Aphids - Stat 433
Professor Rohe, 5/5/2021
Faith Kulzer, Caleb Stevens, Siddhant Thakur, Maxwell Small

Big Data, Small Bugs:

Modeling Midwestern Aphid Population Dynamics with Varimax and Poisson Regression

Introduction

Aphids (members of the insect superfamily *Aphidoidea*) are soft-bodied sucking insects which feed on the sap of a wide variety of agricultural crops, causing reduced plant vigor, stunting, and deformed plant parts. They are also well known as disease vectors for many crops, as well as integral components of many ecosystems as both predators and prey. For example, the invasive soybean aphid was first discovered in the U.S. in 2000, but by 2009 had already spread throughout the Midwest and Canada, quickly becoming one of the most devastating invasive insect pests on soybean plants (Lago-Kutz, et al., 2020). Understanding these rapid changes in aphid population serves two main purposes: aiding human agriculture, and improving our understanding of the ecosystems within which we exist.

For our project, we chose to examine aphid abundance across the Midwest as it corresponds with environmental factors as well as crop presence. We combined data from the soybean aphid suction trap network, crop data from the USDA's National Agricultural Statistics Service CropScape project, and weather data from NOAA's Physical Sciences Laboratory's NCEP-DOE Reanalysis 2. We endeavored to use unsupervised learning, namely Varimax rotation of Principal Component Analysis (PCA), to improve accounting of the environmental factors responsible for aphid population dynamics. We predicted that, after controlling for weather factors, crop monocultures would be associated with lower aphid diversity, with implications for sustainable, effective, and minimally destructive agriculture.

Data

Our data on aphid abundance and diversity was shared with us by researchers in Groves Lab at UW-Madison who helped establish a suction trap network: a network of traps throughout the U.S. which catch insects flying overhead. The aphids were collected weekly from 2005 through 2018 from 49 sites throughout the Midwest, with a total of 263 aphid species recorded. Each row of our data represented one weekly collection at each trap site, with each species identified in a different row, along with the count of each species. To clean our data, we pivoted the table so that the columns listed each species and the values showed the counts; this way we had each row represent one single collection point (Lago-Kutz, et al., 2020).

Our crop data was obtained from the National Agricultural Statistics Service or NASS's CropScape service, which provides annual crop data for the continental United States in raster format, making use of moderate resolution satellite imagery to identify crops (*CropScape*). The attributes of the raster files include the name of the crop, pixel_count, acreage, and value (percentage of an area) of each crop. We used the landscapemetrics package to measure the vegetation from a 50 km circle around each Aphid trap site (*Landscapemetrics*), and joined this to our data. Some sites did not have crop data available for particular years, sometimes due to cloud cover, sometimes for unclear reasons.

