**Final Project QMEE**

**Alejandro Rojas**

**Introduction**

The crop establishment is key factor for soybean growers at the beginning of the season where most of the time, seedlings are subject of the attack of soilborne pathogens, resulting in damping off and poor plant establishment. Soybean seedling diseases impact the plant root system at the beginning of the season, however the root infections can occur at later stages, impacting crop physiology and yields. In 2005, seedling diseases caused losses around 829 tons in the US (Wrather & Koenning, 2006). Different factors could influence the outcome of the interaction of soilborne pathogens and soybean plants, such as seed genetic background, the climate, edaphic factors and their interaction with the soil inhabitants (Broders et al., 2009). These soilborne diseases have been attributed to several pathogens most of them fungi and oomycetes (fungi-like organisms). However, the key species playing a role in disease are not fully known (Broders et al., 2007).

In fact, the oomycetes *Pythium* and *Phytophthora* are two of the main causal agents of soybean root diseases, for instance losses have increased by four-fold in the last ten years (Koenning & Wrather, 2010). The increased incidence is related to some of cultural practices now being used by growers, like no-till and early planting. These practices and the environment interact strongly with the microbial communities present in the soil, causing shifts in the different species that exist in this ecosystem (Arcate et al., 2006). The knowledge of the diversity of species causing soybean seedling diseases in the US Midwest is limited. We initially utilized a culture-based approach to start understanding the diversity of oomycetes associated with soybean seedling diseases in the US sampling in 2011 and 2012. A semi-selective medium was selected to increase the recovery rate of a group of organisms, in this case Oomycetes. The identity of the species was confirmed based on the sequencing of Internal Transcribed Spacer (ITS) of the ribosomal DNA.

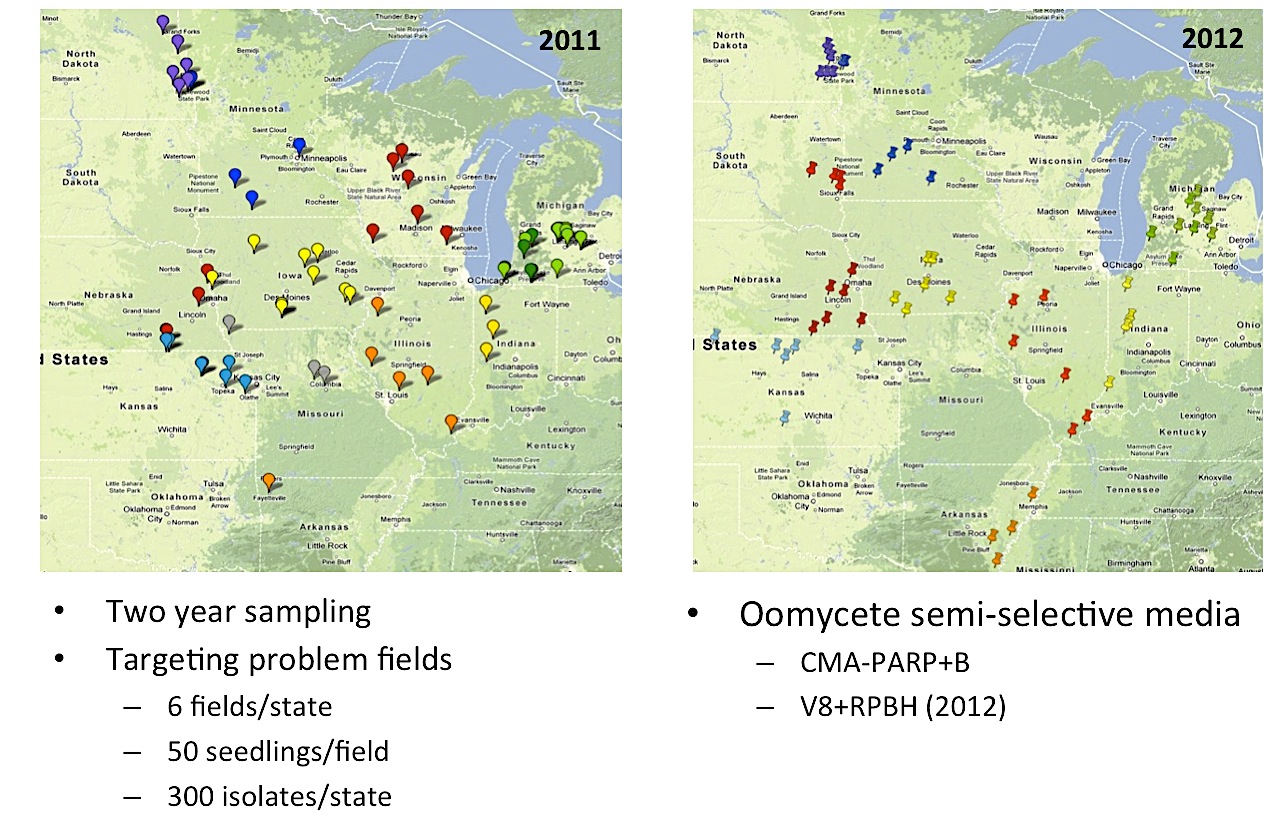
Oomycetes are known to be aggressive plant pathogens, but the number of species associated with different plant species or different ecosystems is not well known. Recent studies have described multiple new species based on soil and water systems surveys (Kang et al., 2010). Different studies have been done co-relating environmental data with fungal species distribution, resulting on the identification of different drivers such a soil properties and latitude (Tedersoo et al., 2014). However, the relationship of species distribution and environmental factors is undetermined for the oomycetes. The research question focuses on what are the oomycete species associated with soybean seedling diseases. Our hypothesis is that seedling diseases are cause by multiple oomycete species, which some of this species could be acting as pathogen complexes to cause disease across the Midwest. In addition, we also hypothesize that there are climatic and edaphic factors driving the presence and abundance of these oomycetes on the different fields.

The study of the diversity of oomycete pathogens associated to crops like soybean will provide a better understanding of the complexes of organisms that might be associated with disease under specific conditions, for instance climatic and edaphic variables. In order to conduct the analysis different linear models will be evaluated to disentangle the role of different factors on oomycete diversity, also including geographical variables such as latitude to evaluate the hypothesis of diversity on relation to distance from the equator. The understanding of microbial communities associated with soybean will be significant since it is major crop with economical importance not only as staple food, but also as biofuel sources to reduce the impact of potential pathogens, as we understand their distribution, frequency and abundance.

