Geospatial Analysis of a Changing Road Network

Will jones Ben Jantzen Yang Shao

2025-11-19

Table of contents

1 Abstract **1**

2 Introduction **2**

3 Methods **2**
3.1 Data Acquisition and Preprocessing 2
3.2 Sampling Scheme and Measurement 3
3.2.1 Terrestrial Points 3
3.2.2 Water Points 3
3.3 Statistical Analysis 4
3.3.1 Permutation Test 6

4 Results **7**

5 Tables **10**

6 Figures **10**
6.1 Abbreviations 14

References **14**

1 Abstract

The 2003 study "How Far to the Nearest Road ?" published in *Frontiers in Ecology and The Environment* studied the road network of the United States (Riitters and Wickham 2003). Specifically, it split the contiguous US into 30x30 m grid cells and classified each as being a

"road" or "non-road" cell, based on if any part of the cell intersects a TIGER/line road feature. Then, a binned classification was used to measure the distance of every cell to its nearest road cell.

Our aim is to expand on this methodology using GIS and statistical methods. Specifically, our objective is to answer the following questions:

1. By randomly sampling points and calculating straight-line (Euclidean) distances, can we measure a more accurate continuous distribution of distance to the nearest road?
2. Using archived Census Bureau TIGER/Line geographic data, can we measure how far the average distance to the nearest road has changed in the contiguous US?
3. What water bodies have experienced the greatest change and become more vulnerable to ecological effects of road construction and operation?

2 Introduction

There are well-documented effects of roads on habitat fragmentation [CITE], plant health [CITE], and species diversity [CITE]. Native stream invertebrate species may be especially at risk. In fact, Gál, et al. found that road crossings had negative effects on the biodiversity of native invertebrates in Hungary, including species richness, abundance, and prevalence of protected species (Gál et al. 2020). Between X and X year, Y miles of roads have been built, an average of X miles of road per year [CITE]. Of course, this period encompasses the time since the last contiguous United States-level measurement of distances to the nearest road in 2003 (Riitters and Wickham 2003). Using GIS methods and a statistical approach based on the energy distance (Szekely, Rizzo, and Székely 2004), we may examine the expansion of the United States road network for future ecological applications. Additionally, through measurement of a paired continuous distribution of distances (2000 and 2024), we may observe the magnitude of change in distances, both at the national scale, and at relevant sub-levels, including watershed, EPA ecoregion, NLCD landcover class, and even individual stream level.

3 Methods

3.1 Data Acquisition and Preprocessing

Six primary spatial datasets were used in our analysis. The 2023 USGS National Land Cover Database (NLCD), containing 11 distinct land cover classes, was used to extract surface type at each of our sampled points ("Annual NLCD Collection 1 Science Products" 2024). Likewise, attribute data for each sampled point was extracted using the EPA's ecoregion level IV map, and watershed membership was extracted using the USGS 3D Hydrography Program (3DHP) data set. 3DHP is a unified lidar-derived data product containing vector data for streams, lakes, ponds, catchments, and other hydrologic features (USGS, 2025).

US Census Bureau's TIGER/Line vector files were used as the road networks in this project. The most recently available 2024 road network data for the conterminous United States was downloaded in bulk using the roads() function from the Python package pygris ("Pygris," n.d.).

Comparable data for the year 2000 was downloaded from the archived Census FTP site (https://www2.census.gov/geo/tiger/tiger2k/), and a multi- step data preprocessing workflow was performed before the data was implemented in the analysis workflow. First, beautifulsoup4 was used to download vector dataset in RT1 format, the historical file format for census geographic data (Richardson 2007). Using GDAL command-line utilities, files were converted to modern shapefile format, then merged into a single dataset using geopandas in python (Jordahl et al. 2020). Finally, feature correction was performed using geometry-snapping and densification processing in QGIS 3.36 "Maidenhead." To rectify data quality issues, namely feature accuracy and spatial mismatch between existing roads in the 2000 and 2024 datasets, the historical vector dataset was densified at regular 25m intervals, then snapped to the nearest matching feature in the current roads dataset. This was done to avoid erroneous measurement differences in distance to the nearest road for each year that may be incorrectly attributed to removal or addition of a new road feature (**?**).

