Smart Streams: Machine Learning for Healthier Streams in Prince George's County

Gwendolyn Zeckowski
May 19th, 2025

Abstract

Since 1993, Maryland Biological Stream Survey (MBSS) has been conducted through the Maryland Department of Natural Resources (MDDNR). It investigates 75-meter stream segments, or reaches, across all of Maryland. It measures the health of the given stream with the macroinvertebrates and fish populations through the Indices of Biotic Integrity (IBI). There are many factors affecting the stream segments such as canopy cover, impervious surfaces, population, and land use. To investigate this outside of the 75-meter stream segment, in Prince George's County (PGCo), Maryland, regression models such as ordinary least squares (OLS), geographically weighted regression (GWR), and forest-based and boosted classification and regression (RFR) were tested on the calibration and validation of the dependent variable, IBI, with the explanatory variables, canopy cover, impervious surfaces, population, and land use. The population was held by the census tracts. They were then joined with the IBI and stream flow direction for the IBI's. The explanatory variables were reclassified into raster's and then they were tabulated to find the percent of canopy cover, percent impervious surfaces, percent agricultures area, percent environmental area, and percent urban area. They were then joined into the census tracts. Now the census tracts hold the values of the IBI's, stream flow direction, percent canopy cover, percent impervious surfaces, percent agricultures area, percent environmental area, and percent urban area. A multilinear regression and an exploratory regression were used to find out which variables were most important by conducting Regression Analysis Check. It was found that population by the census tracts, percent agricultural area, percent environmental area, and percent urban area reached the most significance through the explanatory regression and the multilinear regression. The OLS and GWR for the variables were then conducted. R^2 for the training was 0.286 and R^2 for the testing was 0.220. For GWR the training R^2 was 0.453. These models were underfit. A machine learning model of RFR was then calculated. For 100 trees, training R^2 was 0.835 and testing R^2 was 0.009. The p-values for training were 0.00 and testing was 0.542. For 50 trees, training R^2 was 0.851 and testing R^2 was 0.010. The p-values for training were 0.00 and testing was 0.177. For 25 trees, training R^2 was 0.816 and testing R^2 was 0.206. The p-values for training were 0.035 and testing was 0.366. It was found that 50 trees were the most consistent but all the R^2 values were overfit. If this was to be done again, it would be conducted for the IBI values within the whole state of Maryland.

1 Introduction

1.1 Problem Statement

In Maryland there are studies of Maryland Biological Stream Survey (MBSS) that have been conducted through the Maryland Department of Natural Resources (MDDNR). These studies are conducted on 75-meter stream segments investigating the Indices of Biotic Integrity (IBI) and

the Physical Habitat Assessments (Klauda et al., 1998). These are conducted on 84 stream segments each reaching 75-meters (Klauda et al., 1998). These tests do not explore farther from the 75-meter stream segments. From the IBI's that are found, macroinvertebrates and fish are collected to determine the overall quality of the streams (Klauda et al., 1998). In Prince George's County (PGCo) these studies are collected. Is the stream health also determined by the various factors surrounding the stream? This research project aims to look at a whole county instead of the 75-meters stream segment set out by MDDNR through the MBSS. Prince Georges County stream health will be investigated by exploring the dependent variable, Indices of Biotic Integrity (IBI), with the explanatory variables, canopy cover, impervious surfaces, population, and land use cover.

1.2 Literature Review

1.2.1 Indexes of Biological Integrity

IBI is an important factor in determining the overall stream heath. IBI is defined by the indices of biotic integrity (MD iMap). The combination of the fish index of biological integrity, F-IBI, and the benthic index of biological, B-IBI (MD iMap). IBI is rated as poor, 1.0 to 3.0, fair, 3.0 to 3.9, and good, 4.0 to 5.0 (MD iMap). These ratings were derived from the Maryland Biological Stream Survey, MBSS, to the data which were collected, F-IBI and B-IBI, more collectively. In this paper, IBI will be used as a dependent variable to examine the various explanatory variables – population density, pollution level, agriculture and impervious areas, and canopy cover – and their overall contribution to the stream's health.

1.2.2 Prince Georges County, Maryland

Lessard et al, (2012) investigated PGCo in the Potomac River-Anacostia River (ANA), Potomac River -Non-Anacostia (PNA), and Patuxent River (PTX). It was found that there is very poor quality of streams within the area based on the following reasearch. Researchers used the MBSS in the streams to explore land use changes resulting in the hydrologic characteristic of the stream (Lessard et al., 2012). They analyzed the stressors, where they originate, and how they affect the overall B-IBI of macroinvertebrate and F-IBI of fish factors (Lessard et al., 2012). Results show that the ANA Basin, with B-IBI accounting for 0.13 R^2 value of variation, was the most degraded. The PNA, with a B-IBI R^2 value of 0.17, was similarly degraded compared to the PTX (Lessard et al., 2012). The best model of F-IBI was in PTX where it had 31% being explained by pool substrate, low density residential and pastureland (Lessard et al., 2012). The urban locations were found to have poor