**Methods**

***Sample collection, species ID and climatic data***

Root rot symptomatic soybean seedlings were collected from fields with a history of plant establishment issues and damping-off. A total 50 symptomatic seedlings were collected in W-shaped transect in a field (replication unit). Within a state, 6 fields were sample and a total of 10 to 12 states were sampled in two years 2011 and 2012 (Figure 1). Thus, a hierarchical sampling was conducted collecting plants within single fields within states in the Midwest.



**Figure 1.** Soybean fields sampled across the Midwest during 2011 and 2012.

Symptomatic soybean seedlings were washed with tap water, air-dried and isolations were done placing root tissue on CMA+PARP semi selective media. Plates were incubated at room temperature and visually inspected for hyphal growth after 3 days post isolation. Hyphal tips were transferred to clean media and pure cultures were stored and grown for DNA extraction. DNA extractions were conducted at the genomics center (RTSF – MSU) using a phenol-chloroform automated system (AutoGen Inc., Holliston, MA). Isolates were identified by amplification and sequencing of the Internal Transcribed Spacer (ITS) of the rDNA using primers ITS6 and ITS4 (Cooke et al., 2000). Sequences were assembled using CodonCode Aligner (CodonCode Corp., Dedham, MA, USA) and corroborated against a local database.

Climatic data was collected based GIS coordinates for each sample location and used to query different climatic databases to obtain different environmental parameters. Data was obtained from the database PRISM (<http://www.prism.oregonstate.edu/>) as shape files, and it was imported into DIVA-GIS (<http://www.diva-gis.org/>) for their correlation and extracting climatic parameters for each of the specific locations included in this study.

***Statistical methods***

*Data exploration*

A dataset was constructed based on number of species identified and diversity indices. In addition, based on latitude and longitude also included in the dataset, climate data such as temperature and precipitation were collected. Data was initially explored to determine the transformation of the data or exclusion of data. The field corresponding to Arkansas collected on 2011 was removed from the data due to issues during data collection since only one field was sampled, and it was oversampled deviating from methods followed in other fields. Parameters that contain quantitative data, such as temperature, precipitation and latitude were centered to reduce possible issues with colinearity (Table 1). These parameters were centered and added to the data to explore data and construct models. As part of the exploration, plots of different parameters were examined against operational taxonomic units (OTU) or Shannon index.

**Table 1.** List of parameters utilized to build the different models in this study.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Class** | **Data Type** | **Biological Importance** |
| Operational Taxonomic Unit (OTU) | Numeric | Absolute | Sequence based definition of organisms, for instance based on 97% similarity |
| Shannon diversity index | Numeric | Absolute | Diversity index based on the formula H’= |
| Latitude | Numeric | Interval | Distance fro the equator, it has been shown to correlate with species diversity |
| Temperature | Numeric | Interval | This parameter could affect growth rate and niche of the different oomycete species |
| Precipitation | Numeric | Interval | Accumulated water could play an important role in disease, since it could enhance spore germination |
| Year | Factor | Nominal | Temporal variation of species, also dependent on weather conditions |

*Linear model and model selection*

Different linear models were constructed with two different response variables in order to determine the best fitting model to the oomycete diversity data. The goal is to determine which one of the responses is better to explain the diversity of these organisms, thus we can predict their distribution and associate this data with pathogenicity of different species. Initially, climatic variables were included in the different models in order to determine how well these parameters explain the diversity of the oomycete species. In order to preselect among different models, we initially propose multiple models for both responses: Shannon diversity index (Table 2) and OTU (Table 3). All of the models developed for these two response variables were evaluated using Akaike Information Criterion (AIC), the Log Likelihood, delta AIC and adjusted R2.

**Table 2.** Models preselected for Shannon index as response variable.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Equation** | **Parameters** | **AIC** | **Delta AIC** | **Log Likelihood** | **Adj.**  **R2** |
| Shannon ~ Latitude |  | 2 | 227.9 | 0.0 | -110.99 | 0.056 |
| Shannon ~ Latitude + Temperature |  | 3 | 229.9 | 1.8 | -110.99 | 0.048 |
| Shannon ~ Latitude + Temperature + Latitude\*Temperature |  | 4 | 231.9 | 2.0 | -110.95 | 0.039 |
| Shannon ~ Latitude + Precipitation + Latitude\*Precipitation |  | 4 | 231.6 | 2.0 | -110.79 | 0.042 |

**Table 3.** Models preselected for OTUs as response variable.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Equation** | **Parameters** | **AIC** | **Delta AIC** | **Log Likelihood** | **Adj.**  **R2** |
| OTU ~ Temperature |  | 2 | 805.6 | 0.0 | -399.80 | 0.024 |
| OTU ~ Latitude + Precipitation + Latitude\*Precipitation |  | 4 | 805.5 | 2.0 | -397.73 | 0.042 |
| Shannon ~ Latitude + Precipitation |  | 3 | 805.6 | 2.0 | -398.82 | 0.032 |

The candidate models were evaluated based on different statistical parameters under the general linear model, in order to establish their fitting. Coefficient plots and autocorrelation plots were examined to determine if data was problematic, and to determine the colinearity or other issues that could be affecting the use of the linear models. Two best fitting models were selected based on the lack of correlation and also based on the complexity of the models for further analysis. These models were further evaluated using variance inflation factors (VIF) were calculated for both models, and analyzed using the rule of thumb of VIF>10 to determine the existence of issues in the model. In addition, diagnostic plots were generated for both models under the two different responses.

*Bootstrap residual resampling for parameters*

In order to establish confidence intervals for parameters included in the models selected before, a bootstrap analysis was conducted on the residuals, having in mind the determination of biases of the data collected during the sampling. In this case, we assuming that temperature, latitude and precipitation are fixed, and there limited error in their measurements. The bootstrap analysis was conducted on the two models using Shannon diversity index and OTU number respectively. In this case a total 5000 replications were conducted to determine confidence intervals using bootstrap function.

