

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	3
3.1 Data Acquisition and Preprocessing	3
3.2 Sampling Scheme and Measurement	4
3.2.1 Terrestrial Points	4
3.2.2 Water Points	4
3.3 Statistical Analysis	5
3.3.1 Permutation Test	7
4 Results	8
5 Tables	10
6 Figures	13
6.1 Abbreviations	15
References	15

1 Abstract

This study investigates the expansion of the road network in the contiguous United States from 2000 to 2024 and its implications for landscape ecology. Using GIS tools and statistical

analyses, we examine changes in the proximity of terrestrial and aquatic features to the nearest road. The analysis employs permutation testing and bootstrap resampling to assess the statistical significance of shifts in road proximity over time. The findings highlight the growing influence of road infrastructure on both terrestrial and aquatic ecosystems, with potential implications for habitat fragmentation, water quality, and species connectivity.

2 Introduction

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 tests. 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 the average change in distance to the nearest road for the contiguous US?
3. **What water bodies have experienced the greatest change and become more vulnerable to ecological effects of road construction and operation?**

The role of road networks related to habitat fragmentation, and its subsequent effects on plant and animal life have been studied extensively ([Saura and Rubio 2010](#); [Forman and Alexander 1998](#); [McGregor, Bender, and Fahrig 2008](#)). This is a global topic of interest in landscape ecology, with studies of road impacts across various ecozones in Sweden ([Karlsson and Mörtberg 2015](#)), landscape ecological risk (LER) related to roads in the Central Himalaya ([Mann et al. 2021](#)), giant anteater mortality in Brazil ([Pinto et al. 2018](#)), and many many more.

Functional connectivity, a foundational concept in landscape ecology refers to the degree to which organisms can perform certain biological processes across their landscape. Such processes include dispersal of seeds, migration related to breeding, and genetic exchange. “By severing habitat connections, roads create risk for animals, especially slow-moving reptiles like the wood turtle,” according to the Conservation Planning in the Hudson River Estuary Watershed project at Cornell University ([“Connectivity Planning | Conservation Planning in the Hudson River Estuary Watershed,” n.d.](#))

Within the scope of animals vulnerable to the impacts of road effects, native stream invertebrate species may be especially at risk. In fact, Gál, et al. found that road crossings over

streams (bridges) had negative effects on the biodiversity of native invertebrates in Hungary, including species richness, abundance, and prevalence of protected species (Gál et al. 2020). Despite this, road expansion has continued at a steady rate in the United States, with little long-term monitoring of the growing impact on landscape biodiversity. Between 1980 and 2011, 183,000 miles of roads have been built, an average of 6,500 miles of road per year. Additionally, 32,264 bridges were constructed between 1992 and 2010, bringing the total as of 2010 to 604,460 (“Office of Highway Policy Information - Policy | Federal Highway Administration,” n.d.). 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 and Rizzo 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. (*Note: for this analysis I am using a permutation test, not the energy distance method).

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 the 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.

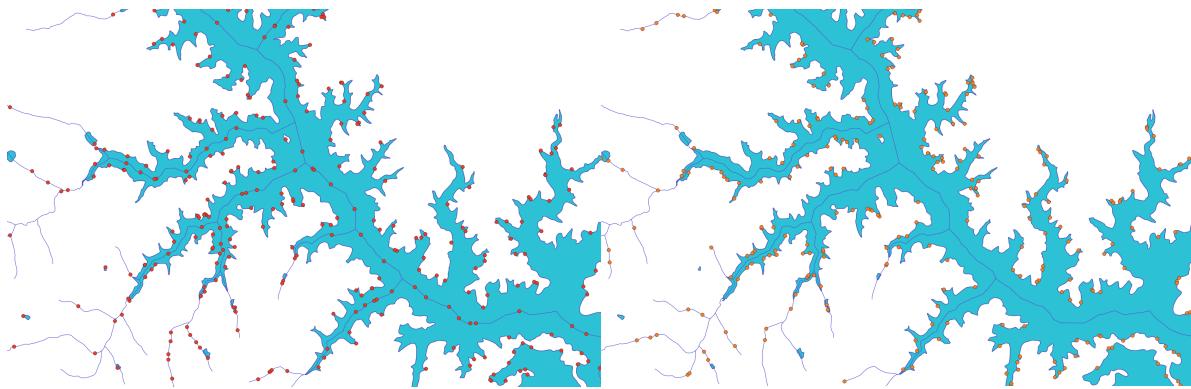
3.2 Sampling Scheme and Measurement

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 Rogerson 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. (*Note: the reason analysis of these water points has not been accomplished is because we have not been able to generate a sufficiently large sample. This is because the geometric operations for a huge set of points are computationally intensive. Our future plan is to do this processing on Virginia Tech’s Advanced Research Computing resources*).