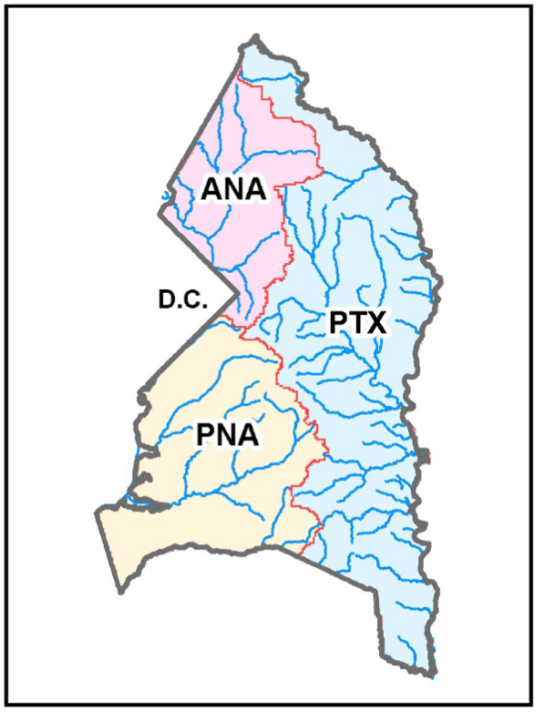


Figure 1. *Drainage Area Map of ANA, PTX, and PNA (Lessard et al., 2012)*

quality, and forested and agricultural areas also had poor quality. In residential areas, such as parks, fields, large yards, golf courses, etc., stressors were more common, as suspected (Lessard et al., 2012). According to this study, returning the streams within PGCo to pre-settlement conditions is not attainable. However, it should continue to restore these streams to make them best compatible with the humans living in the area (Lessard et al., 2012).

Stribling (2024) analyzed the ANA, PTX, and PNA watersheds for erosion and sediment productions. This study was drawn from data collected by TetraTech[1]. It compares data collected from 2000- 2004 with data collected in 2020. These studies were conducted at the cross-sectional[2] geomorphic area, or XS (Stribling, 2024). This study looks from bank to bank within the stream, calculating the width, not the length of the stream. There were 78 streams evaluated in the years 2000-2004 and 2020 (Stribling, 2024). Comparisons were made. The sediment loss totaled 74.4%. Geomorphology, or the XS, there were 45 reaches that experienced no change while 21 reaches have increased the channel instability[3]. Precipitation increase overtime, although Stribling notes this study calls for more time to evaluate the data. For land use, where the land has increased development there were increases in instability. At the same time, areas that were already urbanized had little to no change. This does not mean the quality of the urbanized stream is poor. In an overall synopsis of the erosion and sediment production happening in this Stribling paper, it seems that a stream can only be determined to be in good condition given a positive biological response, which was not explored in this study (Stribling, 2024).

1.2.3 Canopy Cover

Alberts et al., (2017), studied eight streams in northern Kentucky during all seasons: fall, winter, spring, and summer. It explores ecosystem metabolism, "the total energy processed by all the individual organisms that make up an ecosystem" (Springer, 1999). This study investigates ecosystem metabolism by conducting research on canopy treatments during the season aligned with the flow regime on the streams (Alberte et al., 2017). Gross primary production (GPP) was highest in the summer for urban streams. Ecosystem respiration (ER) was highest in the spring and lowest in the fall (Alberte et al., 2017). In all the seasons listed, the strongest correlation of the stream metabolism is most impacted by the summertime, when canopy cover is high (Alberte et al., 2017).

Hession et al. (2003) explores the Piedmont region of the United States – southeastern Pennsylvania, northern Maryland, and Delaware – and looks at forested and non-forested riparian zones[4] in rural and urban areas. Riparian forest restoration is becoming a significant part of the restoration of watersheds (Hession et al., 2003), so it is important to view the differences in the regions. This study examines 26 repaired streams with a large cover of land use areas of both the urban and rural settings, studying the morphological[5] zones in the time frame from 1997 to 1999. Each zone was calculated by the percent impervious land cover taken from Delaware

[1] PGCo uses TetraTech and MBSS in combination to collect data
[2] [4] Cross section data represent the geometric boundary of the stream
[3] [17] Associated with increased sediment supply, land productivity change, land loss, fish habitat deterioration, changes in both short and long-term channel evolution and loss of physical and biological function
[4] [15] Land between land and a stream
[5] [16] Shapes of river channels and how they change in shape and direction over time

Valley Regional Planning Commission (DVRPC) aerial images and the Landsat thematic mapper satellite imagery (Hession et al., 2003). It was found that forested streams are wider than unforested streams and they have greater cross sections. This is regardless of whether they are urbanized or nonurbanized (Hession et al., 2003). In this literature review, there are several factors affecting streams health, as determined by the IBI. As the years pass by, there is an ongoing increase in the population density of the entire world, as in the United States. The increase in population correlates with an increase in impervious surfaces. Increased population results in the need for more housing and commercial development, as well as increased agricultural production to meet new food demand. Expanded agriculture leads to escalated amounts of fertilizer runoff into surrounding streams and waterways. As more people inhabit the earth, they begin cutting down trees subsequently decreasing the overall canopy cover. Thus is the reasoning behind the explanatory variables.

1.2.4 Impervious Surfaces

Urbanization has created flood peaks, increased nutrient load, increased sediment, increased heavy metals, and significant thermal shock on streams. Impervious surfaces in urbanized areas have a major hydrologic impact (Moglen, 2009). Impervious areas are defined as, “any material that prevents the infiltration of water into the soil” (Arnold & Gibbons, 1997). In 1903, 93% of roadways were unpaved, allowing water to seep into the ground although contaminated by various chemicals (Arnold & Gibbons, 1997). From this point on in the 20th century, highways systems and large buildings have only intensified the effects of impervious land surfaces. With this occurrence at hand, hydrologic disruptions giving physical and ecological impacts, increasing the chances of runoff and erosion into streams (Arnold & Gibbons, 1997). Increasing toxic contaminants such as heavy metals, pesticides, and phosphorus lead to algal blooms. In surface waters, and sediment runoff for chemicals being included within those eroded particles these algal blooms keep increasing in size (Arnold & Gibbons, 1997). Impervious surfaces that degrade waterways are involved with the land uses that generate pollution, prevent the natural breakdown by soil, and transport pollutants into waterways. These factors have been battled with our urbanized areas incorporating pervious surfaces.