3.2 Sampling Scheme and Measurement

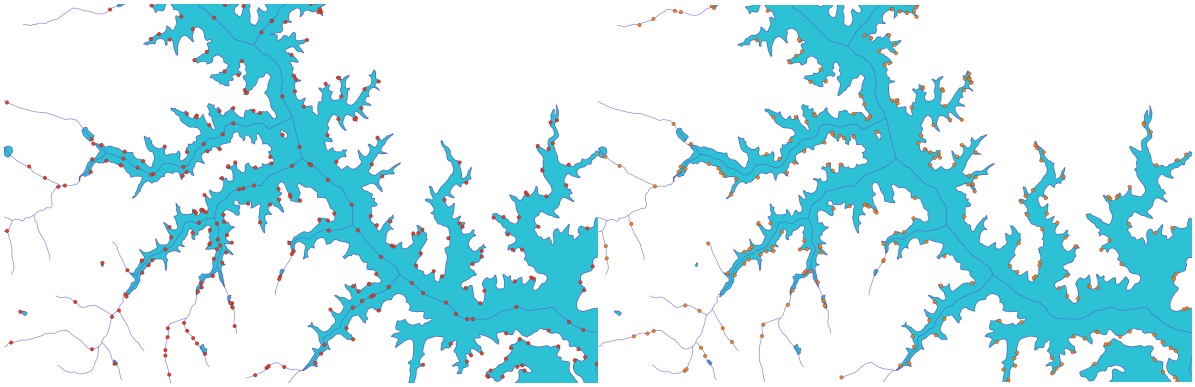
3.2.1 Terrestrial Points

Using the R package sf, 1,000,000 points were randomly sampled across our study area (Virginia, USA). Points were sampled uniformly across space, meaning that everywhere on the continuous surface has equal likelihood of being sampled (Fotheringham and Rogereson 2008). Next, the nearest feature from each road network (2000 and 2024) was identified for each randomly sampled point, and the Euclidean distance between the two geometries was measured. The result was two continuous distributions, each containing a measurement at the same point. This is a case of paired data, analogous to 'before and after' measurements.

3.2.2 Water Points

To perform the corresponding process for water features (lakes, ponds, streams, creeks), a slightly different sampling method was used. It is our goal to randomly select points along linear water features and along the *edges* of waterbodies (polygons). We do not want to sample points within ponds or lakes, because the measurement of interest is the distance between the least central points of each areal feature (i.e., the "shore") and the nearest road feature. This effectively captures where ecological interactions such as runoff into ponds/lakes will occur. To do this, the layers `hydro_3dhp_all_flowline` and `hydro_3dhp_all_waterbody` were extracted from the USGS 3DHP geodatabase, representing flowlines and waterbodies, respectively. Due to the structure of the 3DHP features, `flowlines` are often drawn through

the waterbody that they feed into. To navigate this, and avoid sampling within water bodies, the `sf` package function `st_difference` was used to eliminate these overlapping flowlines from the features that could be sampled on. Lastly, lake and pond boundaries were merged with stream and river features into a single linear feature layer. When using `st_sample`, this ensures that the points are uniformly randomly sampled, meaning that any point in our dataset has equal probability of being sampled.



(a) Sampling on water features before removing overlapping flowlines

(b) Sampling on water features after removing overlaping flowlines

3.3 Statistical Analysis

For the purposes of this class project, 100,000 of the ~1e6 points were chosen randomly for analysis in order to make local processing possible.

To compare the shift in distributions from 2000 to 2024, a permutation test for matched pairs was used (Welch 1990), along with bootstrap resampling (Dixon 2001) to construct a confidence interval for the test statistic of interest, the mean of the differences between the paired points. The mean of the paired differences was chosen as the test statistic for analysis because the data set is dominated by points whose values did not change (i.e., difference is 0).This characteristic of the distribution, along with the fact that the distribution of differences is not symmetric around the mean (see Figure 2), is why we did not use the Wilcoxon Signed Rank test.

We used bootstrapping (Dixon 2001) to construct a sampling distribution and confidence interval for the mean of paired differences. The observed mean of the differences between our paired data is 74.89 meters. After bootstrap resampling, we have a theoretical sampling distribution with a mean of 74.87 meters (95% CI: (73.19, 76.60)) and a standard error of 0.87 meters.