(a) Sampling on water features before removing overlapping flowlines (b) Sampling on water features after removing overlapping flowlines

3.3 Statistical Analysis

For the purposes of this class project, 100,000 of the $\sim 1\text{e}6$ 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 an approximated sampling distribution with a mean of 74.87 meters (95% CI: (73.19, 76.60)) and a standard error of 0.87 meters.

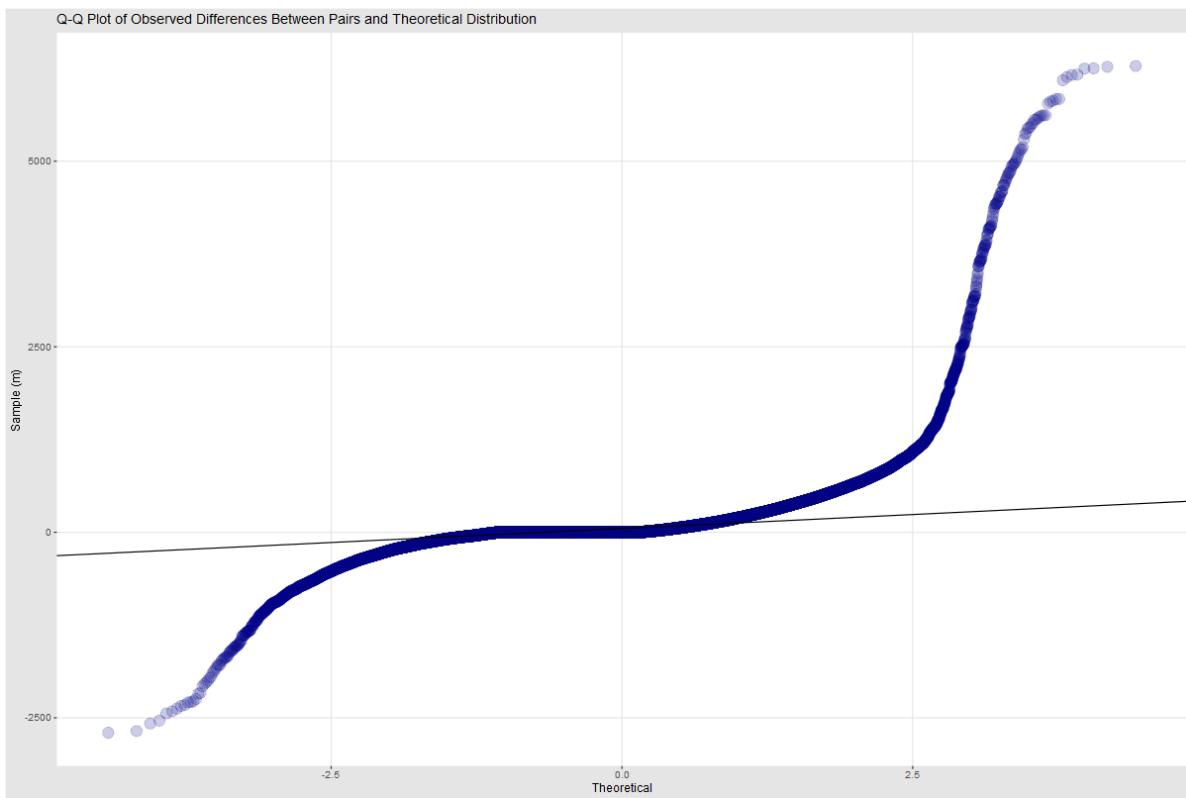
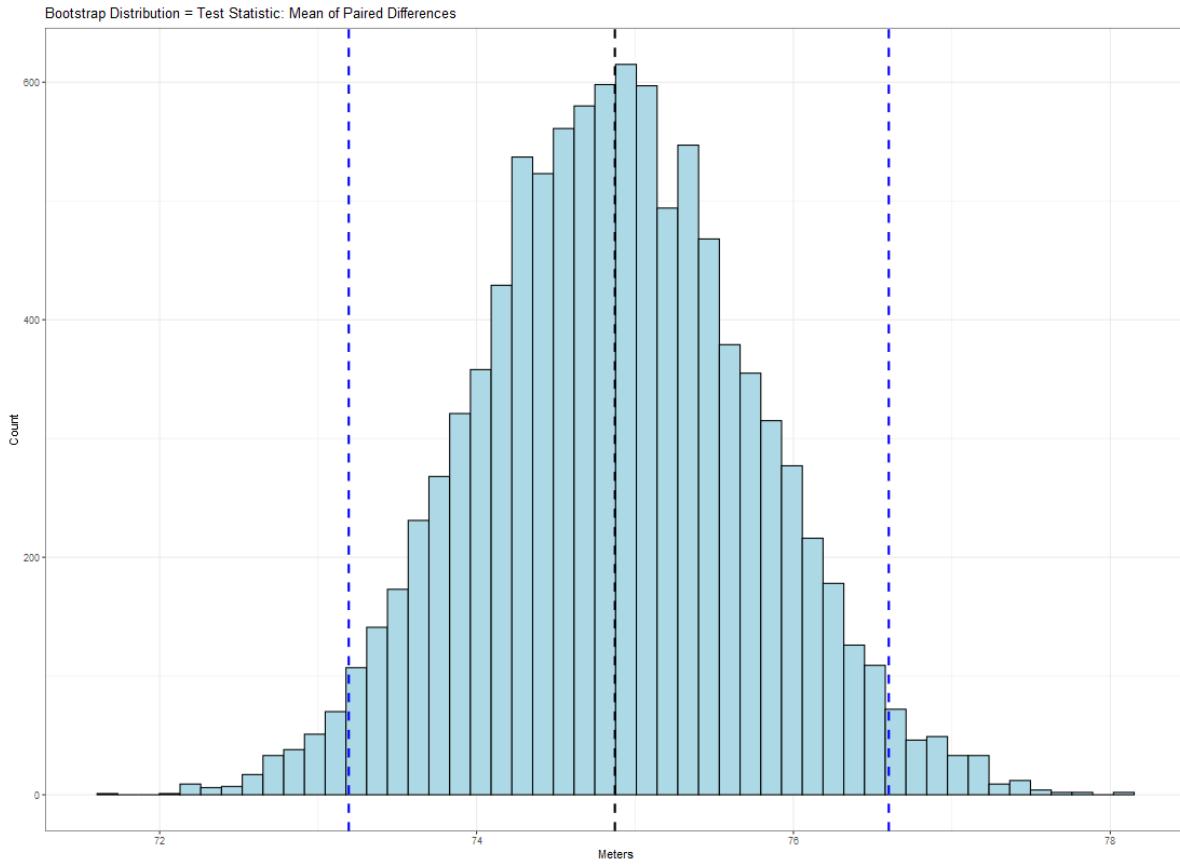


