Impacts of Land Use on Water Quality in Minnesota

https://github.com/lhr12/HDA_Project

Keith Bollt, Jake Greif, Felipe Raby-Amadori, & Lindsay Roth

<Arrow brackets are used for annotating the RMarkdown files. Text within these brackets should not appear in the final version of the PDF document>

<**General Guidelines**> <1. Write in scientific style> <2. Global options for R chunks should be set so that only relevant output is displayed> <3. Make sure your final knitted PDF looks professional. Format tables appropriately, size figures appropriately, make sure bulleted and numbered lists appear as such, avoid awkwardly placed page breaks, etc.>

1 Rationale and Research Questions

<Write 1-2 paragraph(s) detailing the rationale for your study. This should include both the context of the topic as well as a rationale for your choice of dataset (reason for location, variables, etc.) A few citations should be included to give context for your topic. You may choose to configure autoreferencing for your citations or add these manually.>

- Land use has a large impact on nutrient runoff into streams, lakes, and other water bodies
- Minnesota has wide variety of land uses. Inlcudes large urban centers, natural lands, and agricultural area.
- Nutrient management has been a challenge for states in the effort to control harmful algal blooms and coastal dead zones.
- Understanding the causes of nutrient problems will better inform management strategies.

Research questions:

1. What are the predictors of nutrients based on land use in watersheds in the state of Minnesota?
2. How do you characterize seasonal variation between the predictors of nutrients?

Goals:

- Determine how land use, watershed size, and ecoregion explain variation in nutrient loading indicators.
- Discern whether there a seasonal trends in nutrient loading indicators based on land use types, watershed size, and ecoregion.
- Provide insight to inform decisions about nutrient managment practices based on land use types, watershed size, and ecoregion.

<At the end of your rationale, introduce a numbered list of your questions (or an overarching question and sub-questions). Each question should be accompanied by one or more working hypotheses, inserted beneath each question.>

Research questions:

1. What are the predictors of nutrients based on land use in watersheds in the state of Minnesota?

   Hypothesis:

2. How do you characterize seasonal variation between the predictors of nutrients?

   Hypothesis:

2 Dataset Information

<Provide information on how the dataset for this analysis were collected, the data contained in the dataset, and any important pieces of information that are relevant to your analyses. This section should contain much of same information as the metadata file for the dataset but formatted in a way that is more narrative.>

The data used in this analysis include data from the Lake Multi-Scaled Geospatial and Temporal Database (LAGOSNE) and the EPA ecoregion spatial datasets.

LAGOSNE is a collection of several data modules that contain information on lakes in the northern United States. The modules contain data from thousands of lakes in 17 states in the northeastern and midwestern United States, from Missouri to Maine. The dataset includes a complete list of all lakes bigger than 4 hectacres in the 17 state area, and water quality data on a large number of lakes, spanning every state.

Ecoregions are used by planning managers to understand the type of land use that occurs in different regions of the United States. There are different levels of ecoregions. Level 1 divides North America into 15 ecological regions, while Level IV offers fine ecological resolution for each state. This data was published by the U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL). For the purposes of our project, we selected Level III ecoregions because they appear to offer a descriptive narrative of the land use patterns of Minnesota without making a 'distinction without a difference'.

<Describe how your team wrangled your dataset in a format similar to a methods section of a journal article.>

******Not included in original draft*********

<Add a table that summarizes your data structure (variables, units, ranges and/or central tendencies, data source if multiple are used, etc.). This table can be made in markdown text or inserted as a `kable` function in an R chunk. If the latter, do not include the code used to generate your table.>

Column.Name	Description	Units	Variable.Type
chla	Chlorophyll a	mg/L	Dependent
secchi	Secchi depth	m	Dependent
IntenseUrban.pct	Percent med. and high intensity urban land cover	%	Independent
OpenUrban.pct	Percent developed	%	Independent
Barren.pct	Percent areas of bedrock and desert pavement land cover	%	Independent
Forest.pct	Percent deciduous	%	Independent
GrassShrub.pct	Percent areas dominated by shrubs	%	Independent
Wetland.pct	Percent wetlands land cover	%	Independent
Pasture.pct	Percent areas of grasses	%	Independent
RowCrop.pct	Percent cultivated crops land cover	%	Independent
LakeIWS.Ratio	Lake surface area to watershed area ratio	N/A	Independent