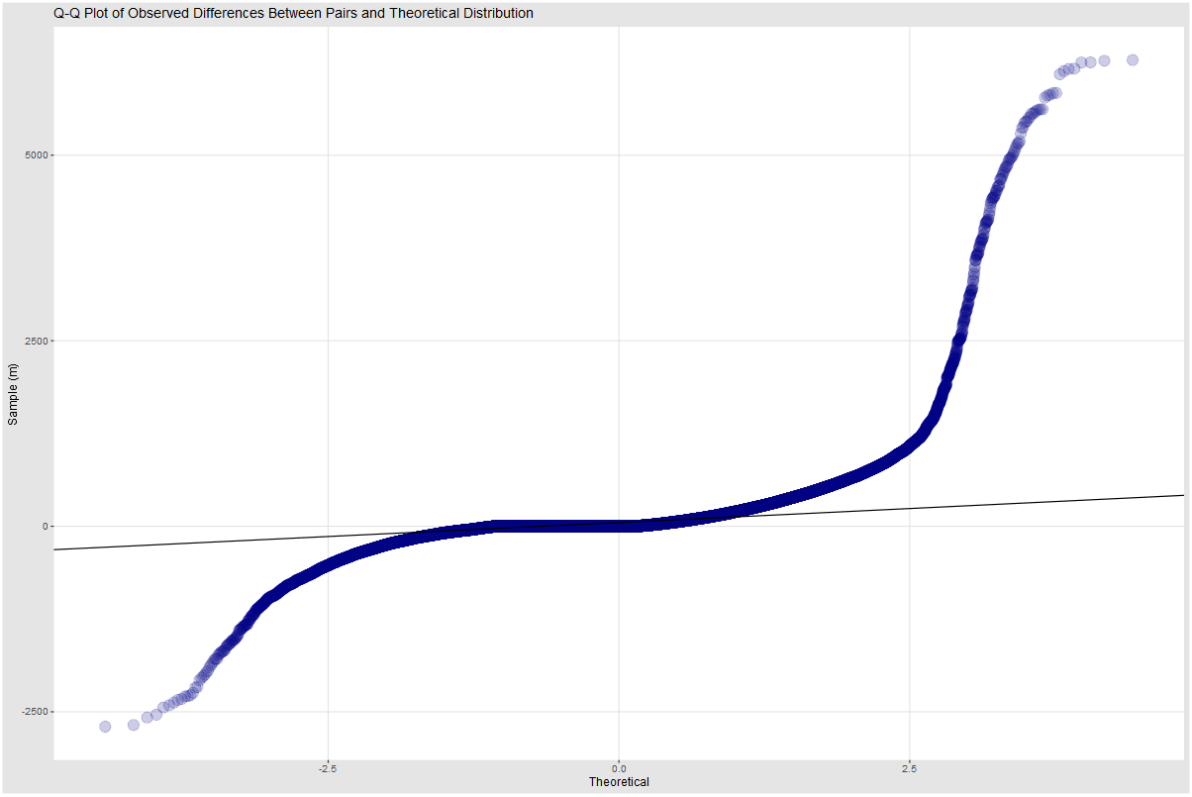
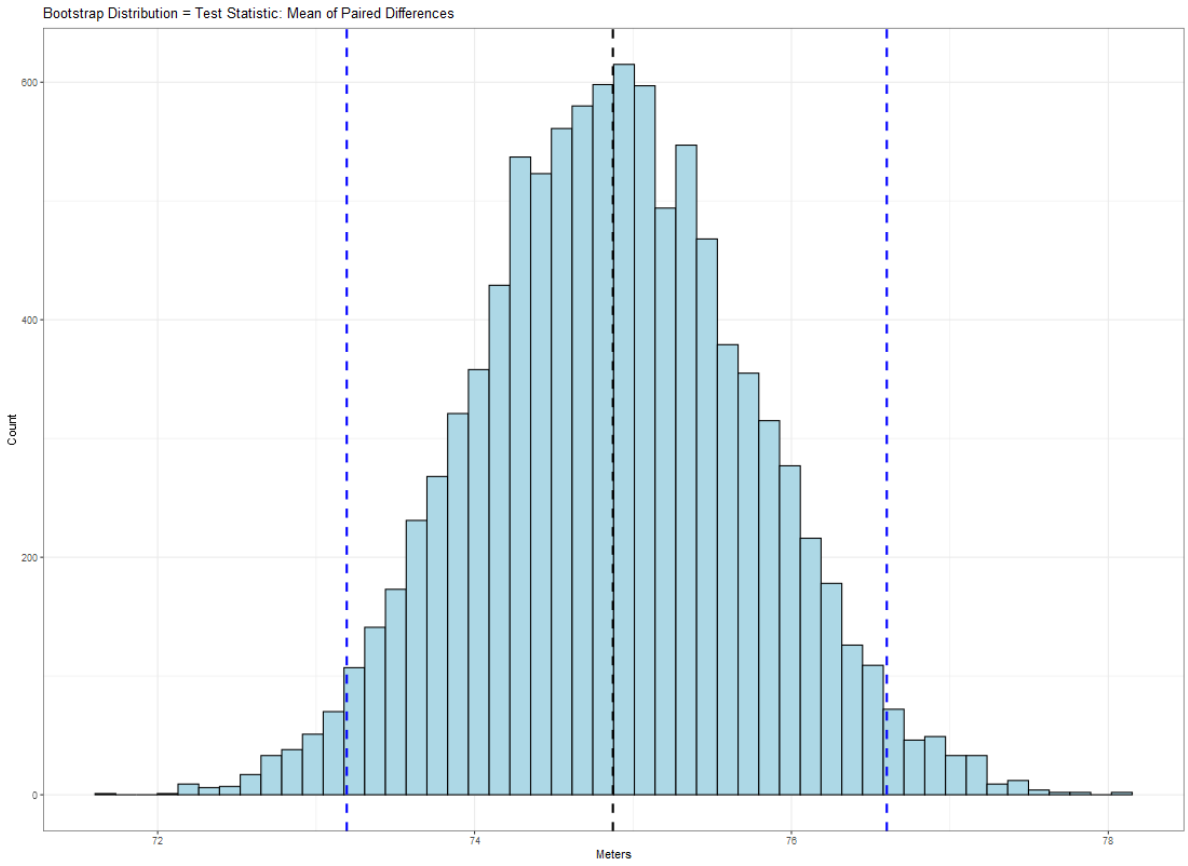