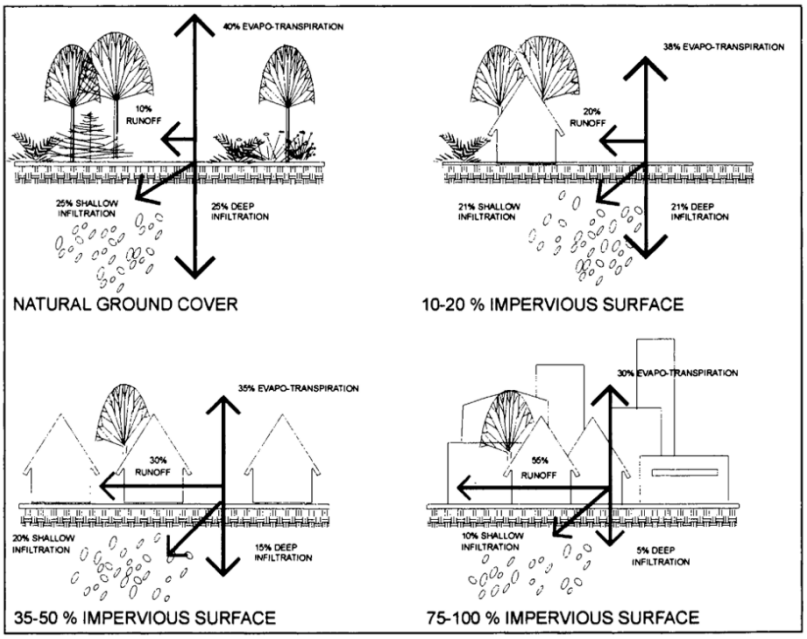


Figure 2. *Image displaying the rate of impervious surfaces increasing (Arnold & Gibbons, 1997)*

1.2.5 Population Density

Bongaarts (2009) put the human population at 6.5 billion in 2005. It states that by 2070, the world population will be approximately 25 billion (Bongaarts, 2009). These studies show that our human population continues to expand exponentially. Since humans are affecting the globe on such a scale, whenever an environmental project is conducted, the population density at that specific time should be considered. Metre et al. (2019) looks at urban growth within the Piedmont area (Virginia, North Carolina, South Carolina, Georgia, and Alabama) witnessing the growth of urban population. Using regression models predicting the future of the Piedmont area into the year 2060, it was found that there is expected to be a 5% urban land use growth within the given watersheds reaching from Alabama up into Virginia. Metre et al. state that urban land use of 50% corresponds to a loss of almost one half of invertebrates and one third of fish. Urban land use will cover 10% to about 30% in the Piedmont area in 2060 (Metre et al. 2019). Boosted regression tree models were also tested and resulted in the cross-validation studies of the R^2 being 0.47 and 0.63 in the year 2009. As seen in the image from the Metre et al. study, in 2060 there will be great loss of most invertebrates and fish in the projected data.

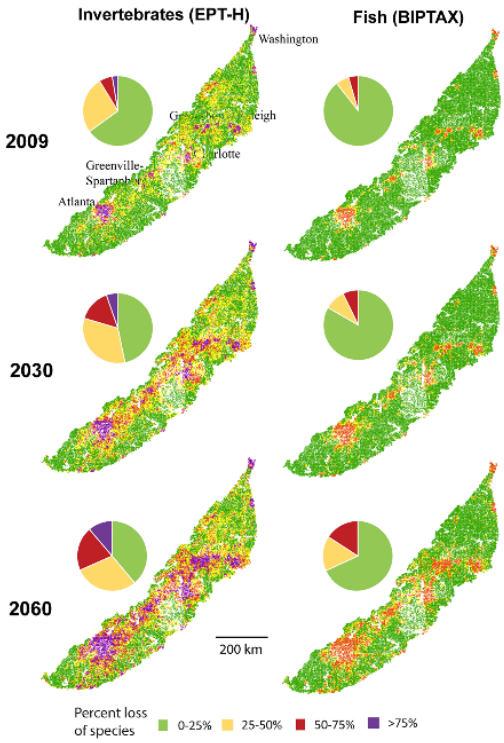


Figure 3. *Maps showing projected stream data in the Piedmont Region (Metre et al., 2019)*

1.2.6 Land Use

Dale (1997) presented the argument of land use changes and how it is affecting the globe. Since the early 1800's, there have been increases of carbon dioxide, methane, and nitrous oxide with recent increases in population. With the rising changes in technology, political economics, political structure, attitudes, and values, there are increases in climate change and the way in which our communities make mention of it (Dale, 1997). Not only are there changes in human structures, but there are also changes on an ecological basis: biodiversity, habitat availability, soil erosion and degradation, productivity, extractable resources, and water quality continuing to increase in a negative light (Dale, 1997). Dale states that there are ways to monitor these changes. On a global scale, vegetation change has been used to project vegetation patterns in the climate processes, plant groups instead of species, and navigating climate and vegetative models in a spatial context (Dale, 1997). Global models can be used to examine land-use change, biomass estimates, and carbon on a land-use scale (Dale, 1997). Regionally, these models can be used to examine scenarios of land management and climate change to determine sensitive variables and features of the region, such as vegetation affected by the carbon budget or a process in which land-use changes play a key role in (Dale, 1997). Dale states that there are some steps that can be taken to illuminate land use change and climate change: interdisciplinary studies of land use, spatially explicit models, relations between cause of land use and actual land

use, ancient organisms and their environmental interactions, agriculture interactions, monitoring the spatial distribution between natural vegetation and human activity, and education (Dale, 1997). Since the writing of this paper in 1997, the hypotheses have become true facts.