Column.Name	Description	Units	Variable.Type
Season	Early/Prime/Late	N/A	Independent
Ecoregion	Level II Ecoregions	N/A	Independent

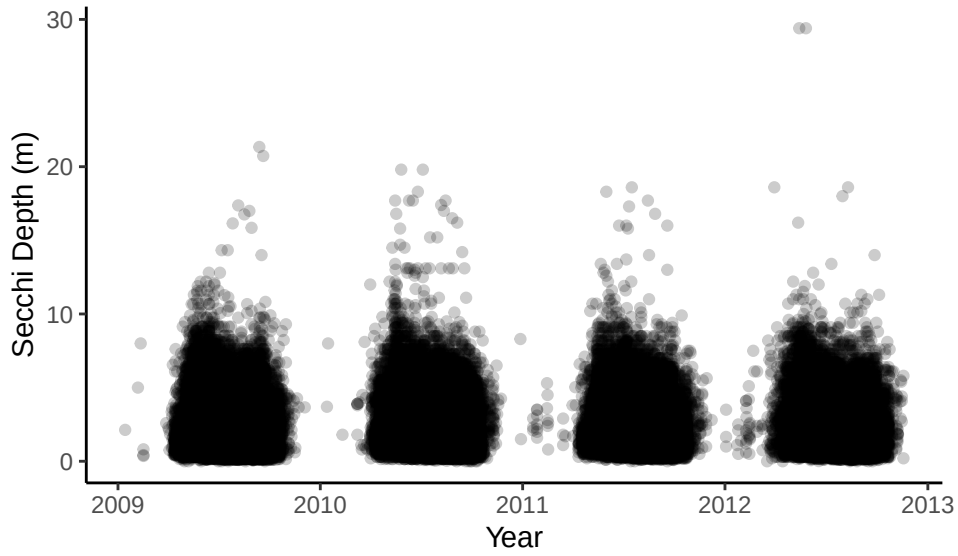


Figure 1: This plot provides a sense of how secchi depth measurements are distributed between and within years, as well as the general range of secchi depth measurements. Sampling occurs during the middle portion of each year, and ceases during the winter season.

3 Exploratory Analysis

<Insert exploratory visualizations of your dataset. This may include, but is not limited to, graphs illustrating the distributions of variables of interest and/or maps of the spatial context of your dataset. Format your R chunks so that graphs are displayed but code is not displayed. Accompany these graphs with text sections that describe the visualizations and provide context for further analyses.>

<Each figure should be accompanied by a caption, and each figure should be referenced within the text>

*******code below is copied directly from originial draft, it needs to be edited to fulfill the above requirements**********

Year	Count
2009	20849
2010	19829
2011	18025
2012	16924

Year	Count
2009	8996
2010	7576
2011	6989
2012	6206


```
## Reading layer `NA_CEC_Eco_Level2' from data source `C:\Users\Felipe\OneDrive - Duke U
## Simple feature collection with 2261 features and 8 fields
## geometry type:  POLYGON
```