Q-Q Plot of Observed Differences Between Pairs and Theoretical Distribution
5000
2500
0
-2500
Sample (m)
-2.5
0.0
2.5
Theoretical

Figure 2



3.3.1 Permutation Test

Then we performed a permutation test for paired data ("Permutation Test for Matched Pairs Data," n.d.) to test for significance in our results:

```
results = c()
for (i in 1:5000) {
  permutedData = sample(c(1,-1),100000,replace=T)*dat$difference
  results = c(mean(permutedData),results)
}
hist(results,col='gray',main="Permutation Distribution",xlab="Simulated difference")
mean(results)
results <- as.data.frame(results)
```

For the standard case of a permutation test, the null hypothesis is that the two groups of interest come from the *same distribution*, and the alternative hypothesis is that the groups

are not from the same distribution. In the case of the matched-pairs permutation test, our hypotheses are:

$$H_0 : \mu_{\text{diff}} = 0$$

In other words, the difference in the before and after values of points is due to random chance

$$H_a : \mu_{\text{diff}} \neq 0$$

The permutation test computes 5,000 permutations of the 100,000 paired points, where before/after values are switched. At each permutation, the test statistic (the mean of all the paired differences) is computed. This creates a roughly normal distribution of mean values. This works on the assumption that if there truly is no difference in the distributions of the before and after groups. Under the null assumption, the observed test statistic (74.89 m, which we calculated from our original sample), should fall somewhere within the "null" or permutation distribution. To calculate our p-value, we take the proportion of all permutation values that are equal to or more extreme than our test statistic. Figure 3 shows clearly that the test statistic obtained from our sample (red line) is far more extreme than what can be expected if there is no difference in the values of before/after groups.

4 Results

here give summary statistics

i think would be good to include change in distance by land cover classes, even ecoregions?

ridge plots could be good, simple bar chart showing mean change in distance for each land cover type

ladder (?) plot showing change by land cover or ecoregion

"based on our analysis, X km of roads were constructed between 2000 and 2024, resulting in a measured shift of X with MEAN ___ and SD ____. take from bootstrap sampling distribution?