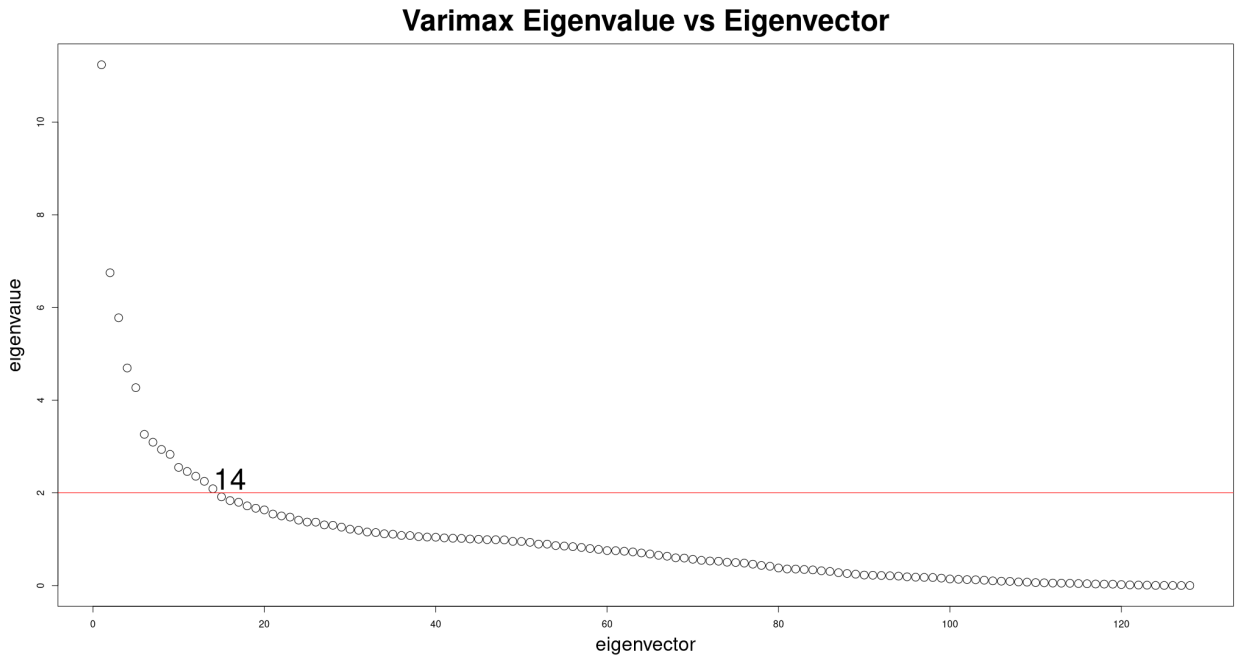
Our weather data came from the NOAA's Physical Sciences Laboratory's NCEP/NCAR Reanalysis 2 Project, which fills in gaps in actual measured weather using weather models (Gleason). This dataset consists of daily spatial weather measurements including wind vectors, air temperature, soil temperature, precipitation, humidity, cloud cover, and more . We used this daily weather data to compute weekly minimums, maximums, and averages as well as cumulative variables where appropriate (e.g. cumulative 'degree days,' precipitation, and wind). We read and cleaned this data using the R-packages 'Raster' and 'Tidyverse.'

We have not included our original weather data, as the data is multiple gigabytes, spread out over many files, and accessing and computing measurements from each file requires several hours of computation. However, the data is publicly available, and our data-cleaning code is available upon request.

Methods

After combining our aphid data with agricultural and weather data, and removing observations without crop data, our design matrix measured about 9,000 rows by 130 columns. We used Varimax rotation of PCA to reduce the dimensionality of our design matrix, in service of creating models for interpretation and prediction. Based on analysis of our design matrix's eigenvectors, we decided to use 14 principal components (PCs) as initial predictors.

Certain of our principal components (namely PC's 1, 2, 3, 6, 13) represent weekly weather conditions. Imagining that the previous week's weather might be related to the current week's aphid population, we included the previous week's values for these PCs as additional predictors, and removed observations lacking a previous week (492 observations, with average measured diversity of 6.16).



(See appendix A for principal component histograms and interpretation)

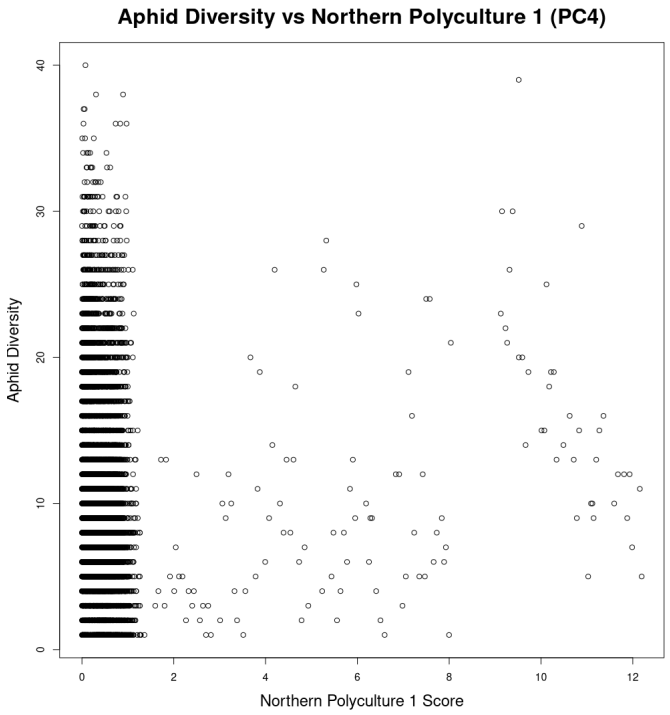
We then used Poisson regression to predict the rate of observed aphid diversity at capture sites in the Midwest. We constructed two models; a parsimonious but statistically flawed model for interpretation, and a (slightly less flawed) predictive model for forecasting. In our predictive model, dispersion was calculated to be 1.31, which we felt was close enough to 1 to that our assumption of $\lambda = \mu = \sigma^2$ for Poisson regression was reasonable.