*Power analysis*

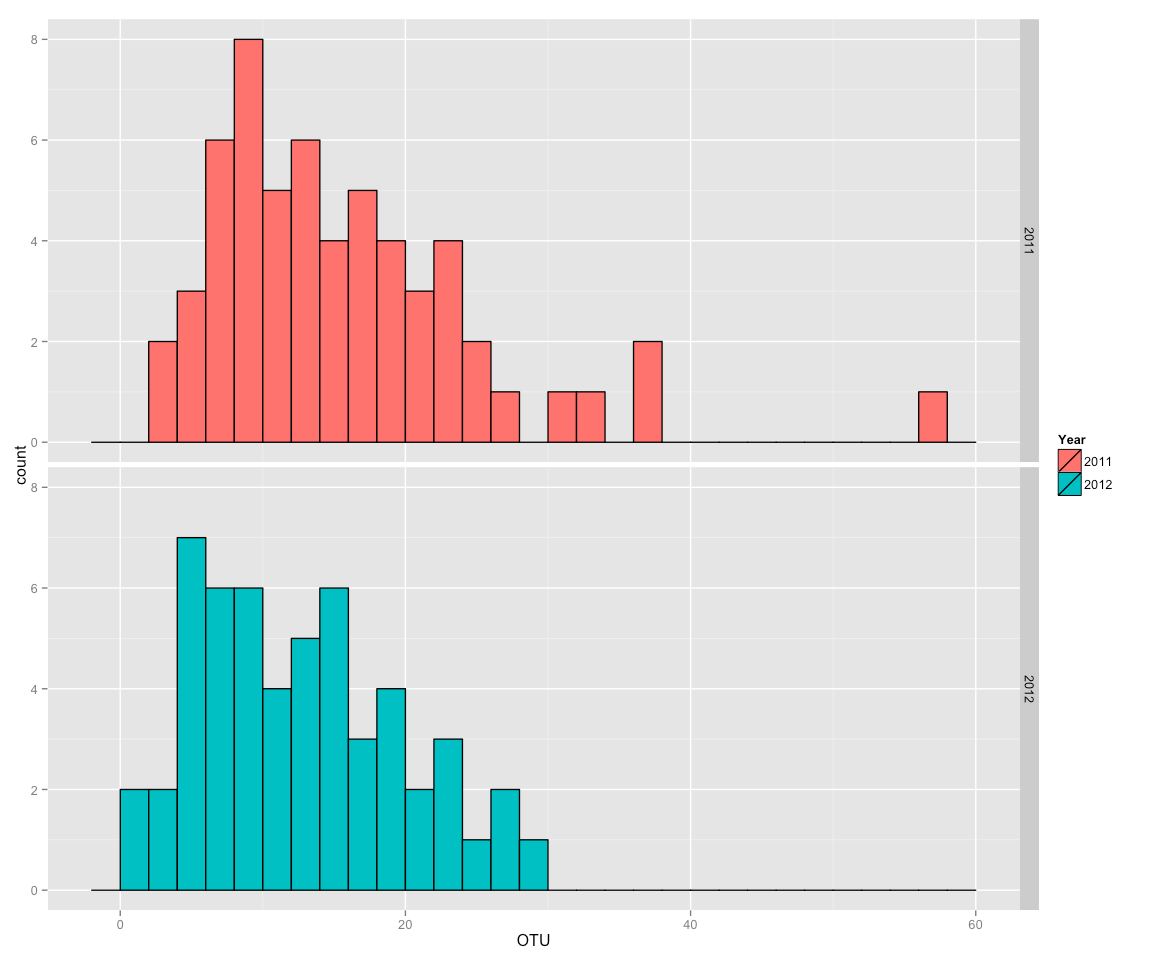
Based on the models previously analyzed and considering the coefficients for these models, we used power analysis to determine the sample size required to have enough power to make any biological inferences based on the model that have been selected. For this approach a monte carlo simulation was conducted using the coefficient of the models, in this case the intercept and assuming a slope that ranges from 0 to 1. A total of 1000 replicates were conducted, increasing sample size by 20 in a range of 10 to 200 samples. Results were plotted as contour plots.

*Mixed models*

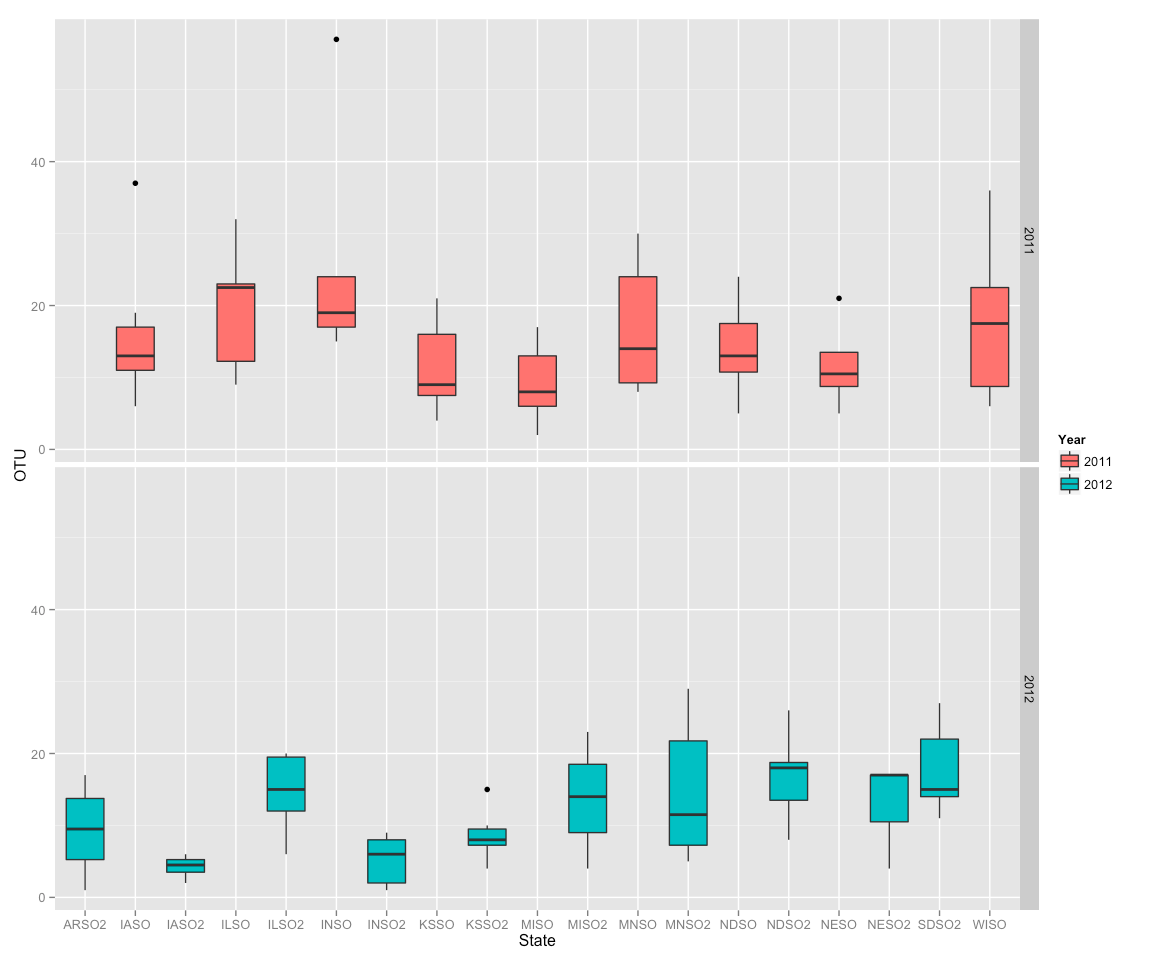
The best fitting model were reevaluated using mix model to account for the differences on sampling between years, since conditions, specifically climate was quite divergent. A covariance structure based on the factor year was considered under the simple covariance method known as compound symmetry. The models were re-evaluated using AIC.