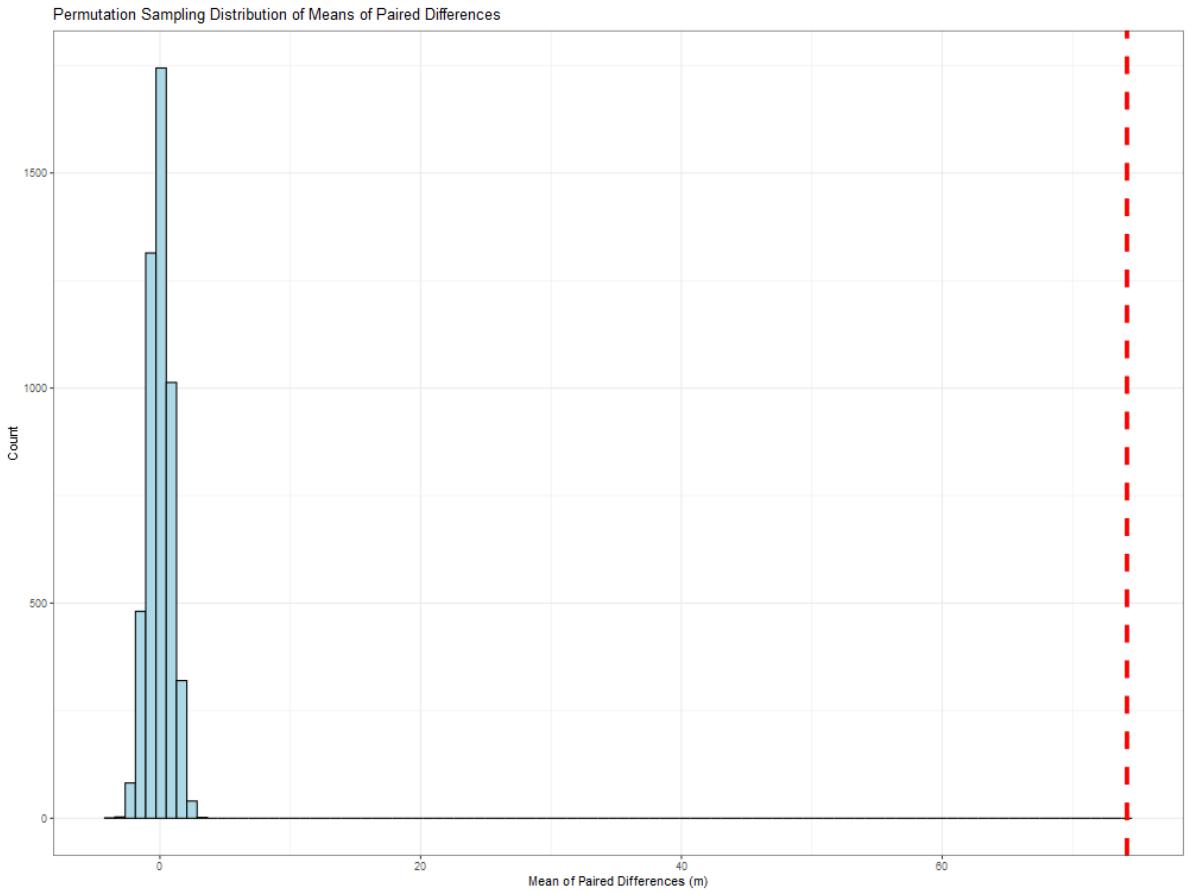


Figure 3: Visualization of permutation test for statistically significant difference in the means of 2000 and 2024 distributions, $p<0.0001$