Figure 2



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)*data$difference
  results = c(mean(permutedData), results)
}

results <- as.data.frame(results)
```

The above code generates 5,000 values stored in the vector `results`. Each value is a simulated test statistic, creating an approximated sampling distribution.

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 / there is no true difference in the average distance between points.

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

The permutation test computes 5,000 permutations of the 100,000 paired points, where the signs of the differences of before/after values are randomly flipped. At each permutation, the test statistic (the mean of all the paired differences) is computed. This creates a (usually) roughly normal distribution of mean values under the null. This works on the assumption that 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

The results of our permutation test are indeed significant, and support the alternative hypothesis: the average change in distance to the nearest road **has** significantly **decreased**. To the best of our knowledge, according to our bootstrap resampling, the true mean change in distance is approximately $74.87 +/ - 0.87$ meters, with a 95% CI of 73.19 m, 76.60 m.

Future work is needed to apply this methodology to freshwater features. This is actually the most important part of our analysis, since our goal is to study the effects of roads on native invertebrate populations. Below, Figure 4 and Figure 5 show visually the impact of road construction in Virginia, separated by land cover class and selected EPA ecoregions. This is a rough way to assess what areas the study area have been most “effected,” or in other words, which areas have had the most points become **closer** to the nearest road. In the figures below, all the points below the dotted 1:1 line (indicating no change), have become closer to the nearest road, which increases the chance of negative ecological effects, such as pollution from runoff, bank erosion, and increased risk of animal collision. It is worth noting that all

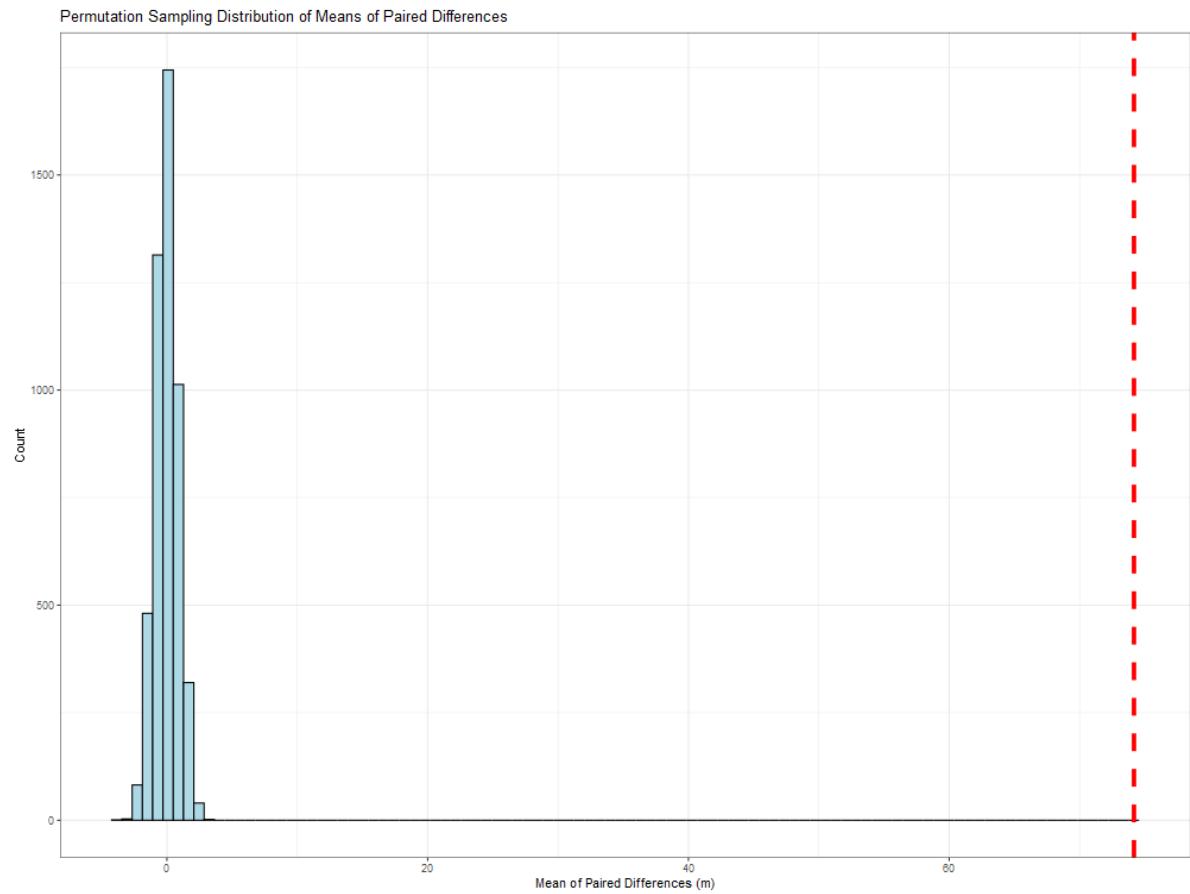


Figure 3: Visualization of permutation test for statistically significant difference in the means of 2000 and 2024 distributions, $p < 0.0002$

scatterplots show a general trend of having **more points under the no change line than over the no change line**. This indicates widespread encroachment of road features across all land covers and ecoregions. However, in all cases, the majority of points fall **exactly on the no change line**. What this means is that the vast majority of points have not changed. This added certain complications in our statistical analysis. For example, the difference in the medians of the