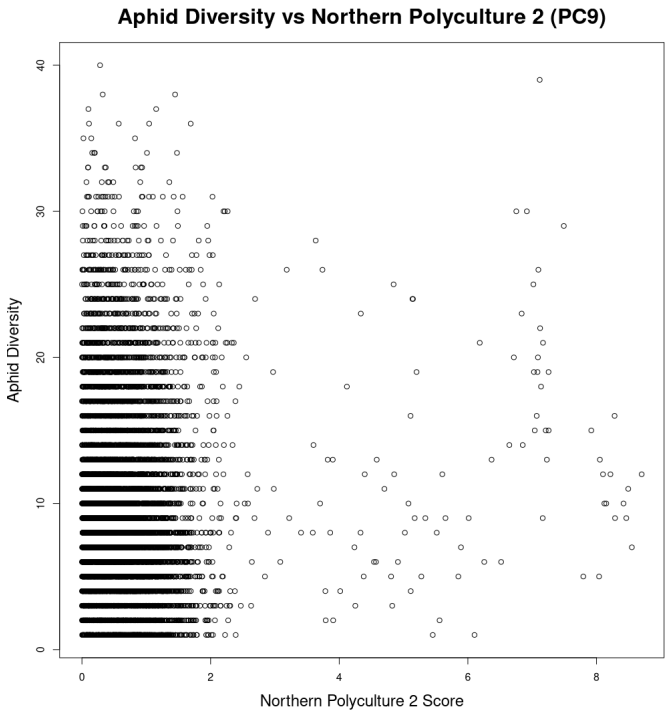
We started with a full model, including all PC's, previous week PC's where appropriate, and interaction effects. We observed a quadratic trend in residuals, and so expanded our model to include squares of PC's, which improved our model fit. At each stage we compared models using Chi-Squared tests to ensure that our newer, larger model actually accounted for more of the variation seen in the data (see appendix B for model assessment, including residual plots and analysis).

So that we could meaningfully compare regression coefficients, we standardized all predictors prior to regression. Finally, we removed four outliers with standardized deviance residuals > 6.

Results

In our model for interpretation, where predictors were standardized to allow comparison, significant predictors (p-values all ~0) most associated with aphid diversity included PC3 (cumulative weather, $\hat{B}$: 1.4), PC4 (northern polyculture 1, $\hat{B}$: 1.15), PC9 (northern polyculture 2, $\hat{B}$: 0.44), and PC1 (hot vs cold, $\hat{B}$: 0.29).