**Results**

The diversity of the oomycetes across the soybean producing area ranges from mostly from 5 to 20 species per field if we consider the OTU parameter (Figure 1 and 2). The diversity index actually suggests a really tight diversity across the entire set of fields sampled. Nonetheless, we have to consider that we are using a very traditional approach to capture some of the diversity present in the field, and we are limited by number of strains that we can recover from infected plants or the soil. In this case, we are using culture-based technique and the set of species that could be recovered is probably limited due to different biases. Thus, when this information on the number of species present per field or the diversity tries to be correlated with other parameters, the models are going to be limited to the data that we have available.

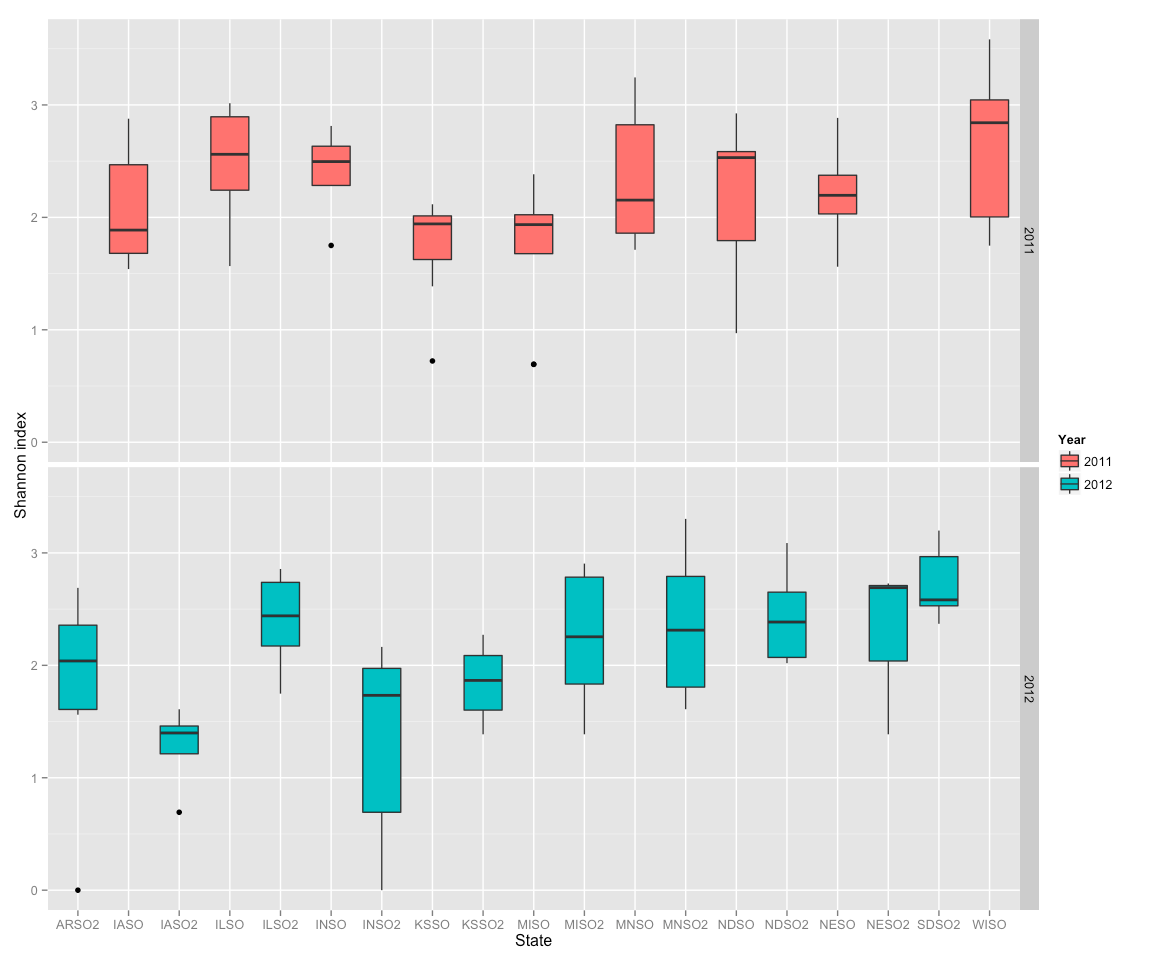


**Figure 1.** Histograms of OTU number found in the Midwest by year.

****

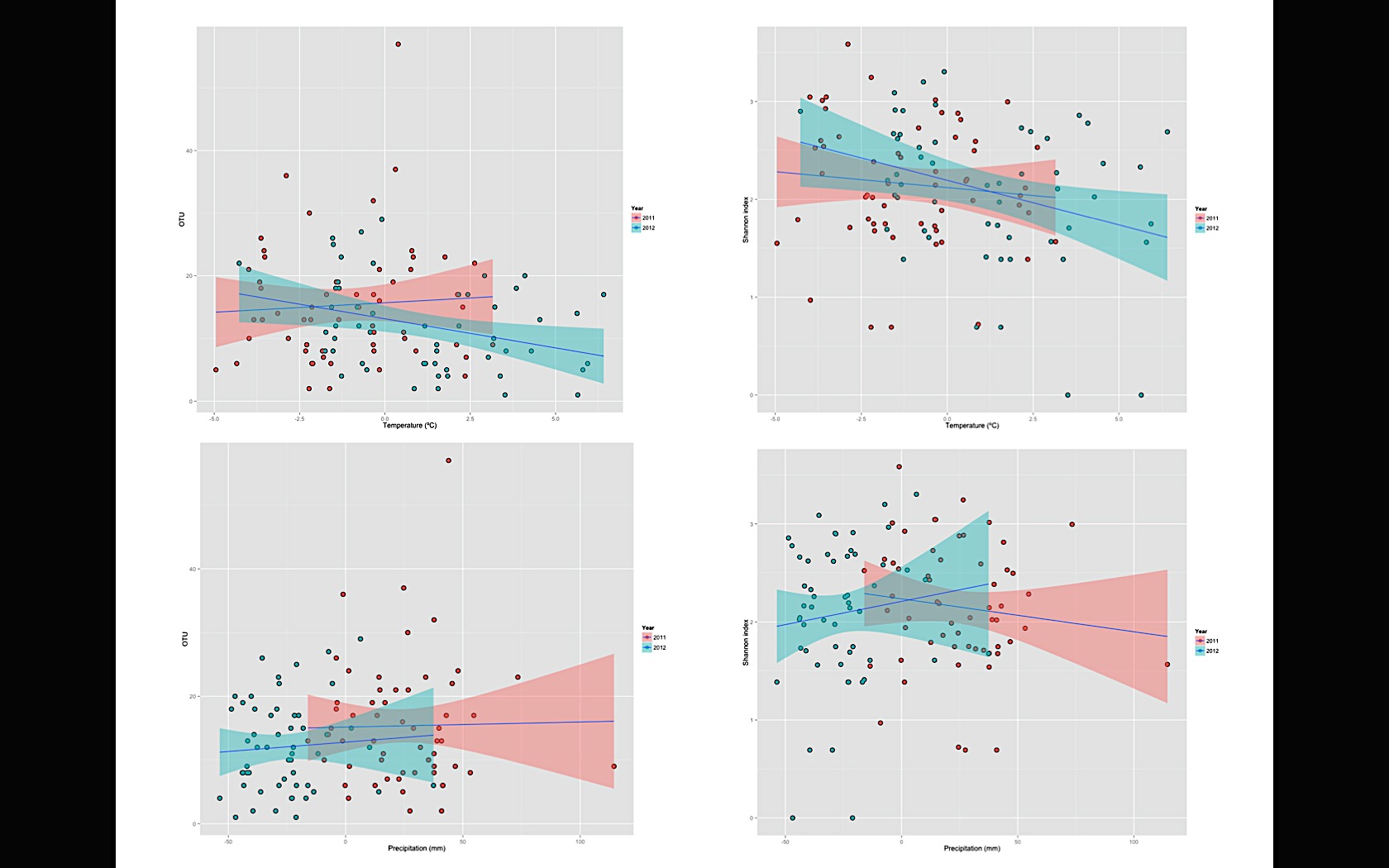
**Figure 2.** Box plot of OTU numbers by year sampled from soybean fields in 2011 and 2012.

In this case, Shannon diversity index is capturing the alpha diversity which a diversity pertaining only to the sample itself, which could be missing other important factors like the differences between communities at two different locations, in this case we are talking about beta diversity. However, we hypothesize that there was a correlation between diversity per field in relation to other factors like climate. In this case metadata was collected using GIS coordinates to construct models based on some climate parameters. Using this information and climate data, linear models were constructed to determine if there was a correlation of the species observed at one field with these parameters. This question will help us to understand what environmental factors could drive the presence and abundance of some species at certain locations. From the data exploration, it was noticeable that some states have a higher variability in the number of OTUs within the same state, while other states the number of OTUs is very tight. This could be a result of sampling. In terms of diversity index, this variability is stabilized since the index accounts for the sample size in each of the populations, so it will be a calculated response in the model, but it helps to improve the models (Figure 3).



**Figure 3.** Box plot of Shannon diversity index by year from soybean fields sampled in 2011 and 2012.

In addition, the exploration of different plots including some of the climate variables in relation to the two responses consider in this study showed some variability across the years (Figure 4). In general, most of the plots revealed a negative relation for some of the parameters; nonetheless it is important to mention that there are differences across the years.