distributions is non-zero, however, the median of the differences between the paired points **is 0**. Since this is a pairwise analysis with unique data structure, we must be careful when trying to draw inferences from our data.

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

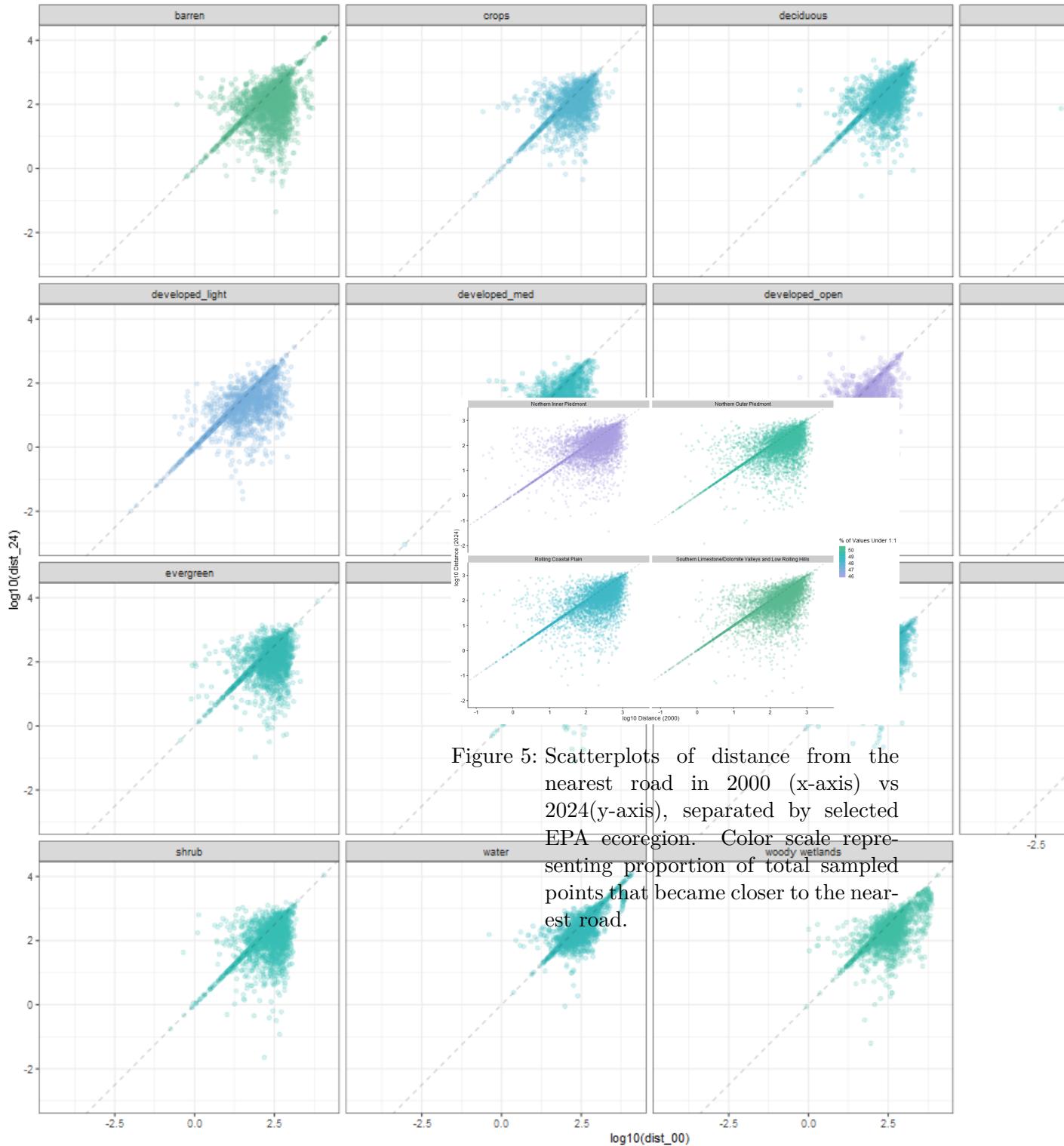


Figure 4: Scatterplots of distance from the nearest road in 2000 (x-axis) vs 2024(y-axis), separated by NLCD¹ class. Color scale representing proportion of total sampled points that became closer to the nearest road.

Figure 5: Scatterplots of distance from the nearest road in 2000 (x-axis) vs 2024(y-axis), separated by selected EPA ecoregion. Color scale representing proportion of total sampled points that became closer to the nearest road.

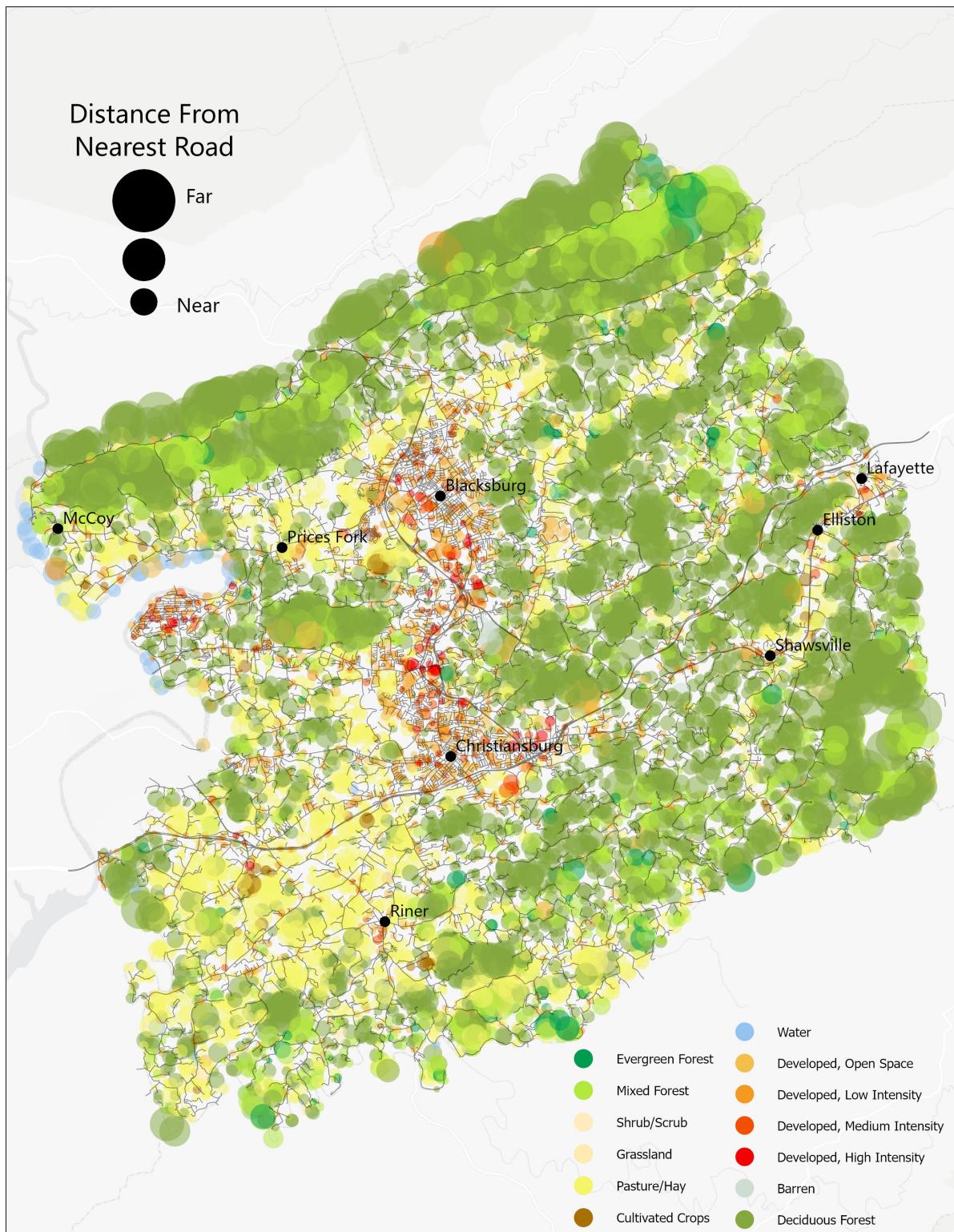


Figure 7: Map of Montgomery County, Virginia. Point color represents land cover classification, and point size represents its distance to the nearest road

6 Figures

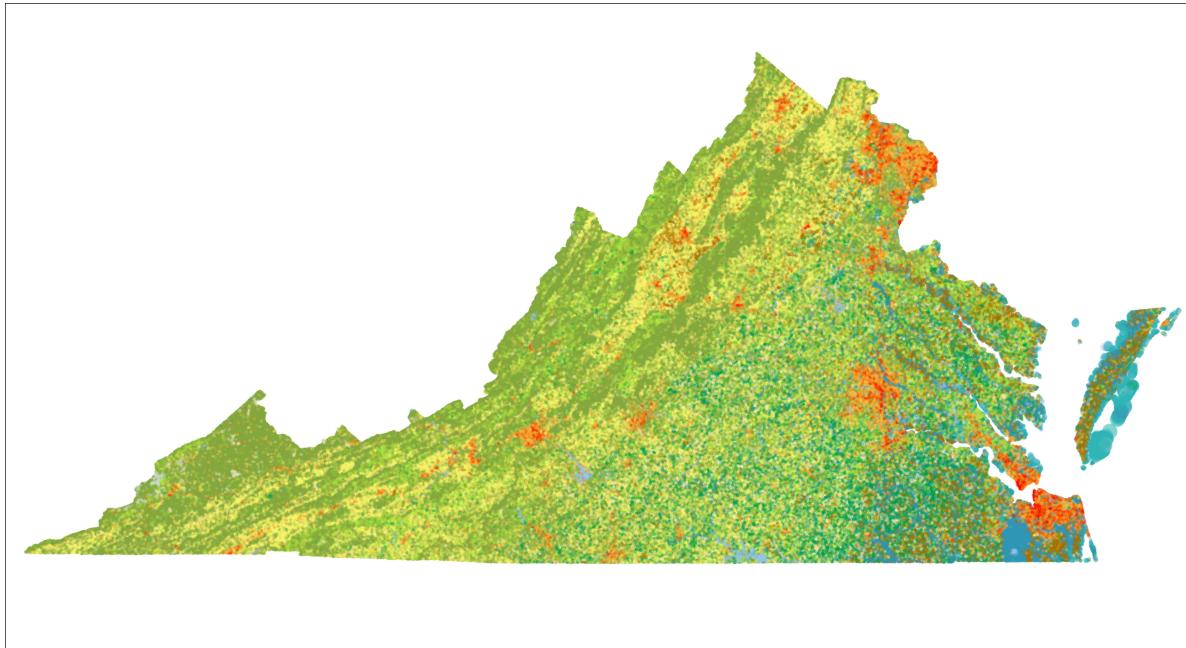


Figure 8: Map of data across Virginia. Symbology is the same as in Figure 7.

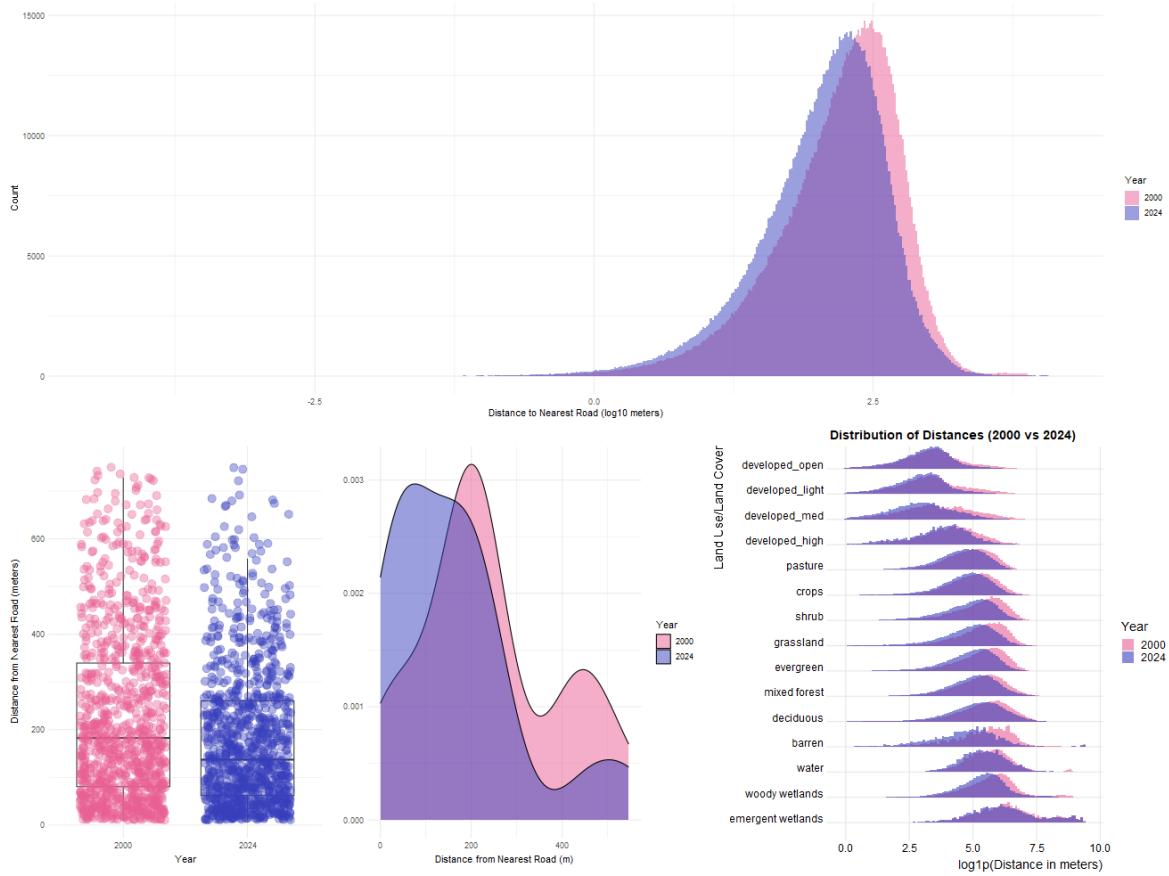


Figure 9: (a). Histograms for two paired distributions, for the year 2000 and 2024, showing a left-shift from the measured values, indicating that the measured points became closer to the nearest road, (b). Boxplot showing the distributions of distance to the nearest road in each year, (c). Density plot 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 10: Example of points sampled along water features in Franklin County, VA

