Statistics 520, Fall 2025

Take Home Exam

Of interest to limnologists (scientists who study water chemistry and water quality) are the broad effects of land use, geology and landforms on the relations between nutrients, primary productivity, and other water characteristics in lakes and reservoirs. Geographic regions with similar landforms, geology, and evolutionary history (primarily glaciation) are called *physiogeographic regions*. Physiogeographic regions are also related to land use patterns, in particular agricultural versus non-agricultural uses. Our concern in this exam will be to examine the relations between several key water chemistry and water quality variables in the state of Missouri, USA, and determine whether those relations are affected by physiogeographic region.

1 Limnological Data

The data available to us consist of measurements made in 135 reservoirs in Missouri. Each reservoir was sampled three to four times a year, and over a period of 4 to 21 years between 1978 and 2002. A lake-level value for each limnological variable was constructed as detailed in Jones *et al.*(2004). Although a larger number of water chemistry and water quality variables were measured, our concern will be with three particular variables, Total Nitrogen, Total Chlorophyll, and Secchi Depth. Total Nitrogen (TN) in mg/L is one of the primary plant nutrients, Total Chlorophyll (Chl) in μg/L is an measure of algal biomass and, hence, indicates how productive a lake is. Secchi Depth in meters is a measure of water clarity. The data are available on the course web page in the Data module under the file name `takehomedata_2025.txt`. One observation with an extreme value of TN was excluded from this file, leaving a total of 134 reservoirs in our data set.

2 Physiogeographic Regions and Land Use in Missouri

There are generally considered to be about four or five physiogeographic regions in Missouri, which are sometimes further subdivided into smaller divisions. We will use a system that defines only broad regions, giving four areas known as the *Disseted Till Plains*, the *Osage Plains*, the *Ozarks* (including the Springfield Plateau), and the *Southeastern Lowlands* as depicted in Figure 1. We will not have data from any lakes in the Southeastern Lowlands and will combine the two regions of Plains, to arrive at only two physiogeographic regions of interest, Plains, and Ozarks. For

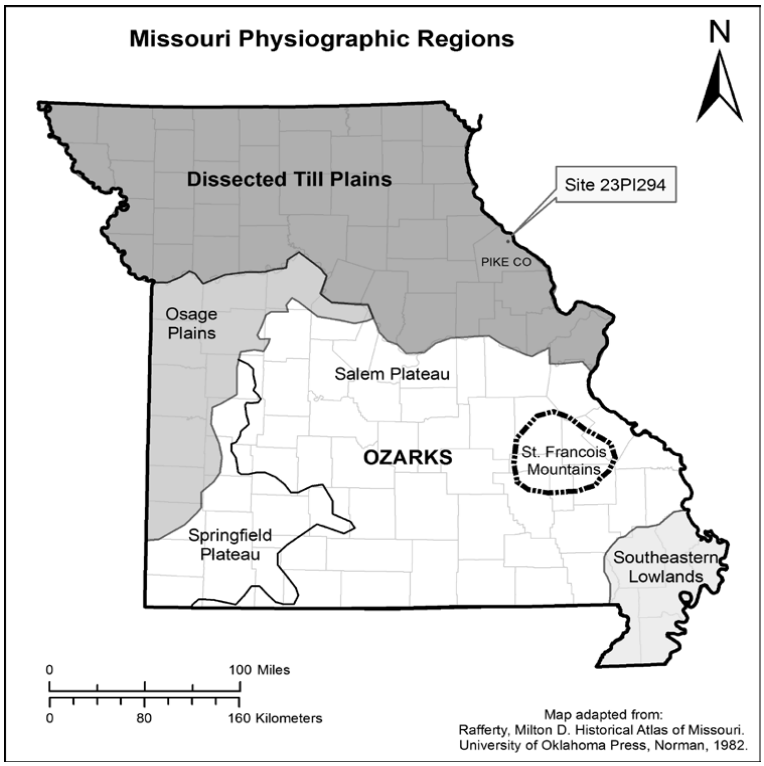

Figure 1: Physiogeogrphic regions in Missouri.