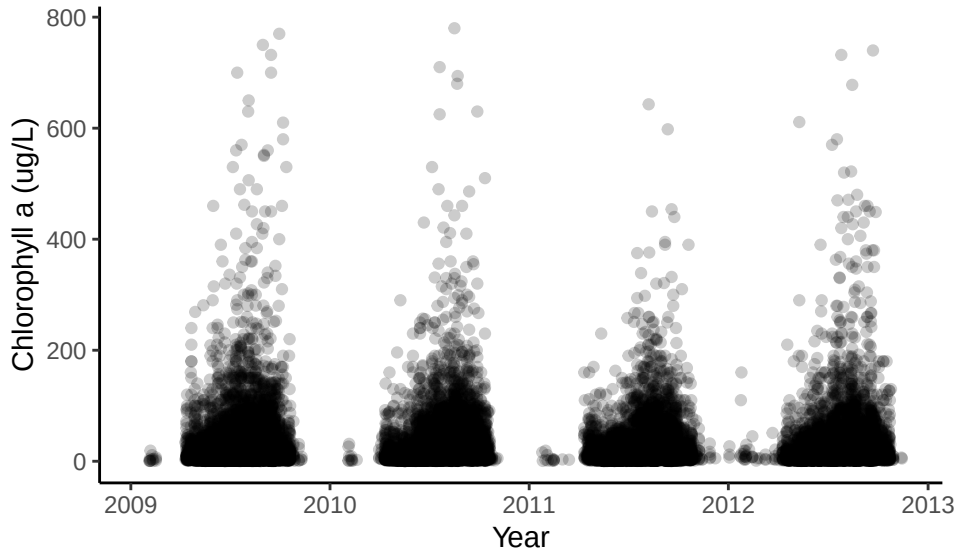


Figure 2: Chlorophyll a vs. Time. This plot provides a sense of how chlorophyll a measurements are distributed between and within years, as well as the general range of chlorophyll a measurements. Sampling occurs during the middle portion of each year, and ceases during the winter season.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.080   0.900   1.680   1.898   2.590   7.500
```

Figures XX and YY explore the trends in secchi depth and chlorophyll a by the percentage of each land use type in the late season. The late season was chosen for exploration because it had to lowest and highest mean secchi depth and mean chlorophyll a, respectively.

FRA Comment: Next plot main purpose is to check is we have enough data in each ecoregion.

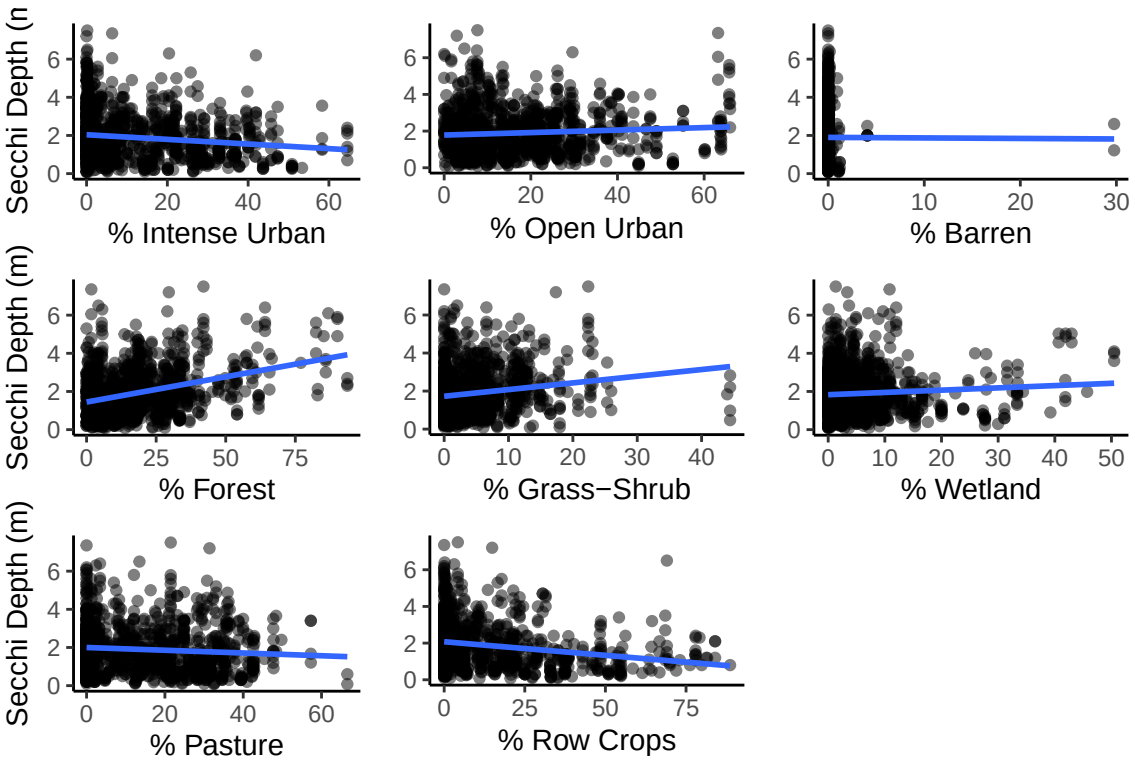