1.2.7 Regression Modeling – OLS, GWR, and RFR

Sheehan (2011) studied stream habitat spatial modeling using OLS and GWR. The study area was two steam sites located in the Greater Yellow Stone Ecosystem. Sheehan measured water depth, water velocity, and benthic substrates in the streams. Sampling periods were between July 2008 and August of 2008. The OLS indicated that both flow and depth were significant in the model (Sheehan, 2011). GWR revealed higher adjusted R^2 than OLS in both stream settings. GWR captures the natural spatial patterns and regional variation better than the OLS could (Sheehan, 2011). It is stated that using GWR in steam habitat modeling is important due to the implications in three things: data collection, habitat assessment, and habitat prediction (Sheehan, 2011). GWR provides the ability to obtain more accurate results.

Pumhirunroj et al, (2023) released a study using forest-based classification and regression in the Phon Na Kaeo district of Sakon Nakhon Providence in northern Thailand. It studies liver flukes and cholangiocarcinoma parasites that inhabit humans causing bile cancer (Pumhirunroj et al, 2023). From 2018 to 2021, 12,063 cases were reported after consumption from a nearby watershed (Pumhirunroj et al, 2023). Pumhirunroj et al, therefore studied an infected fish within two model sites of the stream using the forest-based classification and regression. It resulted in a 0.964 stating that both models had the highest area under the curve when running the forest-based classification and regression (Pumhirunroj et al, 2023). Although these are great results, Pumhirunroj et al, states that the forest-based classification and regression can be improved by training the model and testing it.

1.3 Objectives

The objectives for this research study in Prince George's County, Maryland from the period of the IBI data, 2000 to 2017, are as follows:

- Data Collection and Processing: The collection of the dependent variable (IBI) and explanatory variables (canopy cover, impervious surfaces, population, and land use) will be aggregated based on a spatial modeling unit. The spatial modeling units are, percent canopy cover, percent impervious surfaces, percent agricultural area, percent environmental area, percent urban, and population of the census tracts. These units will be run with multilinear regression and exploratory regression to determine the most significant variable. The data will then be divided into calibration and validation data.
- Generate Models: The calibration data will be run on the most significant explanatory variables with the dependent variable, IBI, to determine OLS, GWR, and RFR. The dependent variables impact on the explanatory variables will be estimated by the models.
- Evaluate the Models: The validation data will be run on the most significant explanatory variables with the dependent variable, IBI, to determine OLS, GWR, and RFR. The dependent variables impact on the explanatory variables will be estimated by the models.

The purpose of the validation data is to test if the regression models and determine its fit (over fit, well-fit, or underfit) with the validation data.

- Compare and Determine the Results: The calibration and validation results for R^2 will be used to determine the models fit.
- Develop: An Esri StoryMap with ArcGIS Online Mapping components to best interpret the results and outcome of models, OLS, GWR, and RFR.

2 Methodology

2.1 Study Area

This study uses county-level data based on PGCo, Maryland. The 24 watersheds within this data are important to account for [Figure 1]. This area of study is important when accumulating data since the data will need to be positioned in PGCo. All other counties in Maryland are valid; this area was chosen due to the author conducting MBSS within PGCo.