each of the 134 reservoirs for which we have water chemistry data, the proportion

of the watershed in land use categories of *crop*, *forest*, *grassland*, and *urban* was recorded from a GIS data base (see Jones *et al.*, 2004). The median proportions of land use in these categories for each region are given in Table 1. Although the

Land Use	Plains	Ozarks
Crop	0.263	0.046
Forest	0.146	0.574
Grassland	0.364	0.263
Urban	0.000	0.000

Table 1: Median proportion of watersheds in various land use categories for two physiogeographic regions in Missouri.

median proportion of watershed land use in the category urban was zero for both regions, there were 9 watersheds in the Plains region with urban land use greater than 0.50 (primarily the Kansas City and St. Louis areas). There were no such watersheds in the Ozarks region, although there were two with proportion of urban land use between 0.20 and 0.26. It is perhaps difficult to get a true picture of how land use varies between regions from Table 1 alone. To obtain a fuller picture of the distributions of land use across watersheds in the two regions, Figure 2 contains side-by-side boxplots of land use categories for the Plains and Ozarks regions. The plots of this figure show rather dramatic differences in the proportions of watersheds dedicated to crop and forest land uses, with the Plains region generally having much more land devoted to crop and the Ozarks region having much more land devoted to forest. There might be some *a priori* suspicion that this difference in overall land use between the regions could be a potential cause of differences in limnological characteristics. This is because one of the principle crops grown in northern Missouri is corn, and corn is often heavily fertilized with nitrogen. The information on land use and pysiogeographic region is presented for your benefit and possible use in

interpretation of results. It is not anticipated that you will actively use data on land use in your analysis.

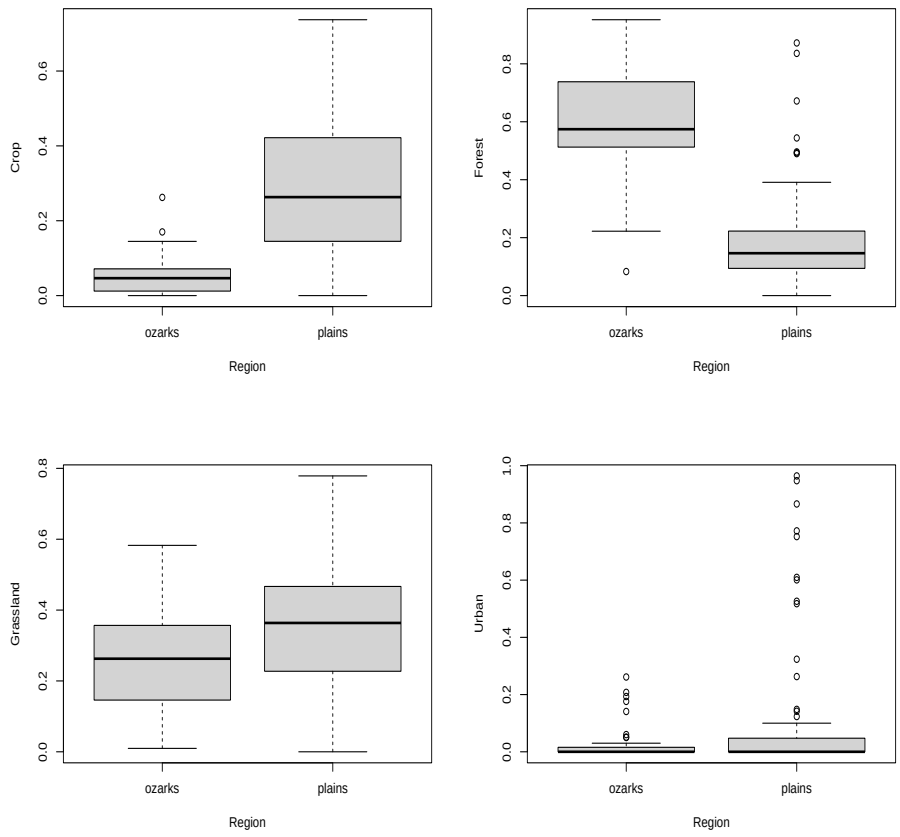

Figure 2: Side-by-side boxplots of proportional land use in the Plains and Ozarks regions.