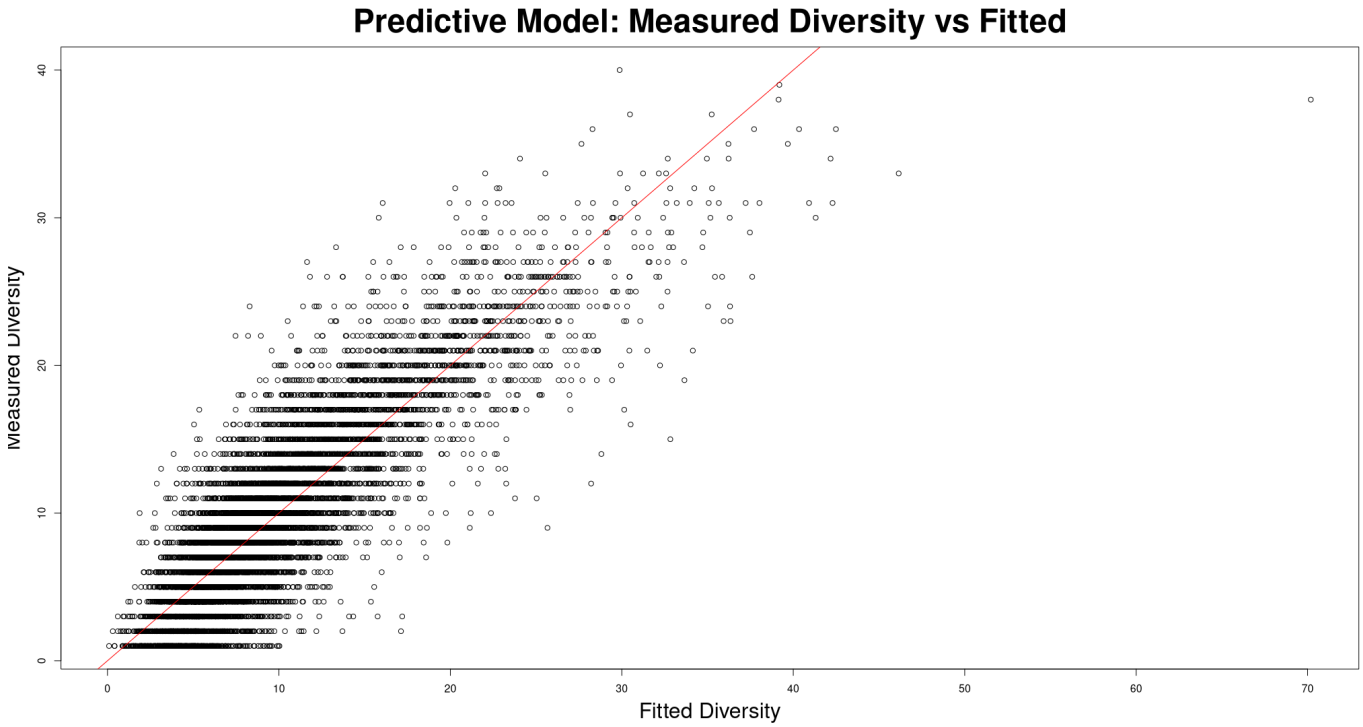


Surprisingly, in this model, southern polyculture (PC7, $\hat{B}$: -0.40) and fruit polyculture (PC11, $\hat{B}$: -0.38) were negatively associated with aphid diversity (p-val ~0). Additionally, wind from the south (PC13, $\hat{B}$: -0.91), which we understood to carry aphids from warmer southern fields into northern fields in the spring, was highly significant (p-value ~0), but was the predictor most strongly associated with a *lack* of aphid diversity. Water/forest (PC14, $\hat{B}$: -0.63) was the next predictor most strongly associated with a *lack* of aphid diversity, again surprisingly. Each of these factors, considered by itself, was correlated with an increase in aphid diversity, but when considered with all other factors, each was associated with lower diversity.

We were also interested to note that one form of development measured in our crop data was significantly associated with an increase in aphid diversity (‘development 1’, PC5, p-val ~0, $\hat{B}$: 0.11), while another form of development was also significant, but associated with a decrease in diversity (‘development 2’, PC10, p-val ~0, $\hat{B}$: -0.09).

Our predictive model is more difficult to interpret, but among significant predictors ($p < 0.01$) with the 15 largest $\hat{B}$ values, 14 out of 15 are either directly a northern polyculture principle component, or an interaction involving one or more northern polyculture PC's, again suggesting northern midwestern polyculture's close association with aphid diversity.

Our predictive model provides less in terms of understanding, but could allow a farmer to predict with a reasonable degree of accuracy the degree of aphid biodiversity likely to be found in their field at a given point in time, based on surrounding crop data and weather data from the season, and appears fairly able to predict mean aphid diversity under given conditions (Model analysis in Appendix B).



Conclusion

We hypothesized initially that polycultures would be associated with greater aphid diversity, and monocultures with lower aphid diversity. We found strong evidence that *northern* polycultures in particular are associated with aphid diversity (see *Results*). We hypothesized that aphid populations might be reflective of broader insect ecosystem health, with diverse systems of agriculture associated with aphid diversity across the board.