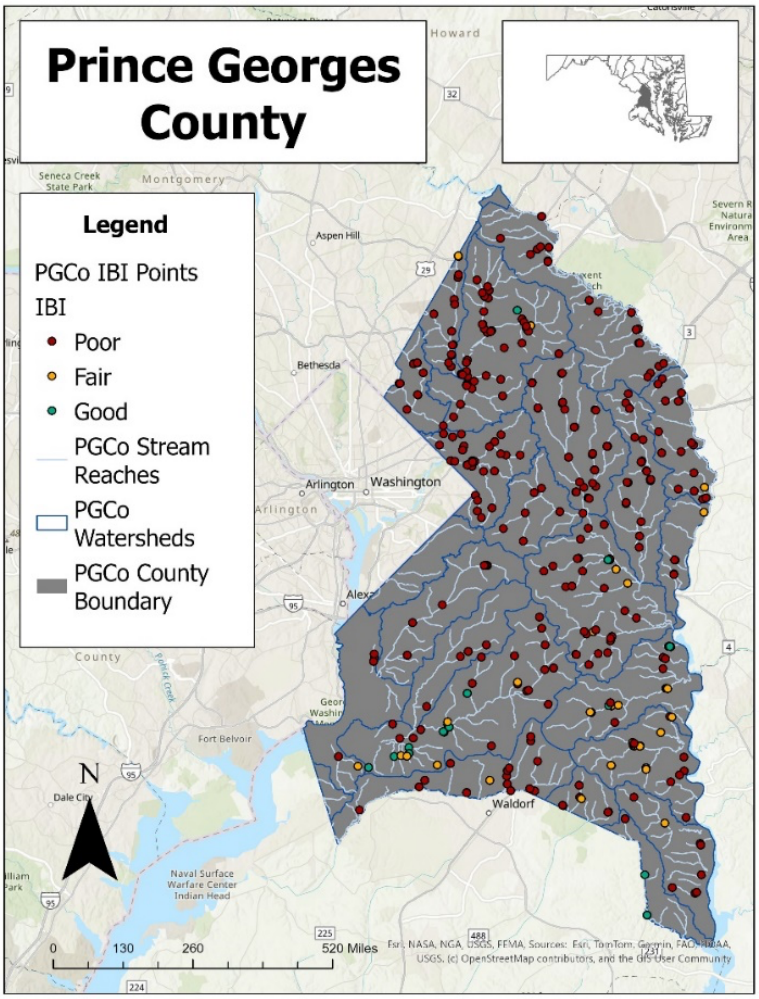


Figure 4. *PGCo Map Study Area*

PGCo is 428.7 square miles in area, placing it sixth largest, out of 23, counties in the state of Maryland (U.S. Census Bureau). From the 1700's PGCo had an agriculturally based economy with the leading cash crop being tobacco (Virta). The 1800's brought a shift in the way PGCo operated its agriculture methods. Great machines, mass production, and hundreds of workers all came from England. But by 1860, approximately 90% of the agricultural production was done by African American slaves (Virta). After the Civil War, average farm size services provided people with the ability to expand their knowledge and broaden their horizons (Virta). The population of PGCo continued to grow from 30,000 in 1860 to 60,000 by 1930 to 350,000 by 1960 (Virta). Since the 1970's the population has skyrocketed 660,000 in 1970 to 950,000 in 2022 (USAFacts, 2025). Because urban setting increased exponentially, farms disappeared, and agriculture vastly decreased

(Virta). As the population grows, it is known to influence the environment greatly. Trees must be cut down to build houses. Impervious surfaces continue to increase as roadways and sidewalks are built.

There are hundreds of streams in PGCo associated with the 24 different watersheds. Since the 1800's the quality of these streams has decreased. Due to lack of trees, impervious surface creating run-off, and trash thrown into the streams, only to be carried out to a larger body of water, potentially the Chesapeake Bay. It is important for us as humans to watch over the streams and interpret their quality based on these factors.

2.2 Data Sources

Table 1 summarizes the data to be used in this project, covering a wide variety of vector data – points, lines, and polygons – and raster data. It primarily relies on data from 2020 onward, except for the MBSS IBI from 2000 to 2017. This was the only data that could be found. Although the data is inaccurate, there could be issues with access to various points and the display could slightly be off the newer points since updates have been made to ArcGIS since 2000.

Each of these data sources are useful for the project. For PGCo Population data as a CSV, which holds the population for the census tracts, was joined with PGCo Population Data as a shapefile joining the accumulation of two data points to receive the population within the census tracts. It will then be projected to NAD1983 State Plane MD FIPS1900, as the other sources of data are set to. PGCo 12 Digit watersheds were clipped to the PGCO boundary to guarantee only PGCo will be in use. The Maryland Stream Health data was shifted to the projection of NAD1983 State Plane MD FIPS1900, to match the other sources of data. The land use raster of the entire USA was cropped using PGCo boundary PGCo. The other data sources are all accurate and ready for use.

***Table 1**. Chart describing the datatypes, sources, and their quality and projection for use in this project.*

Data	Data Type	Source	Quality and Projection
Prince George's County – Boundary (2015)	Shapefile – Polygon	GIS Open Data Portal	**Very Accurate.** NAD1983 StatePlane MD FIPS1900.
Prince George's County – 12 Digit Watersheds (2024)	Shapefile – Polygon	NRCS Geospatial Data Gateway	**Very Accurate**. NAD1983 StatePlane MD FIPS1900.
Maryland Stream Health MBSS Stream Reach File (2018)	Shapefile – Line	Maryland.gov	**Accurate.** WGS 1984 Web Mercator (auxiliary sphere); to be changed. Data from 2018.
Maryland Stream Health MBSS IBI Data (2017)	Shapefile – Points	Maryland.gov	**Accurate.** WGS 1984 Web Mercator (auxiliary sphere); to be changed. Data from 2000 to 2017.
Prince George's County – Population Density (2020)	Shapefile – Polygon	United States Census Bureau	**Accurate**. NAD 1983 UTM Zone 18N; to be changed.
Prince George's County Population Data (2020)	CSV	United States Census Bureau	**Very Accurate**. To be added into the Population Density per Census Tract.
Prince George's County – Canopy Cover (2020)	Shapefile – Polygon	GIS Open Data Portal	**Very Accurate.** NAD1983 StatePlane MD FIPS1900. No data for government facilities or airports.
Prince George's County – Impervious Surface Area (2020)	Shapefile – Polygon	GIS Open Data Portal	**Very Accurate**. NAD1983 StatePlane MD FIPS1900. No data for government facilities or airports.
Land Cover of Entire USA (2020)	Raster 30m – TIFF	MRLC NLCD Land Cover	**Very Accurate**. Entire United States of America covered.
Prince George's County DEM (2021)	Raster 2m - GRID	GIS Open Data Portal	**Very Accurate**. NAD1983 StatePlane MD FIPS1900.

Since the MBSS data is in point form of the midway points of the 75-m stream reach, not the actual stream lengths, this could result in a slight issue in the representation of points instead of lines. Stream points are 365 out of 6,951 within PGCo, of the 23 counties in Maryland, if each one were to average 302 streams, 365 seems to be a good approximation for the number of streams that fall within PGCo. Although, this could be an over approximation for the number of streams within all of Maryland.

2.3 Methods

Regression analysis with OLS, GWR, and RFR will be completed in ArcGIS Pro. It will incorporate the independent variable, IBI, and the explanatory variables, canopy cover, impervious surfaces, population, and land use cover. The final will result in R^2 values and Adjusted R^2.