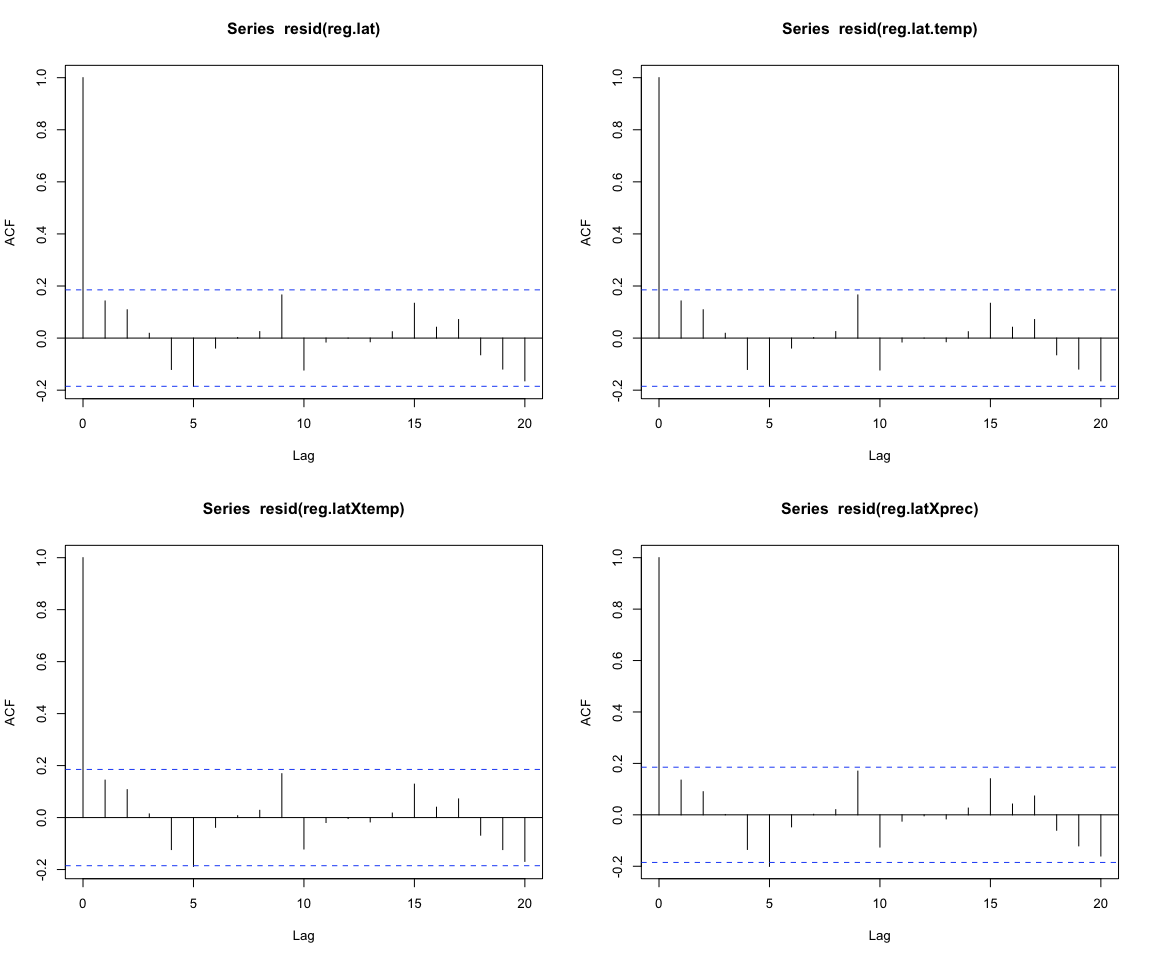


**Figure 4.** Plot of OTU number and Shannon diversity index with the two climatic parameters: precipitation and temperature by year.

The possible explanation for the different conditions could be that the conditions in 2012 were dryer (less precipitation and high temperature) than in 2011. These conditions could have a greater effect on the species observed across the locations. In the other hand, latitude seems to have a positive correlation with diversity of the species, where higher latitudes had a greater diversity than those at lower latitudes. The OTU number shows that 2011, it has negative correlation, based on only on exploratory analysis, suggesting a different behavior during this year, corroborating the effects observed with temperature and precipitation (Figures in the appendix).

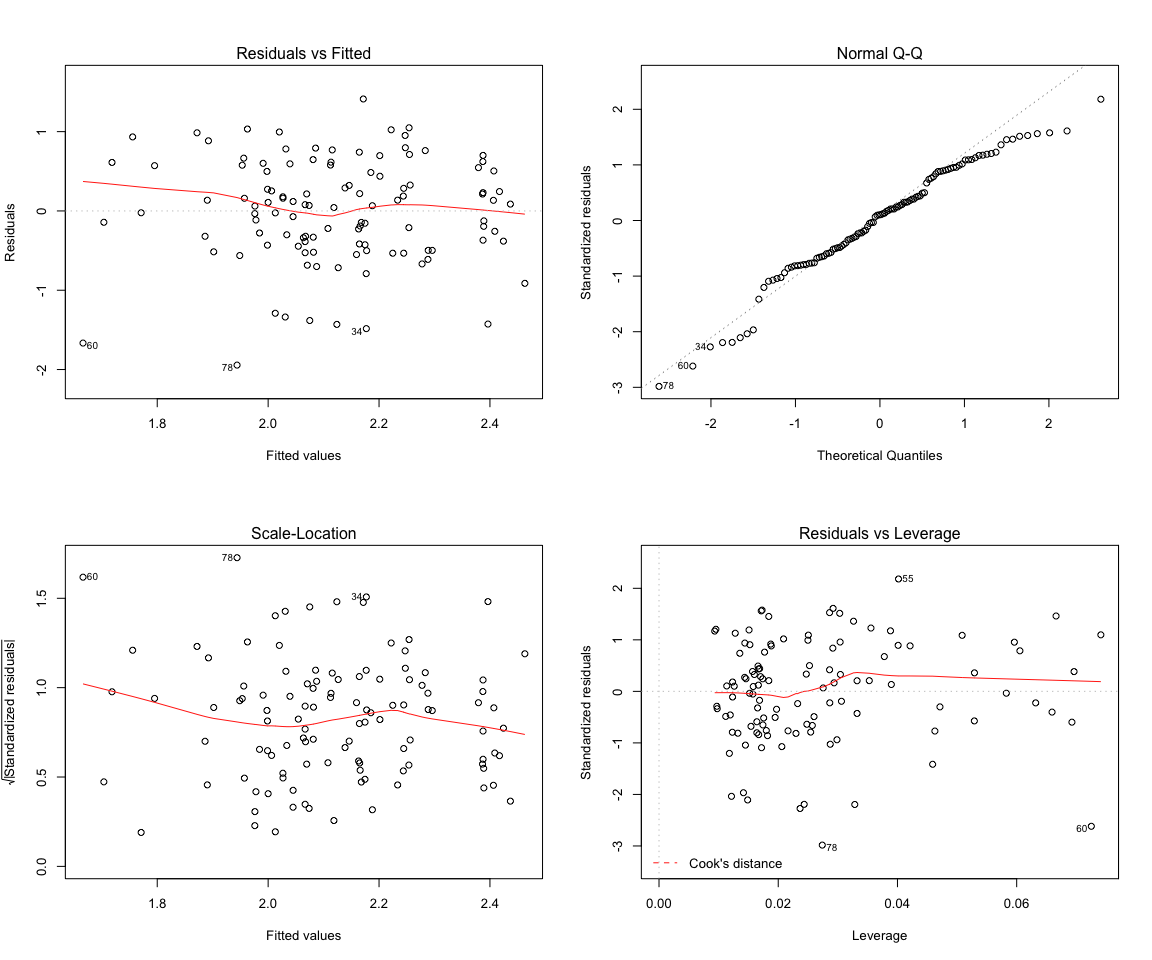
Based on these observations, different models were constructed considering the possible interactions of the different climatic and geographical parameters. Two sets of models were constructed using the two different responses: OTU number and Shannon diversity index (Table 2 and 3, complete list of models in the appendix). The different models were evaluated using Akaike Information Criterion (AIC), based on this parameter the list of models was mine also considering other parameters such as: Log Likelihood, delta AIC and partial R2. These analyses revealed a total of 4 best fitting models for Shannon diversity index as a response and 3 best fitting models for OTU number as response (Table 2 and 3). Despite of being the best fitting models, in general, the correlation of the different variables with response was low, for instance the best fitting model for OTU number was only based on latitude with partial R2 equal to 0.068.