In reality, aphid population dynamics appear more complicated, with several polycultures (southern, fruit) observed to have a significant negative association with aphid diversity after

accounting for weather and crop data. We imagined that wilderness environments would be associated with higher aphid diversity, whereas the opposite was true, while some types of human development were associated with higher diversity.

Additionally, what we understand to be aphids' main mode of transportation (wind moving north from the American South) was strongly associated with *lower* aphid diversity. This suggests to us that aphids often migrate from the South in communities of low diversity, swarming en masse to and from resources, rather than forming stable ecosystems that remain in place. Taken as a whole, our findings suggest that aphids are more closely intertwined with human patterns of agriculture and settlement, and in less predictable ways, than we initially imagined.

Our initial hypothesis was largely incorrect, but our findings—that outside of the northern Midwest, aphids often move in waves of relatively few species— reinforced the importance of being able to model and predict aphid population dynamics. Our goal in this paper was to investigate patterns of aphid diversity, but the same modeling techniques could be applied to predicting counts of specific aphid pests.

Our analysis has a number of limitations, including its fairly limited scope. We only examined aphid diversity, which has a limited utility compared to modeling surges in aphid pest populations. Additionally, our weather data came from a fairly low-resolution source. With more time and computation resources, a more granular analysis could be performed. Finally, with further data wrangling and creative modeling, our predictive model could be fit more accurately than its current state.

We remain interested in extending our modeling approach to investigate population dynamics among specific aphid pest species, particularly well-known pests including the soybean, bean, cabbage, corn leaf, green peach, potato, melon, and pea aphids. We hope this type of further investigation could be useful to farmers, and help reduce the use of insecticides to only the most necessary and effective applications.

Shiny @ `runGitHub( "Aphids", "mcnugg3t")`

Citations

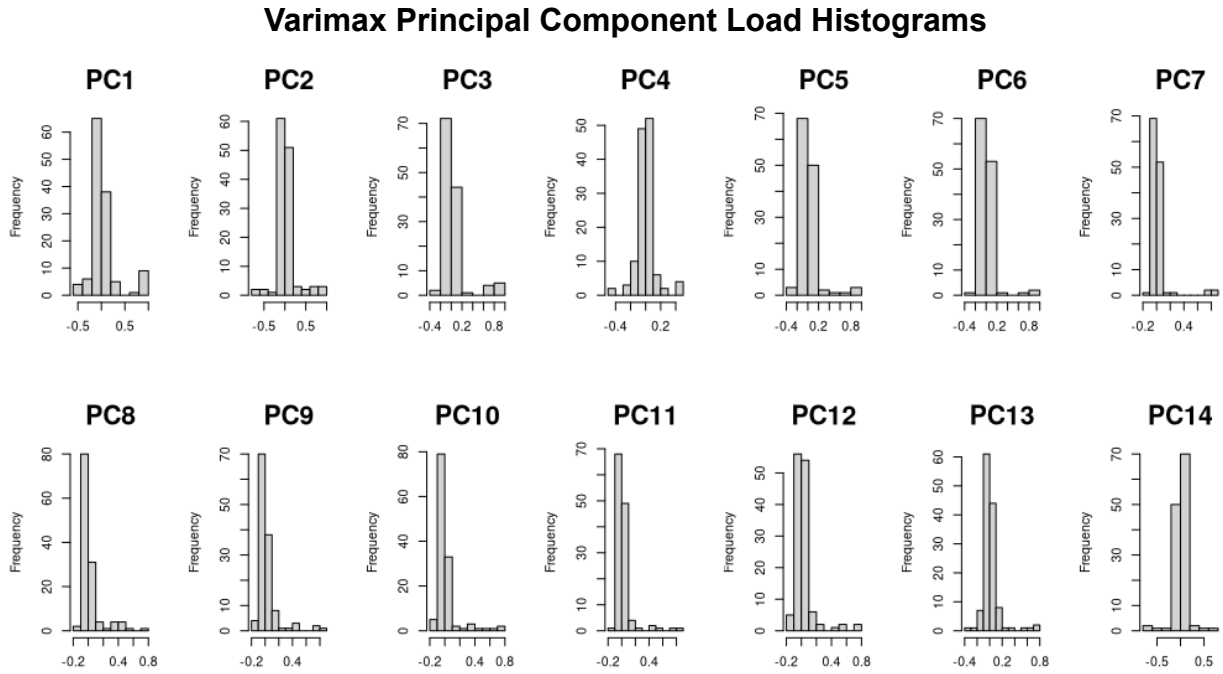
CropScape - NASS CDL Program, nassgeodata.gmu.edu/CropScape/