Figure 3: These plots shows that there are trends in secchi depth relative to the given land uses during the 'Late Season'. The presence of trends supports the inclusion of the above land uses in the analysis. Note that the x-axes are on different scales.

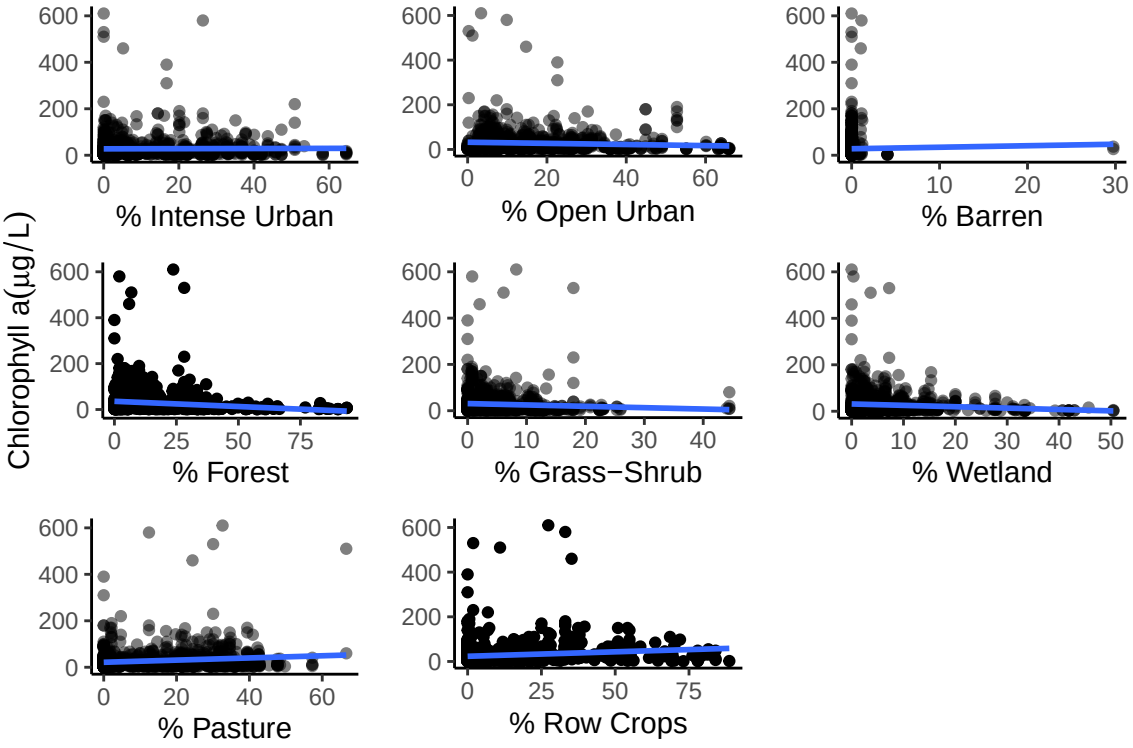


Figure 4: These plots shows that there are trends in chlorophyll a concentrations relative to the given land uses during the 'Late Season'. The presence of trends supports the inclusion of the above land uses in the analysis. Note that the x-axes are on different scales.

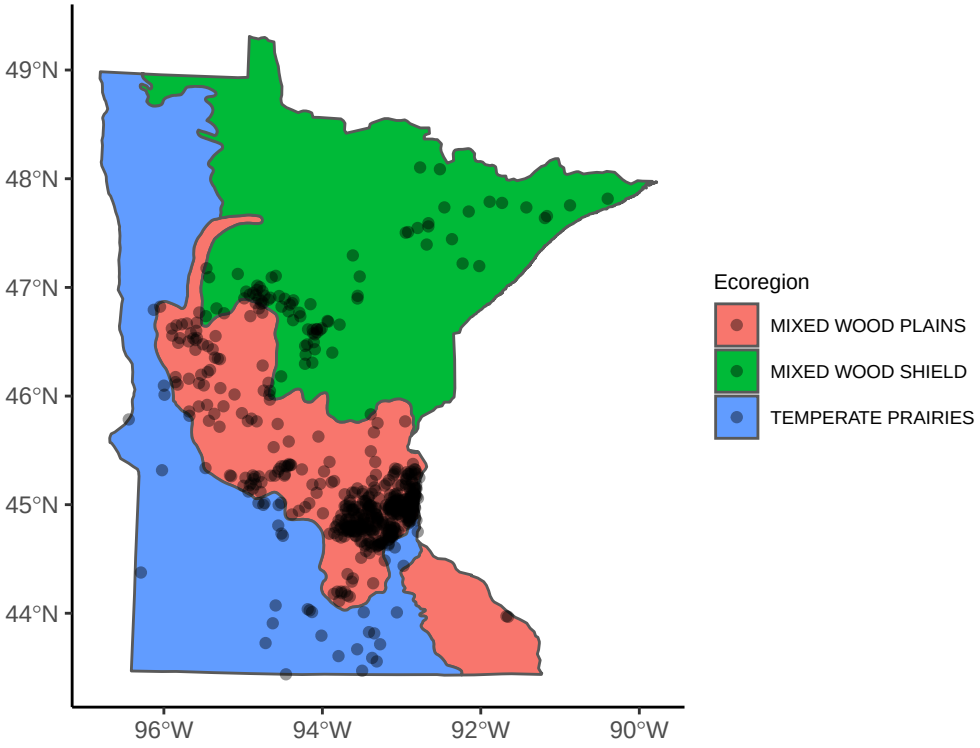
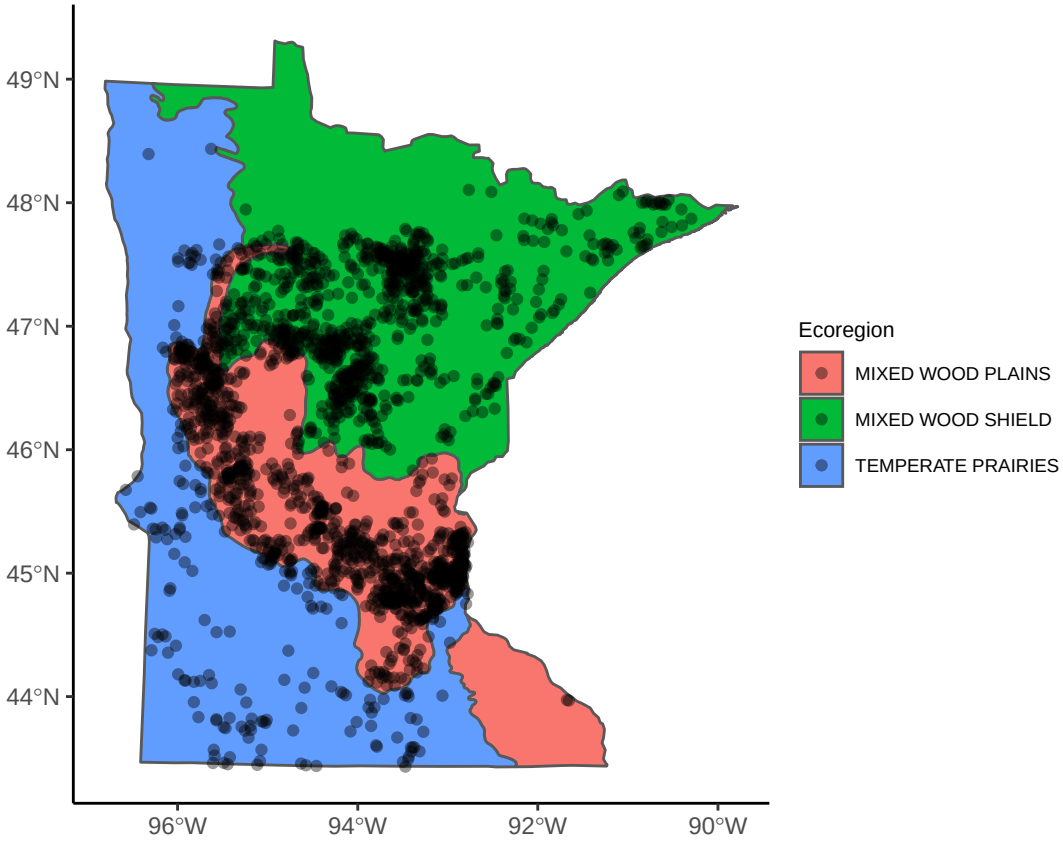
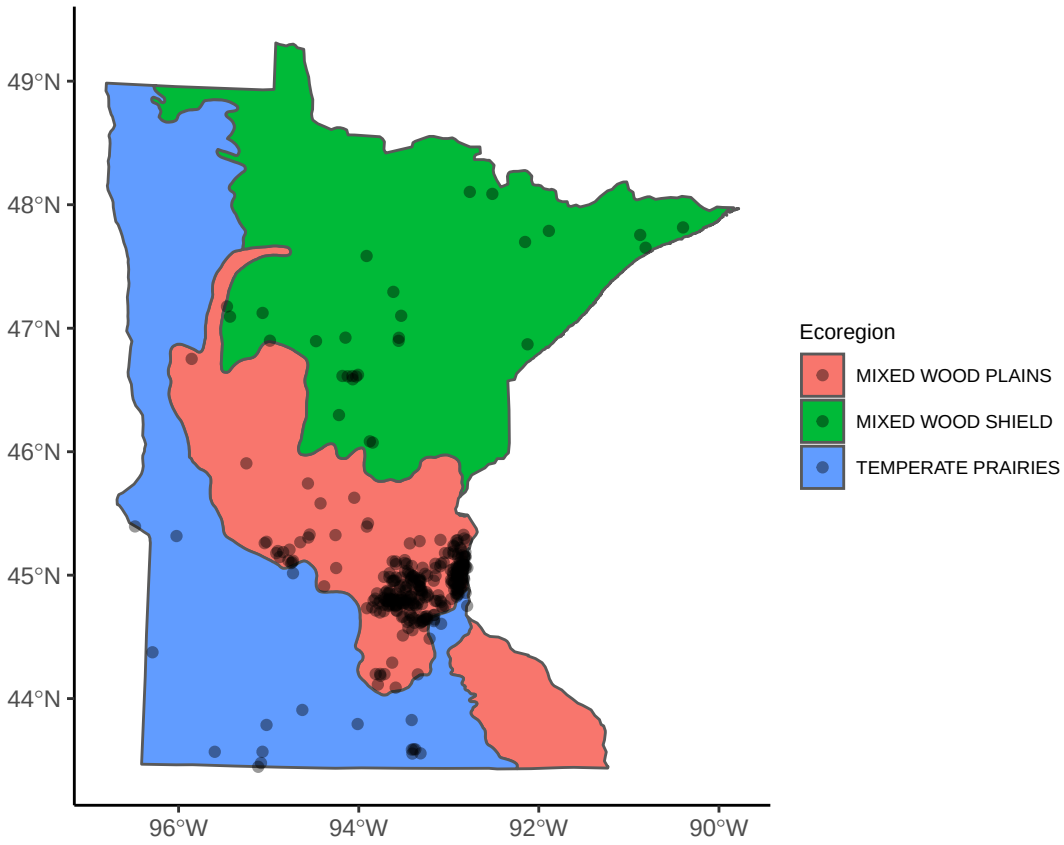


Figure 5: These plots shows that there are trends in chlorophyll a concentrations relative to the given land uses during the 'Late Season'. The presence of trends supports the inclusion of the above land uses in the analysis. Note that the x-axes are on different scales.


49°N
48°N
47°N
46°N
45°N
44°N
96°W
94°W
92°W
90°W
Ecoregion