Looking closer to the coefficients of the models, both approaches agreed in the intercepts confidence intervals being consistent across the models and regression estimates latitude in general has the stronger effect to explain the variability or the diversity observed across the samples (Appendix). It is important to mention that these quantitative parameters were centered in order to reduce the chances of autocorrelation or colinearity, and examining the models it was conclusive that all the models had none to little issues with colinearity and autocorrelation (Figure 5 and appendix).

****

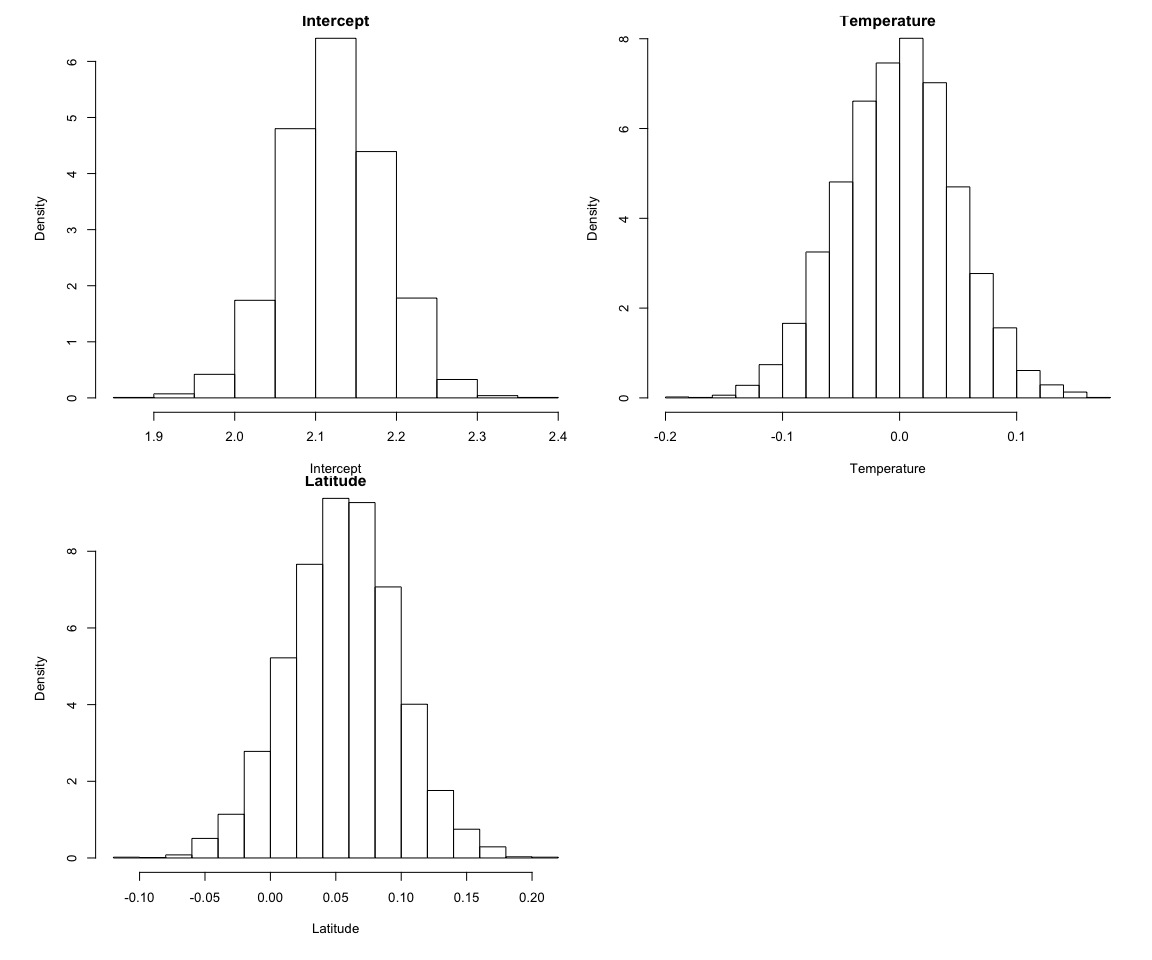
**Figure 5.** Autocorrelation plots for models using Shannon diversity index as in response in the best four fitting models.

In addition, variance inflation factor was examined for the best fitted complex models, in this case: **‘Shannon ~ latitude + temperature’** and **‘OTU ~ latitude + precipitation +latitude\*precipitation’,** the VIF was 4.21 and 1.32-1.53 respectively. This indicates for both models there is low colinearity discarding any possible issues related to this factor. A closer look was taken on both models of the models, in this case the Shannon model based on latitude and temperature; the different plots suggested a lack of pattern on the residual plots. However, there are some residuals that are outliers in some of these plots, which suggest that some sites sampled have different behavior (Figure 6). The case is not same for the OTU model since there are some patterns across the different plots (appendix). Actually, some of the sites had a increased number of OTU different form the rest of the samples, which could be a result of the method use recover the isolates since media could have biases, meaning that most of the fields we were able to recover similar species and levels of diversity, but in certain specific cases other species were favored generating these outliers.



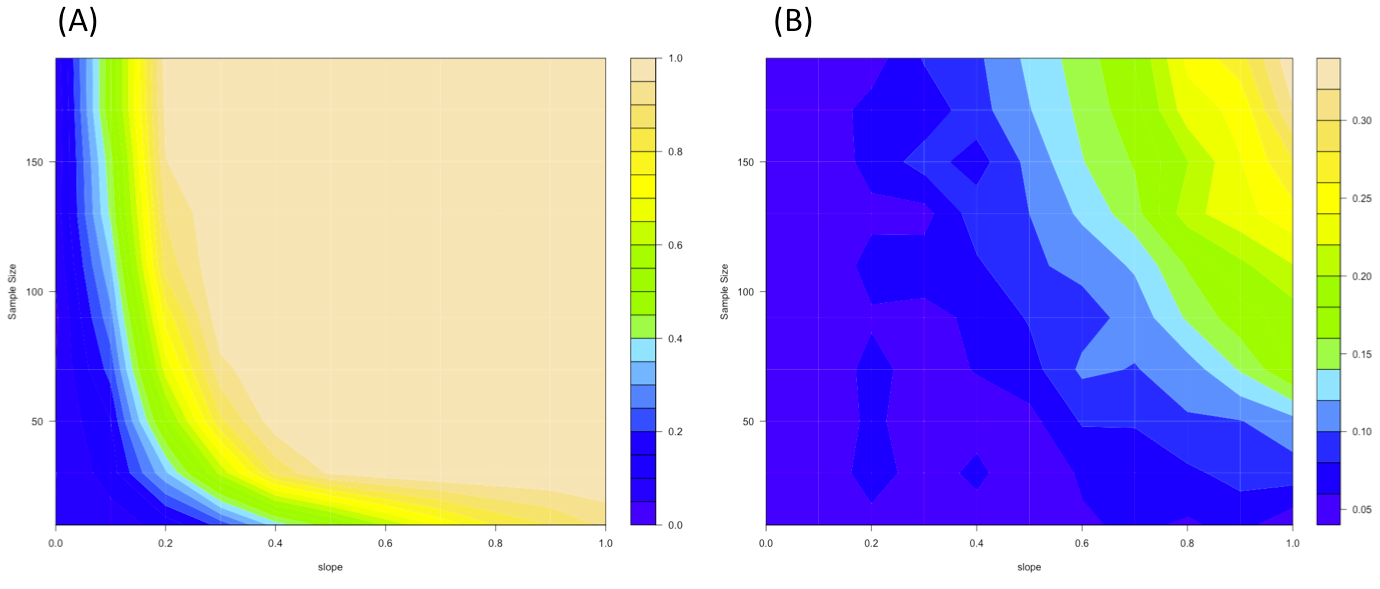
**Figure 6.** Diagnostic plots for the model ‘Shannon ~ latitude + temperature’.

