the Poisson Model is the best fitted model. Graph 1 indicates the Poisson regression model and graph 2 indicates the negative binomial model.



**Discussion**

It was apparent that the greenhouse effect simulation affected the abundance of Megostigmata and the type of ecosystem also influenced how much the greenhouse effect simulation contributed to the abundance of Megostigmata. This was a small data set, so we chose our best set of predictors based on the AIC values and Mean Square error rather than splitting the data into testing and training data. Along with the general SAS and R programming we have used the Shinny app to show the prediction of the mites, which accepts user input for the best set of predictors. Further visualization analysis was carried out which can be seen in the Analysis tab in the app.

**References**

(Marion et al., 1997; Aronson and Mcnulty 2009)

Dr Rafael de Andrade Moral (2020) “MiteResponse InfoFile” Maynooth university notes by Dr Rafael de Andrade Moral Available at: https://2020.moodle.maynoothuniversity.ie/pluginfile.php/329527/mod\_resource/content/

Dr Rafael de Andrade Moral (2020) “Data Cleaning and Manipulation in SASFile” Maynooth university notes by Dr Rafael de Andrade Moral Available at: https://2020.moodle.maynoothuniversity.ie/pluginfile.php/329527/mod\_resource/content/0/ST662\_Section3\_CleaningManipSAS.pdf