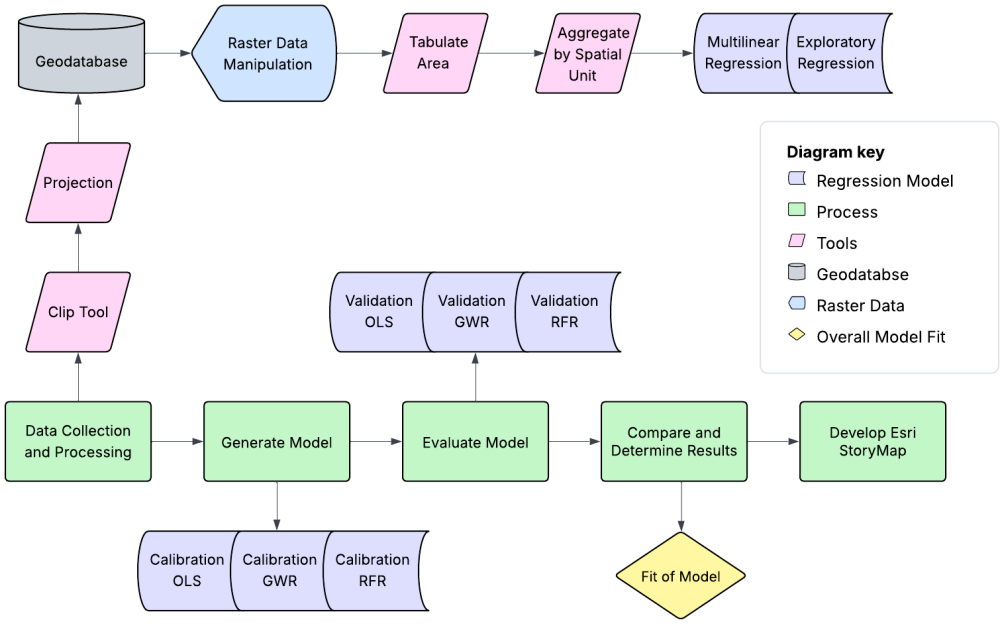


Figure 5. *Workflow for Methodology*

2.3.1 Data Collection and Processing

The data Prince Georges County Population Density (2020) in the form of a polygon shapefile of the census tracts and the Prince Georges County Population Data (2020) in the form of a CSV need to be combined. It was found that the Prince Georges County Population Density (2020) contained only the census tracts, no population data. For the consolidation of these two files the first row of the CSV needs to be taken out. This was done by opening Excel and removing the first row of data then uploading the CSV into ArcGIS Pro. Then the polygon attribute table was opened in ArcGIS and a new field was added. The following CSV file format was correctly input to the new field, P1_001N representing the data for the population in each of the census tracts.

The clip tool was used for data to make it more accurate to PGCo. The following data was entered: Prince George's County – 12 Digit Watersheds (2024), Maryland Stream Health MBSS Stream Reach File (2018), and Maryland Stream Health MBSS IBI Data (2017). They were clipped to the Prince George's County – Boundary (2015).

A couple of data shapefiles have the projections adjusted to them: Maryland Stream Health MBSS Stream Reach File (2018), Maryland Stream Health MBSS IBI Data (2017), and the Prince George's County – Population Density (2020). These were corrected through the Project Tool in ArcGIS Pro. They were situated as NAD1983 StatePlane MD FIPS1900 (US Feet) as all the others were adjusted too.

Once all those tasks were completed, a geodatabase, GDB, was made for the corrected data. The GDB holds a feature class with the following data. For the polygons, County Boundaries,

Canopy Cover, Census Tracts, Impervious Surfaces and the watersheds within PGCo. For the lines, PGCo Stream Reaches. For the points, PGCo IBI points. All were input on a map.

Table 2. *Reclassification of NLCD Data*

Reclassification	Area
0	Large Bodies of Water
1	Urban Areas
2	Agricultural Areas
3	Environmental Areas

Raster data was then input into the map's contents, NLCD Land Cover of Entire USA (2020) and Prince George's County DEM (2021). The land cover data was then selected and clipped to the raster of PGCo. The reclassifications are shown in table 2.

The reclassification of the land use rater and the DEM for flow direction were added to the GDB through data and raster. The DEM data was situated using the fill tool and then the flow direction tool to receive a map of the flow direction of the raster: one is east, two is southeast, four is south, eight is southwest, sixteen is west, 32 is northwest, 64 is north, and 128 is northeast. This was then placed into the extract values by point; PGCo IBI points as input feature and flow direction raster as the output. The attribute table was now matched with the flow direction for each point given. The flow direction of the IBI points was to match the stream flow from a contributing factor: canopy cover, impervious surfaces, or land use.

The polygon to raster tool was then used for the canopy cover and the impervious surfaces polygon. The value fields were the OBJECTID, and the cell size was two, to make sure it was as detailed as possible, and it aligned with the previously dealt with DEM raster. The environments were set to the input of the field in the output coordinate system and the extent, latitude and longitude, were set as a full PGCo. The following was entered into the raster calculator:

Table 3. *Raster Calculator Calibrations*

Raster Calculator	Equation
Raster for Canopy Cover	Con(IsNull("CanopyCover_polygon2raster"), 0, 1)
Raster for Impervious Surfaces	Con(IsNull("ImpSurfaces_polygon2raster"), 0, 1)

The census tract population needed to be first joined with IBI points and the flow direction. This was done by joining the flow directions to the given IBI points. Next the IBI points and flow direction to the census tracts populations by adding a spatial join. Within the census tracts it now held the IBI points, flow direction, and the population data per census tract.