3 The Exam – Part I

As indicated previously, nitrogen is one of the primary plant nutrients (the other is Phosphorus). Thus, it is of interest to ask how productivity in reservoirs in the Midwest United States (including Missouri) may be related to nitrogen availability

and, given the difference in agricultural land use between regions, whether there is a difference in the chlorophyll-nitrogen relation between regions. In Part I of the exam you are asked to develop what you believe is a suitable regression model to relate Chlorophyll as a response to TN as a covariate. Present evidence that supports your modeling decisions, including whether to approach this task under the framework of a basic generalized linear model or an additive error model, choices of distributions, expectation functions, and/or models for variances, as appropriate. Determine whether you would conclude there are differences between the two regions in this regression, and present at least a basic model assessment to justify the modeling choices you made. See the final section of this exam for some detailed instructions on estimation and comparisons between the regions.

As indicated by Figure 2, higher TN values were observed in the Plains region than in the Ozarks. We expect Chlorophyll to be positively related to TN and, indeed, it can be seen in the summary values of Table 2 that the Plains region contains higher Chlorophyll values than the Ozarks region. And yet, scatterplots reveal that both

Region	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Plains	2.668	11.372	15.629	19.172	24.013	56.203
Ozarks	1.076	2.957	5.857	9.592	13.401	35.984

Table 2: Summary values for Chlorophyll in the Plains and Ozarks regions.

regions contain observations with $0.40 < TN < 1.0$. Let Y_i be a random variable associated with Chlorophyll concentration corresponding to TN x_i in the Plains region, and let Z_i be a random variable associated with Chlorophyll concentration corresponding to TN x_i in the Ozarks region. Based on your fitted model for this part of the exam estimate (point estimate only),

$$Pr(Y_i > Z_i | x_i = 0.70)$$

Given that you have this calculation programmed, compute $Pr(Y_i > Z_i|x_i)$ for a sequence of values $x_i \in \{0.4, 0.41, 0.42, \ldots, 1.0\}$ and produce a graph with TN concentration on the horizontal axis and the estimated probabilities on the vertical axis.

Given all of the pieces you have prepared for this part of the Exam, state whatever conclusions you have reached about the relation between Chlorophyll and TN within these two regions.

4 The Exam – Part II

Secchi depth is measured by taking a disc painted in black-white-black-white quarters, and lowering it into the water. The depth at which the disc just disappears from view is called the Secchi depth, and is a measure of how clear the water is. Chlorophyll in the water column of a lake or reservoir adds to overall turbidity which, in turn decreases water clarity so we might expect a negative relation between chlorophyll concentration and Secchi depth. In Part II of the exam you are asked to develop a regression model to relate Secchi depth as a response to Chlorophyll as a covariate. As with Part I, present evidence for each decision you make at each step of your development, a determination of potential differences between the two regions, and basic model assessment. Again, see the final section of this exam for some additional information.

5 The Exam – Part III

In Part III of the exam you are asked to use full maximum likelihood estimation with the model from Part II, allowing you to conduct a likelihood ratio test for a difference between the Plains and Ozarks regions in the relation between Secchi depth and Chlorophyll. This will give us a definitive answer to the question of

whether the regions differ in the relation between Chlorophyll and Secchi depth, which we will have to assess by some less formal methods in Part II of the exam.