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 Geological Survey

References

- “Annual NLCD Collection 1 Science Products.” 2024. US Geological Survey (USGS).
“Connectivity Planning | Conservation Planning in the Hudson River Estuary Watershed.” n.d. <https://hudson.dnr.cals.cornell.edu/conservation-planning/inventory-and-planning/connectivity-planning>.

- Dixon, Philip M. 2001. "Bootstrap Resampling." *Encyclopedia of Environmetrics*, October. <https://doi.org/10.1002/9780470057339.VAB028>.
- Forman, Richard T. T., and Lauren E. Alexander. 1998. "Roads and Their Major Ecological Effects." *Annual Review of Ecology and Systematics* 29: 207–31. <https://doi.org/10.1146/ANNUREV.ECOLSYS.29.1.207>.
- 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>.
- Karlson, Mårten, and Ulla Mörtberg. 2015. "A Spatial Ecological Assessment of Fragmentation and Disturbance Effects of the Swedish Road Network." *Landscape and Urban Planning* 134 (February): 53–65. <https://doi.org/10.1016/J.LANDURBPLAN.2014.10.009>.
- Mann, Deepika, Mangalasseri Mohammad Anees, Shalini Rankavat, and Pawan Kumar Joshi. 2021. "Spatio-Temporal Variations in Landscape Ecological Risk Related to Road Network in the Central Himalaya." *Human and Ecological Risk Assessment* 27: 289–306. <https://doi.org/10.1080/10807039.2019.1710693>.