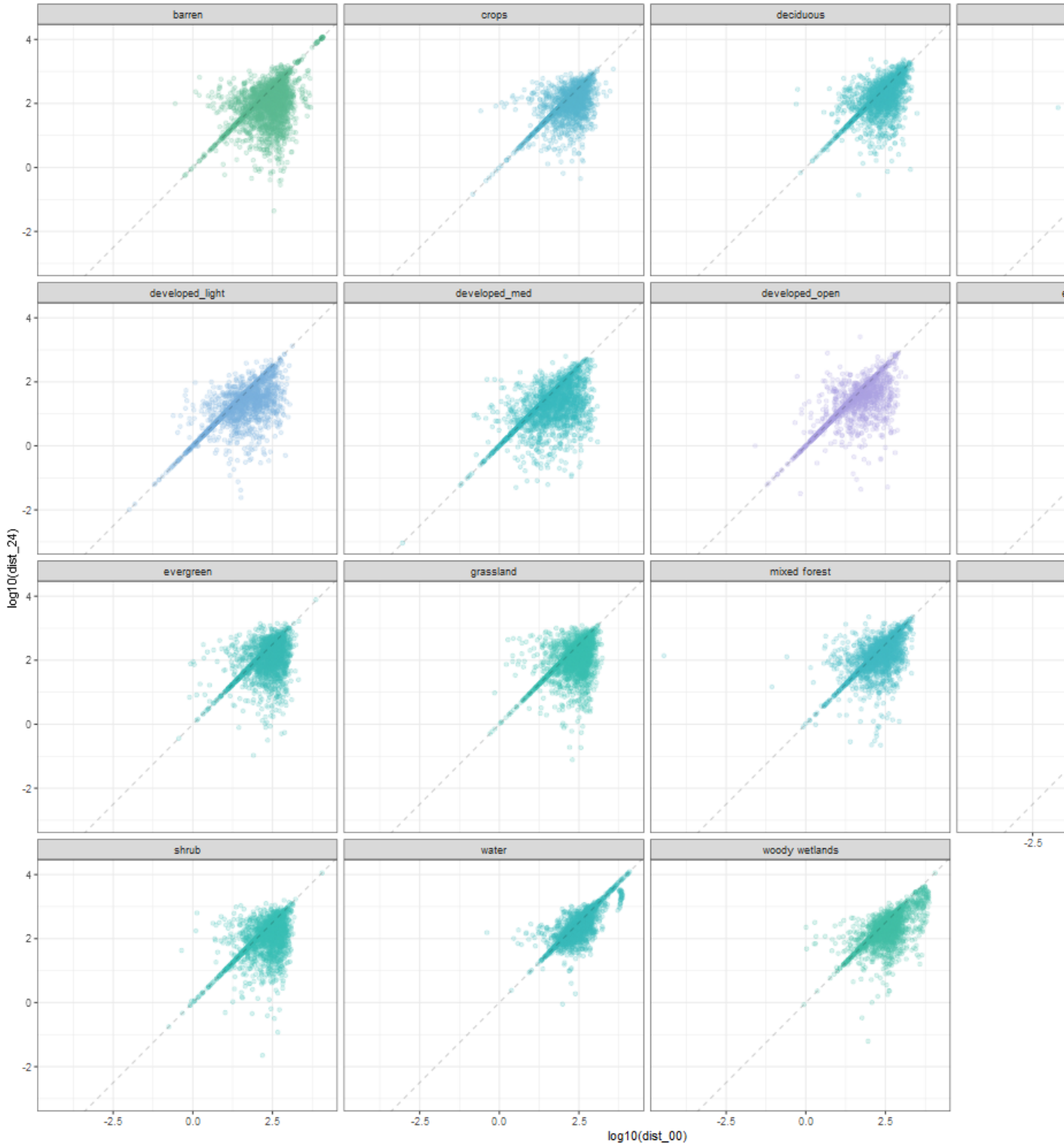


Figure 4: Scatterplots of distance from the nearest road in 2000 (x-axis) vs 2024(y-axis), separated by NLCD class. Color scale representing proportion of total sampled points that became closer to the nearest road.

5 Tables

Summary Statistics (m)							
	Mean	SD	Min	1st quartile	Median	3rd Quartile	Max
2000	289.29	425.65	0.00004	74.80	188.47	376.25	12963.61
2024	214.83	342.06	0.00031	52.93	132.74	270.36	12892.76
Difference	74.46	83.59	−0.00027	21.87	55.73	105.89	70.85

Figure 5: Table showing summary statistics of each distribution, 2000 and 2024, in meters