The canopy cover, land use, and impervious surfaces raster needed to be accurately tabulated into their spatial modeling units. The spatial modeling units are the percent canopy cover area, percent impervious surfaces area, percent agriculture area, percent environmental area, and percent urban area. The tabulate area was conducted for the input vector as the census tracts with the zone field as object ID. The output of the class field of value and the output raster was canopy cover, impervious surfaces, and land use. The total area was calculated as all values. The

percentage of canopy cover was calculated by the (output value of one / the total value) * 100. The percentage of impervious surfaces was calculated by the (output value of one / the total value) * 100. The land use, for percent agriculture area was (output value of two / the total value) * 100. The land use, for percent environmental area was the (output value of three / the total value) * 100. The land use, for percent urban area was the (output value of one / the total value) * 100. Addition of add join of these tabulated areas with the census tract polygon was then made.

The first regression model for multilinear regression, or MLR, on OLS was run to determine what was the most important variable: percent agriculture, percent canopy cover, percent environmental, precent impervious surfaces, percent urban area, or the flow direction.

Then an exploratory regression tool was run to determine what was the most important variable: percent agriculture, percent canopy cover, percent environmental, precent impervious surfaces, percent urban area, or the flow direction. These were determined through Regression Analysis Check. This is the determination of the six factors used when studying a regression model. The signs of the coefficient indicate whether there is a positive sign, indicating a positive relationship or negative sign, indicating there is a negative relationship. The VIF value, if larger than 7.5, indicates that one than one variable is telling the same story; it is not necessarily a bad thing. An asterisk next two the probability or the robust probability tells whether a variable is significant or not. When the Jarque-Bera Statistic has an asterisk next to it, this suggests that the model that model is significantly significant, and the model can be biased. Therefor there should not be an asterisk beside this. The model's performance or the R^2 value, running between zero and 1, usually when it is 0.5 or higher, suggests a "good" model has been run. Finally, determining the spatial autocorrelation, the spatial clustering of over and under predictions. These six factors will be used when looking into the regression models' accuracy with the exploratory regression tool and MLR.

The variables are combined and the exploratory regression and MLR showed which variable was most important. A process of subset features on the census tract data was completed. Through the calculated field tab, the join count was subjected to a sequential number to make all numbers different. In the subset feature tool, the data was subject to a 70-30 split. Of the 103 points of data taken into consideration, no nulls considered, 69 points of data were for a set for the calibration data and 34 were for a set of validation data.

2.3.2 Generate Models

The OLS and GWR training data were run for the four most likely explanatory variables: population per census tract, percent agriculture, percent environmental, and percent urban. The training model, for the four most optimal explanatory variables, was placed into RFR. Optimizing the parameters were selected to narrow down the parameters of tree size. The tree size was run per 100 trees, 50 trees, and 25 trees to determine which was the best for the data.

2.3.3 Evaluate Models

The OLS and GWR training data were run for the four most likely explanatory variables: population per census tract, percent agriculture, percent environmental, and percent urban.

2.3.4 Compare and Determine Results

The data for the OLS, GWR, and RFR were compared for both the calibration (training) and testing (validation). Adjusted R^2 and R^2 were explored to determine if the models were under-fit, well-fit, or over-fit in evaluation. The under-fit model is determined if the calibration error and validation error are high. The well-fit means is determined if the calibration error and validation error are low. Over-fit is determined if the calibration error is low while the validation error is high.

2.3.5 Develop Esri Products

The Esri StoryMap creation, is a synopsis of this entire report. It was composed of a synopsis of the literature review, the objectives, the study area, data and the results. ArcGIS Online will be used to represent the results in a format that is cohesive to the given results of OLS, GWR, and RFR. Below is the Esri StroryMap consolidation.

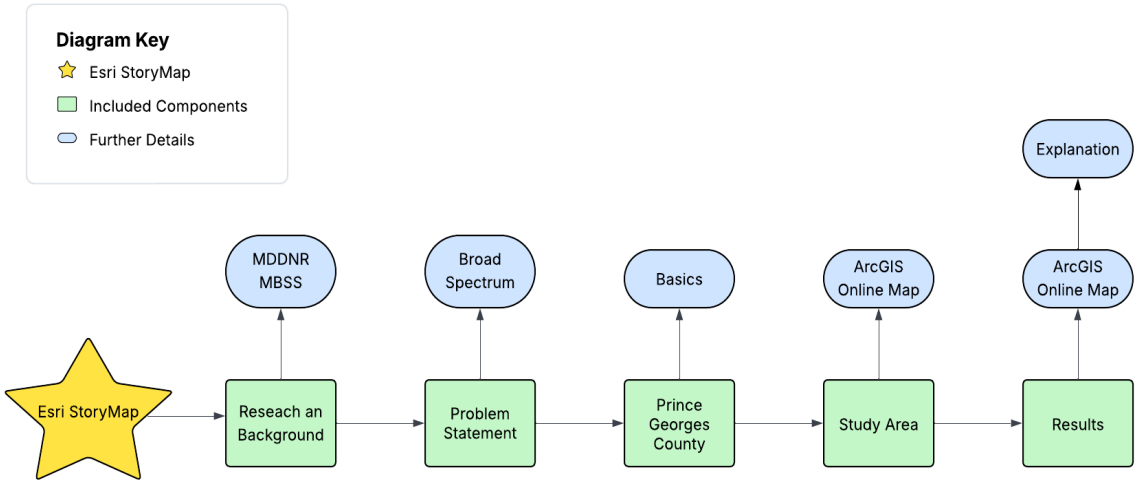


***Figure 6**. Esri StoryMap*

3 Results and Discussion

3.1 Exploratory Regression Analysis

Exploratory regression analysis needs to be made to analyze the explanatory variables. There are two important factors when this comes into play in determining which is the most statistically significant explanatory variable: the summary of variable significance and the VIF violations covariates. The summary of variable significance was population had the highest significance, 71.93%, and a 100% positive, Percent environmental was second with 33.33% significant, and

100% positive. Third was percent agriculture with 29.82% significance, and 91.23% negative. It is important to note that although the flow direction had a 98.25% negative trend, it was 0.00% significant. With the VIF violations covariates, population of the census tracts, flow direction of the IBI, and the percent agricultural, were the only three explanatory variables that did not suffer from multicollinearity. Multicollinearity refers to there being multiple explanatory variables affecting the regression model making it difficult to determine what effect is on the dependent variable. Most of the multicollinearity was percent environmentally affecting the other.