- McGregor, Rachelle L., Darren J. Bender, and Lenore Fahrig. 2008. "Do Small Mammals Avoid Roads Because of the Traffic?" *Journal of Applied Ecology* 45 (February): 117–23. <https://doi.org/10.1111/J.1365-2664.2007.01403.X>.
- "Office of Highway Policy Information - Policy | Federal Highway Administration." n.d. <https://www.fhwa.dot.gov/policyinformation/pubs/hf/pl11028/chapter1.cfm>.
- "Permutation Test for Matched Pairs Data." n.d. <https://people.hsc.edu/faculty-staff/blins/classes/spring19/math222/Examples/MatchedPairsPermutationTest.html>.
- Pinto, Fernando A. S., Alex Bager, Anthony P. Clevenger, and Clara Grilo. 2018. "Giant Anteater (*Myrmecophaga Tridactyla*) Conservation in Brazil: Analysing the Relative Effects of Fragmentation and Mortality Due to Roads." *Biological Conservation* 228 (December): 148–57. <https://doi.org/10.1016/J.BIOCON.2018.10.023>.
- "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.
- Saura, Santiago, and Lidón Rubio. 2010. "A Common Currency for the Different Ways in Which Patches and Links Can Contribute to Habitat Availability and Connectivity in the Landscape." *Ecography* 33 (June): 523–37. <https://doi.org/10.1111/J.1600-0587.2009.05760.X>.
- Szekely, Gabor J, and Maria L Rizzo. 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>.