6 Figures

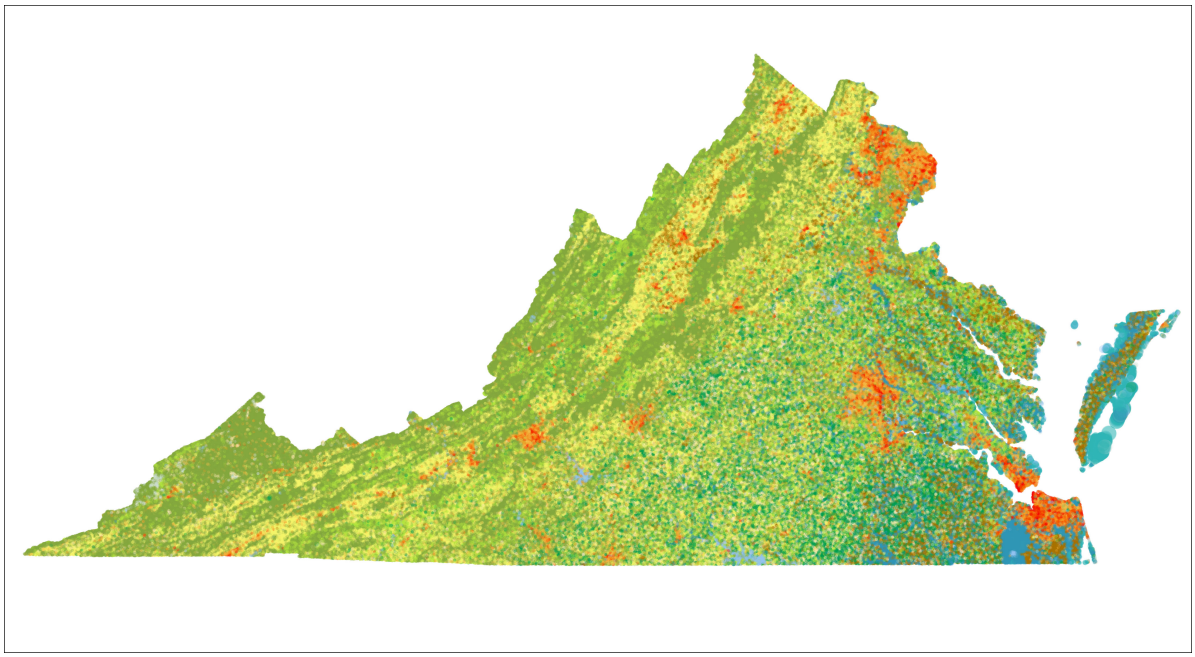


Figure 6: VA Map

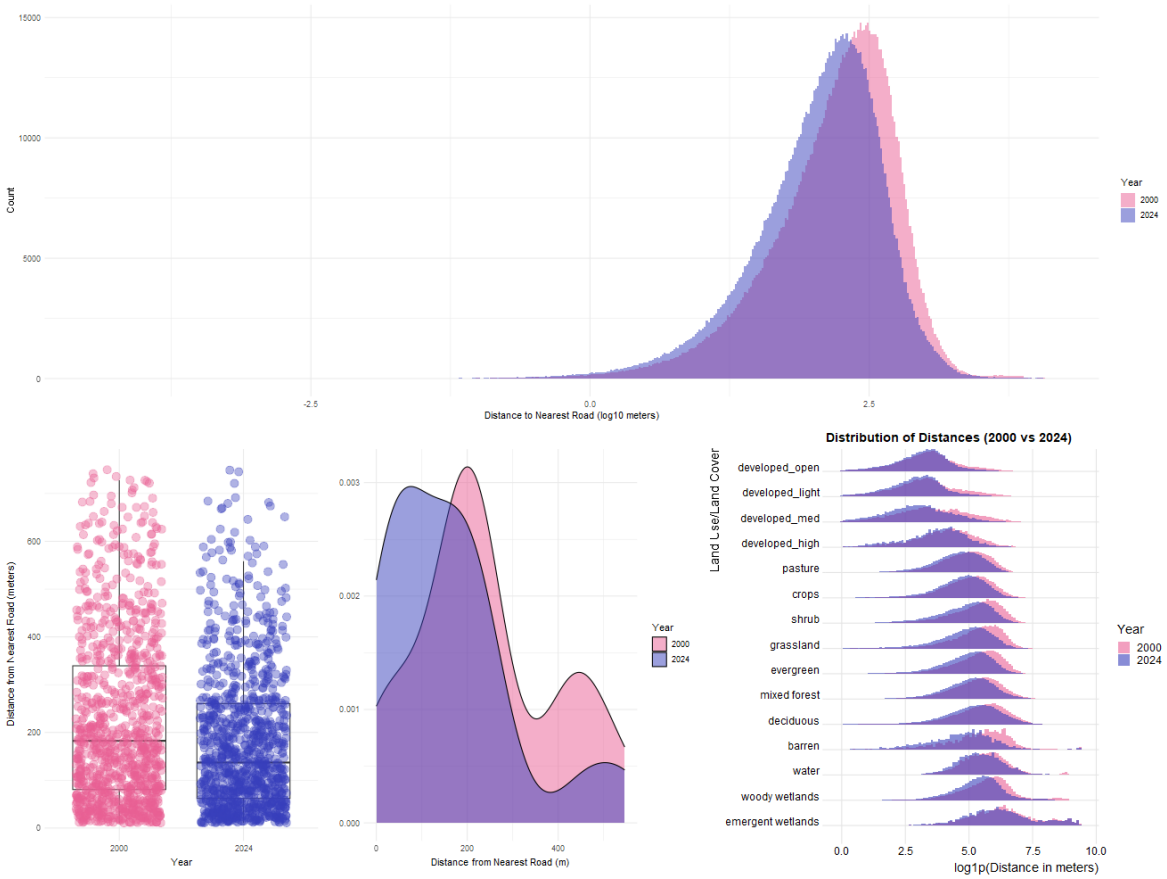


Figure 7: (a). Histograms for two paired distributions, for the year 2000 and 2024, showing a left-shift from the measured values, indicating that the measured points became closer to the nearest road, (b). Boxplot showing the distributions of distance to the nearest road in each year, (c). Density plot for a specific EPA level IV ecozone, Limestone Valleys and Coves, indicating distances to the nearest road, (d). ridgeplot of the measured values for 11 NLCD land use land cover classes