The confidence intervals for different parameters were estimated using bootstrap approach to evaluate the range of potential values that we have in both models, in general, we expect a average diversity 13 species (2.1 for Shannon index) and this values seem not to vary greatly, however for the other parameters, latitude was the other parameter with smaller interval, whereas precipitation seems to vary greatly (Figure 7). This suggest that temperature is not a strong effect and could be related to species, but so far it does not perform really well in the models as it is with latitude.



**Figure 7.** Histograms of coefficients after bootstrap residual resampling for shannon ~ latitude + temperature.

A similar effect was also observed in the OTU model that temperature has a wider confidence interval, probably for the reduced effect that this parameter can have on the prediction of this response (Appendix figure). Again latitude has a narrower confidence interval, and it suggest again the power of latitude to predict the species diversity in the model, this agrees with different studies that concluded that latitude could have a greater effect on species distribution, but it is also important to understand that some climatic parameters could be related to latitude.



**Figure 8.** Power analyses for model (a) ‘Shannon ~ latitude + temperature’ and (b) ‘OTU ~ latitude + precipitation +latitude\*precipitation’.

A power simulation analysis was conducted for both models and it was perceptible that the model based on Shannon as response has more power to detect differences across samples, whereas the OTU model has very limited power despite the number of samples (Figure 8). In the first model, using Shannon index, a total of 150 fields could be sampled to have a power 0.6 that could be enough to describe the diversity of an area. Nonetheless, if we used approaches that have less biases like amplicon community analysis, where more species can be detected in a sample and better understanding of the community can be generated. In that case a reduced number of samples and more replicates per field could be done increasing the power and the behavior of model.

**Table 4.** Summary of mixed models accounting for Year as fixed effect.

|  |  |  |
| --- | --- | --- |
|  | Dependent variable: | |
|  | shannon | OTU |
| ct.lat | 0.068 | 0.138 |
|  | (0.055) | (0.311) |
| ct.temp | 0.017 |  |
|  | (0.073) |  |
| ct.precp |  | -0.005 |
|  |  | (0.042) |
| Year2012 | -0.06 | -2.649 |
|  | (0.181) | (2.589) |
| ct.lat:ct.precp |  | -0.015 |
|  |  | (0.01) |
| Constant | 2.152\*\*\* | 15.115\*\*\* |
|  | (0.108) | (1.494) |
| Observations | 112 | 112 |
| Log Likelihood | -118.633 | -401.557 |
| Akaike Inf. Crit. | 249.266 | 817.113 |
| Bayesian Inf. Crit. | 265.359 | 835.823 |

In addition, mixed models were considered for both approaches including year a fixed factor in the models. The rationale behind this is that across the two years there was differences in climate that could affect the behavior of both models, so the idea was to account for variance across the years. Using simple covariance method both models were analyzed using year as a fixed factor, and a summary of both models is presented in table 4. The first model actually lost power on predicting the species diversity and it had reduced fitting, worsening the model. This could be also related to response variable, that considers the species present and their abundance normalizing some variability across samples. Whereas the model using OTU number does not account for this differences, therefore, the second model including year actually increases the fitting and improves the model in general becoming a better model than the previous models selected.

In general, the diversity of the oomycete species so far is difficult to model with current parameters evaluated, nonetheless it follows similar patterns to other organisms previously studied like plants and fungi, where latitude plays an important roles on the species distribution and diversity. The effect of latitude was observed in this study, but the correlation was very low still, thus it is necessary to collect more metadata to be included in the models an increase the predictive power of the different models. The determination of species distribution and how climate parameters affect them will be important to determine the risk of existence of certain species in soybean fields, generating an understanding of distribution of pathogenic and non-pathogenic species.

**References**

Arcate JM, Karp MA, Nelson EB, 2006. Diversity of peronosporomycete (oomycete) communities associated with the rhizosphere of different plant species. *Microbial Ecology* **51**, 36-50.

Broders K, Wallhead M, Austin G*, et al.*, 2009. Association of soil chemical and physical properties with Pythium species diversity, community composition, and disease incidence. *Phytopathology* **99**, 957-67.

Broders KD, Lipps PE, Paul PA, Dorrance AE, 2007. Characterization of *Pythium* spp. Associated with Corn and Soybean Seed and Seedling Disease in Ohio. *Plant Disease* **91**, 727-35.

Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM, 2000. A molecular phylogeny of Phytophthora and related oomycetes. *Fungal genetics and biology : FG &amp; B* **30**, 17-32.

Kang S, Mansfield M, Park BB*, et al.*, 2010. The promise and pitfalls of sequence-based identification of plant-pathogenic fungi and oomycetes. *Phytopathology* **100**, 732-7.

Koenning SR, Wrather JA, 2010. Suppression of Soybean Yield Potential in the Continental United States by Plant Diseases from 2006 to 2009. *Plant Health Progress*.

Tedersoo L, Bahram M, Põlme S*, et al.*, 2014. Fungal biogeography. Global diversity and geography of soil fungi. *Science (New York, N.Y.)* **346**, 1256688.

Wrather JA, Koenning SR, 2006. Estimates of disease effects on soybean yields in the United States 2003 to 2005. *Journal of nematology* **38**, 173.