Gleason, Enloe. "U.S. Wind Climatology." *National Climatic Data Center*, www.ncdc.noaa.gov/societal-impacts/wind/overview

Lagos-Kutz, Doris, et al. "Soybean Aphid Suction Trap Network: Sampling the Aerobiological 'Soup.'" *OUP Academic*, Oxford University Press, 12 Mar. 2020, academic.oup.com/ae/article/66/1/48/5803396?login=true

Hesselbarth MH, Sciaini M, With KA, Wiegand K, Nowosad J (2019). "landscapemetrics: an open-source R tool to calculate landscape metrics.", cran.r-project.org/web/packages/landscapemetrics/index.html

Appendix A: Varimax Result Interpretation



Interpretation:

(Contrasts in red)

PC1 - **'hot vs cold'** - temperature, humidity vs wind from north or east, wind speed, froze

PC2 - **'moisture vs evaporation'** - precipitation, soil water, clouds vs sun, evaporation

PC3 - **'time'** - week, day of year, cumulative wind and precipitation

PC4 - **'northern polyculture 1'** - (% land) oats, clover/windflowers, barley, mixed forest, peas

PC5 - **'development 1'** - (% land) developed low, medium, and high intensity, developed open

PC6 - **'east vs west environment'** - Lat, spring wheat, wetlands, wind from E vs Lon, deciduous, rain

PC7 - **'southern polyculture'** - (% land) cotton, rice, soy + cotton, watermelon

PC8 - **'spelt polyculture'** - (% land) sugar beets, beans spelt, dry beans, cucumbers, winter wheat, Lon

PC9 - **'northern polyculture 2'** - (% land) potatoes, evergreen, carrots, shrubland, christmas trees

PC10 - **'development 2'** - (% land) non-agricultural undefined, developed, forest, water

PC11 - '**fruit polyculture'** - (% land) cherries, grapes, blueberries, apples, alfalfa, apples

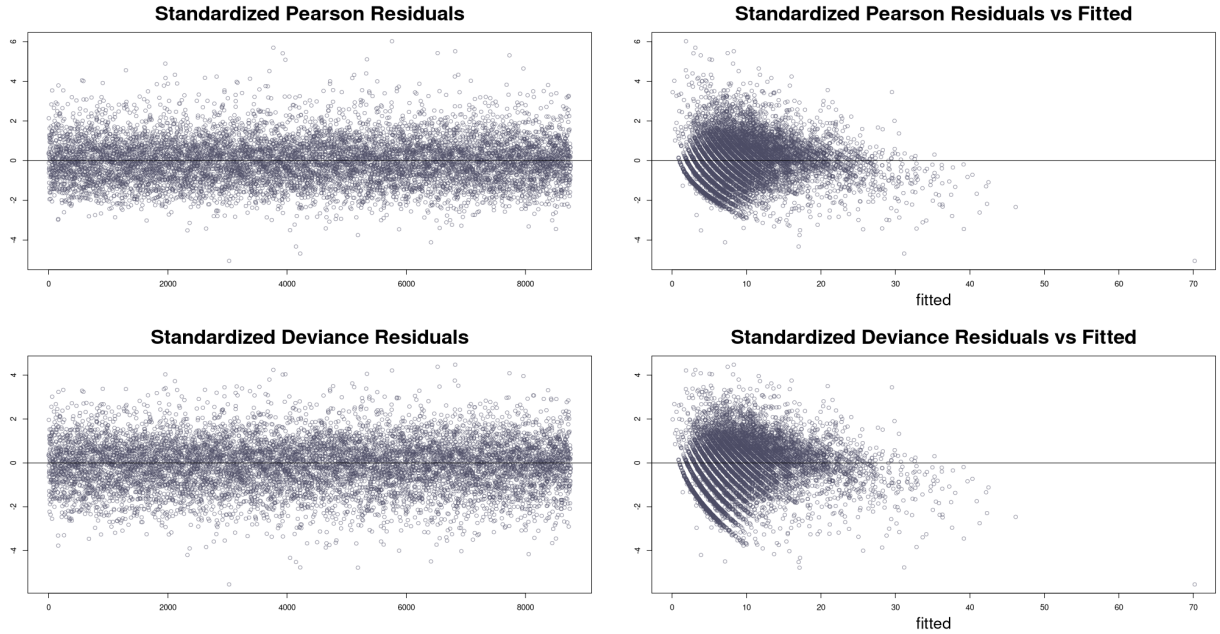
PC12 - **barren polyculture'** - (% land) squash, barren, peppers, durum, sweet corn, cranberries