MIXED WOOD PLAINS
MIXED WOOD SHIELD
TEMPERATE PRAIRIES



FRA Comment: The next plot where the one we did for the proposal. We have to see if we want them

4 Analysis

<Insert visualizations and text describing your main analyses. Format your R chunks so that graphs are displayed but code and other output is not displayed. Instead, describe the results of any statistical tests in the main text (e.g., "Variable x was significantly different among y groups (ANOVA; df = 300, F = 5.55, p < 0.0001)"). Each paragraph, accompanied by one or more visualizations, should describe the major findings and how they relate to the question and hypotheses. Divide this section into subsections, one for each research question.>

<Each figure should be accompanied by a caption, and each figure should be referenced within the text>

- First we will create correlation plots in order to eliminate variables with a correlation greater than 0.8.
- Then we will run Shapiro-Wilkes tests to determine normality and the need for possible data transformations.
- After determining the distributions of the data, then we will generate mixed effect linear models with chlorophyll a and secchi depth as response variables, land use and watershed size as fixed effects, and ecoregion as a random effect.

Final figures will include:

- 6 maps of the state, each showing the relationship between land use and both response variables. Ecoregion will be included as a base layer for each map.
- Scatter plots showing the strongest relationships between land use and the response variables.
- Table showing results of linear model.

4.1 Question 1:

4.2 Question 2:

5 Summary and Conclusions

<Summarize your major findings from your analyses in a few paragraphs. What conclusions do you draw from your findings? Relate your findings back to the original research questions and rationale.>

6 References