Figure 8: Example of points sampled along water features in Franklin County, VA

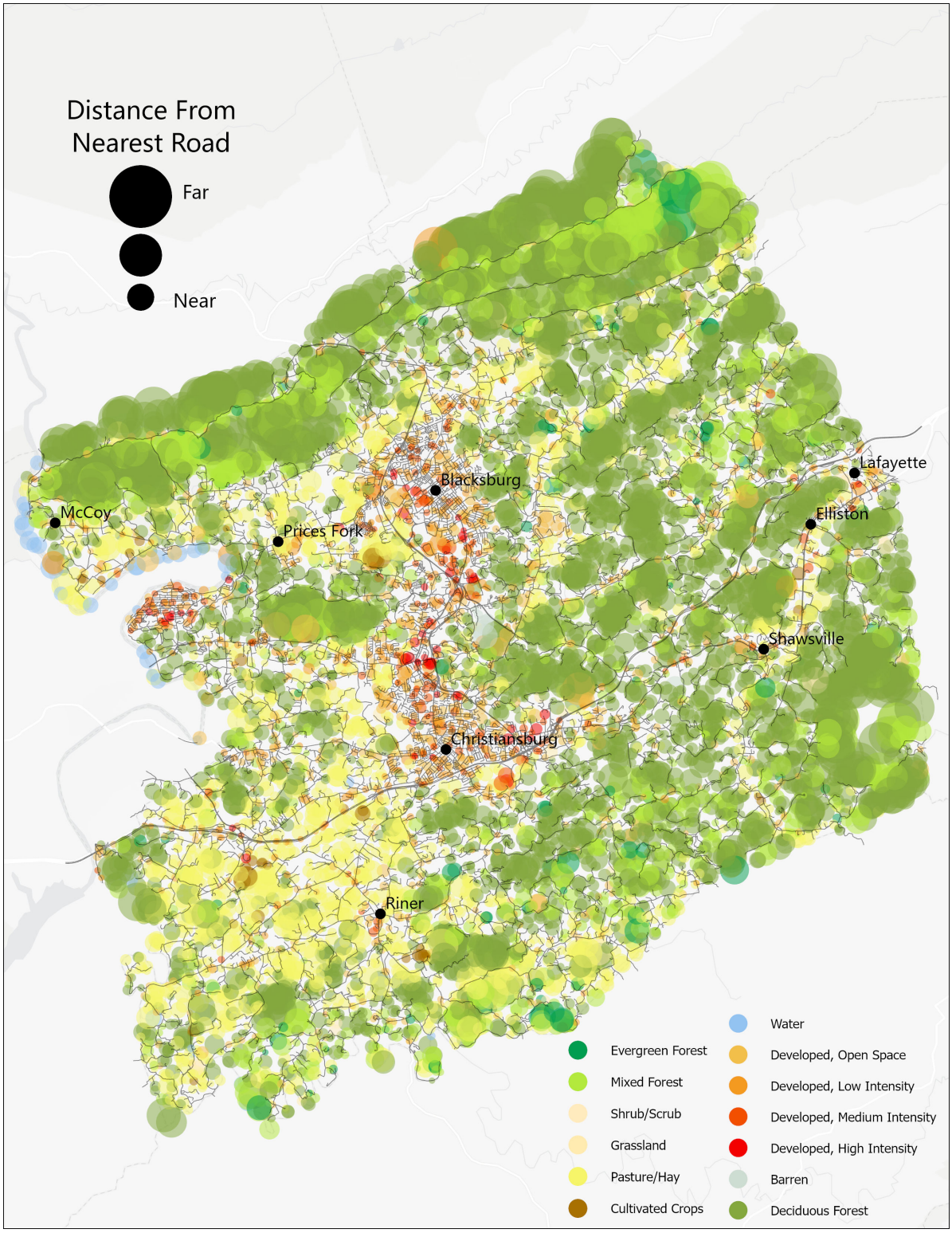


Figure 9: Map of Montgomery County, Virginia. Point color represents land cover classification, and point size represents its distance to the nearest road

6.1 Abbreviations

- **3DHP** 3D Hydrography Program
- **EPA** Environmental Protection Agency
- **GDAL** Geospatial Data Abstraction Library
- **HUC** Hydrologic Unit Code
- **NLCD** National Land Cover Database
- **TIGER** Topologically Integrated Geographic Encoding and Referencing
- **USGS** United States Geoglogical Survey

References

"Annual NLCD Collection 1 Science Products." 2024. US Geological Survey (USGS).

Dixon, Philip M. 2001. "Bootstrap Resampling." *Encyclopedia of Environmetrics*, October. https://doi.org/10.1002/9780470057339.VAB028.

Gál, Blanka, András Weiperth, János Farkas, and Dénes Schmera. 2020. "The Effects of Road Crossings on Stream Macro-Invertebrate Diversity." *Biodiversity and Conservation* 29 (March): 729–45. https://doi.org/10.1007/S10531-019-01907-4.

Jordahl, Kelsey, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, et al. 2020. "Geopandas/Geopandas: V0.8.1." Zenodo. https://doi.org/10.5281/zenodo.3946761.

"Permutation Test for Matched Pairs Data." n.d. https://people.hsc.edu/faculty-staff/blins/classes/spring19/math222/Examples/MatchedPairsPermutationTest.html.

"Pygris." n.d. https://walker-data.com/pygris/.

Richardson, Leonard. 2007. "Beautiful Soup Documentation." *April*.

Riitters, Kurt H, and James D Wickham. 2003. "How Far to the Nearest Road?" *Ecology and the Environment*. Vol. 1.

Szekely, Gabor J, Maria L Rizzo, and Gábor J Székely. 2004. "Testing for Equal Distributions in High Dimension." https://www.researchgate.net/publication/228918499.

Welch, William J. 1990. "Construction of Permutation Tests." *Journal of the American Statistical Association* 85: 693–98. https://doi.org/10.1080/01621459.1990.10474929.