PC13 - **'wind from south / west'**- wind from south (that week), wind speed, evaporation, wind from west

PC14 - **'water / forest'** - (% land) open water, woody weblands, forest, evergreen

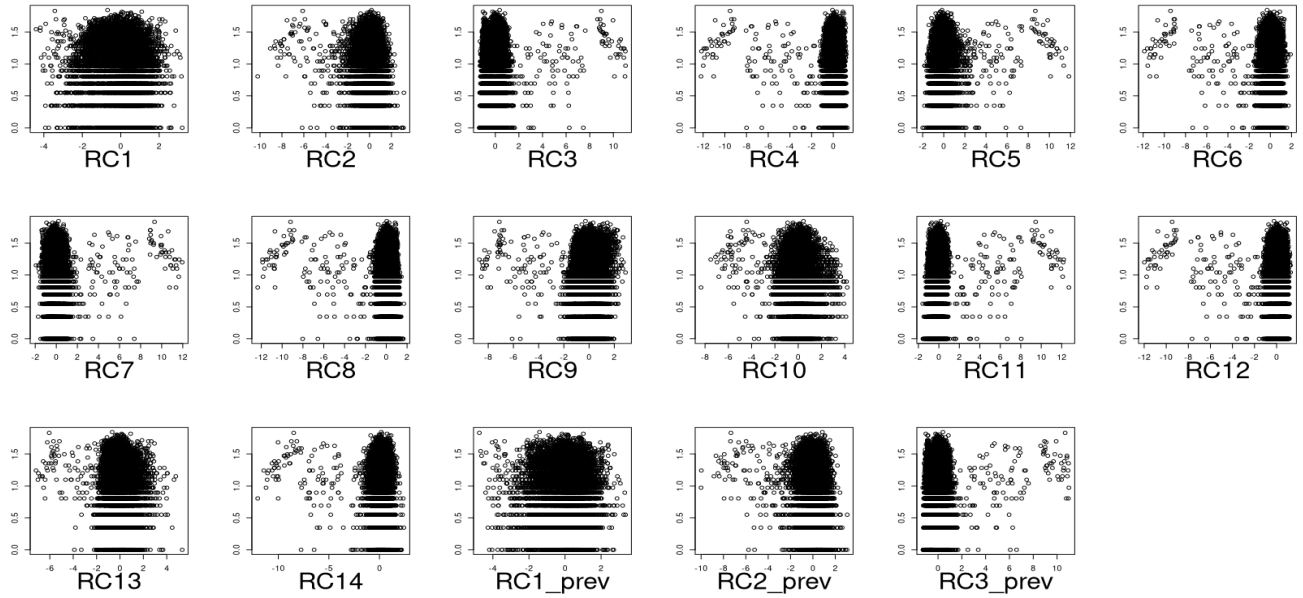
Appendix B: Model Assessment

Poisson Regression Residuals



Residuals vs fitted plots show some signs of heteroscedasticity, as well as a non-linear relationship, which challenges our model assumptions. However, we were unable to fully resolve either of these issues with transformations or addition of polynomial terms.

Log response vs Predictors:



We observe that our assumption of linearity through the link function (Log) is not entirely valid, and that each predictor (PC) has ‘lumps’ of data, especially around 0.