6 Some Details

1. I anticipate this exam being completed using non-Bayesian approaches to estimation and inference. For estimation of an additive error model, the first option should be generalized least squares for regression parameters combined with moment-based estimation of σ^2. If you use a model that has a parameter other than the regression parameters in the variance model, try to choose a fixed value during model formulation. For a basic generalized linear model, the first option should be maximum likelihood for regression parameters combined with moment-based estimation of the dispersion parameter (if there is one). As indicated by the description of Part III of the exam, it is possible with either generalized linear models or additive error models to conduct full simultaneous maximum likelihood estimation of all model parameters, which then broadens the scope of inferential procedures available to us to include, in particular, likelihood ratio tests of nested models.

2. You may try more than one model with the same regression problem, such as a generalized linear model with several random model components. If this is the case, see if you can make a case than one choice is preferred to another possibility. Recall, however, that all data contain a finite amount of information, and sometimes there really are several models that appear to be good choices.

3. Modeling curves that are decreasing and sort of L-shaped can be a challenge. Some possibilities include,

- Reciprocal link functions such as $g(\mu_i) = 1/\mu_i^{1/2}$, $g(\mu_i) = 1/\mu_i$, or $g(\mu_i) = 1/\mu_i^2$. Similarly, for additive error models, $g(x_i, \alpha) = 1/(x_i + \alpha)$ or $g(x_i, \boldsymbol{\beta}) = 1/(\beta_0 + \beta_1 x_i)$.
- Exponential models such as glms with link function $g(\mu_i) = \log(\mu_i)$ or, for additive error models, $g(x_i, \beta) = \exp(-\beta x_i)$ or $g(x_i, \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_i)$.
- Specialty functions in additive error models, such as one based on the Parteo distribution, $g(x_i, \alpha) = 1 - 1/x_i^{\alpha}$ or what is known as a Freundlich model, $g(x_i, \alpha, \beta) = \alpha x_i^{\beta}$.

4. Our basic approach to the comparison of groups, whatever the model within each group is, has been to find a formulation that appears reasonable for both groups and then estimation of a full model with separate parameters for the groups and a reduced model with common parameters for the groups, followed by a likelihood ratio test. This will, in fact, be the approach taken in Part III of the exam. But, if you adhere to the prescriptions of Part I and Part II, a likelihood ratio test will not be available to you. Thus, we must rely on other, perhaps less formal, methods of comparison between regions. Consider what quantities are available to you. For either generalized linear or additive error models we have estimated asymptotic covariance matrices that can be used in the construction of confidence intervals and point-wise confidence bands for regression functions or systematic model components. For generalized linear models we also have the possibility that situations may differ in the random model component, which leads to perhaps even greater difficulties for model comparison, at least through the use of formal inferential procedures. Consideration of the following points might help you as you contemplate comparison of regressions between regions in Part I and Part II of the exam.

   (a) Given the procedures available to use at this point in our study, compar-

ison of regressions that differ in random model components or distributional form must be based on diagnostics such as residual plots.

(b) We know that an interval for the difference in two parameters is a more powerful way to examine for a meaningful difference than is determining whether separate intervals overlap.

(c) A comparison of regression functions can be made by examining the estimated functions with confidence bands. Simultaneous bands would be preferred for this comparison, but point-wise bands provide another possibility that is somewhat more conservative.

(d) To assist with visual examination of fitted curves and confidence bands in nonlinear models, it may be beneficial to plot only portions of the overall covariate range, which may then also change the overall scale across which the response function varies.

(e) If two procedures have been used to examine potential differences in regressions and the conclusions do not agree, we are obligated to attempt to determine a reason for the contradictory results.

Reference:

Jones, J.R., Knowlton, M.F., Obrecht, D.V. and Cook, E.A. (2004), Importance of landscape variables and morphology on nutrients in Missouri reservoirs. *Canadian Journal of Fiheries and Aquatic Sciences.* **61**: 1503-1512.