Summary of Variable Significance

Variable	% Significant	% Negative	% Positive
PGCO_CENSUS_TRACTS_POPULATION_EXPORTFEATURES.P1_001N	71.93	0.00	100.00
TABULATEAREA_POPULATION_LAND_USE.PERCENTENV	33.33	0.00	100.00
TABULATEAREA_POPULATION_LAND_USE.PERCENTAG	29.82	91.23	8.77
TABULATEAREA_POPULATION_IMPERVIOUS_SURFACES.PERCENT_IMPERVIOUS	26.32	100.00	0.00
TABULATEAREA_POPULATION_LAND_USE.PERCENTURBAN	26.32	92.98	7.02
TABULATEAREA_POPULATION_CANOPY_COVER.PERCENT_CANOPY	14.04	57.89	42.11
PGCO_CENSUS_TRACTS_POPULATION_EXPORTFEATURES.RASTERVALU	0.00	98.25	1.75

Summary of Multicollinearity

Variable	VIF	Violations	Covariates
PGCO_CENSUS_TRACTS_POPULATION_EXPORTFEATURES.P1_001N	1.15	0	--------
PGCO_CENSUS_TRACTS_POPULATION_EXPORTFEATURES.RASTERVALU	1.15	0	--------
TABULATEAREA_POPULATION_IMPERVIOUS_SURFACES.PERCENT_IMPERVIOUS	8.21	4	TABULATEAREA_POPULATION_CANOPY_COVER.PERCENT_CANOPY (14.81), TABULATEAREA_POP
TABULATEAREA_POPULATION_CANOPY_COVER.PERCENT_CANOPY	11.86	26	TABULATEAREA_POPULATION_LAND_USE.PERCENTENV (96.30), TABULATEAREA_POPULATION_
TABULATEAREA_POPULATION_LAND_USE.PERCENTURBAN	24.46	33	TABULATEAREA_POPULATION_LAND_USE.PERCENTENV (96.30), TABULATEAREA_POPULATION_
TABULATEAREA_POPULATION_LAND_USE.PERCENTAG	2.41	0	--------
TABULATEAREA_POPULATION_LAND_USE.PERCENTENV	31.11	41	TABULATEAREA_POPULATION_CANOPY_COVER.PERCENT_CANOPY (96.30), TABULATEAREA_POP

***Figure** 7. Summary of Variable Significance and Multicollinearity*

3.2 Regression Analysis Check

In the process of the Regression Analysis Check, through multilinear regression, the six processes must be checked. If the regression model fits the six checks, it is a great regression model. Unfortunately, this was not the case. The model fits some of the Regression Analysis check as you can see in the below six model checks.

3.2.1 Coefficient Sign

The coefficients sign could either be negative or positive. The population in the census tracts appeared to have the best positive correlation, 0.000082. The environmental percentage, with 0.17751, made a close second. For the negative correlations, the flow direction of the given IBI scores had the best negative correlation with -0.001882. The urban percentage, with -0.004501, made a close second.

3.2.2 VIF

VIF is the second item to check. To the right is a table suggesting the VIF scores. The flow direction of the IBI points, 1.153, and the population of the census tracts, 1.173, is low. The environmental percentage, 31.912, and urban percentage, 25.349, are very high, higher than the limit, suggesting there are several factors involved with its determination, resulting in an overcount bias.

3.2.3 Statistically Significant Variables

The percent agricultural and the population of the census tract, are the only two variables that contain asterisk next to the numbers of the probability and robust probability. This suggests they are statistically significant. The Koekner Statistic is not statistically significant so we cannot rely on the number for the robust probability. There goes along with the probability; 0.0369 for the percent agricultural 0.0226 for the census tract population.

3.2.4 Jarque-Bera Statistic

This relies on whether the statics can form a normal regression model are normally distributed or follow a bell curve. This is marked by an asterisk. In the situation at hand, there is no asterisk. When this statistic has an asterisk at the end, it shows that it is significant. No asterisk means that there is some bias. Since there is no asterisk in this regression, it shows that the model is biased and an explanatory variable or two may not be included.

3.2.5 AICc and R^2

The AIC is in this multilinear regression is 205.229. The multiple R^2 is 0.244 and the adjusted R^2 is 0.188. This does not mean it is bad or good. Essentially the R needs to be in the range of zero to one, which it falls under. Generally, a higher than 0.5 R^2 is good, but depends on the variable.

3.2.6 Spatial Autocorrelation Report

The spatial auto correlation report should report a clustering, or high concentrations of various explanatory variables are present. In this situation, the report showed a random correlation with a z-score of 1.584 and a p-value of 0.113. Not showing it leaning either way. If the p-score was high say 2.5 and the p-value was low, 0.01, this model would be clustered.

After the tests have been conducted, it seems that population from the census tract, the percentage agricultural, the percentage environmental, and the percentage urban are the most significant variables. Population affected almost all factors of the Regression Analysis Check. Although the percentage environmental and the percentage urban fall within a high VIF rating. This only means there are several other factors involved with determination and their determination.

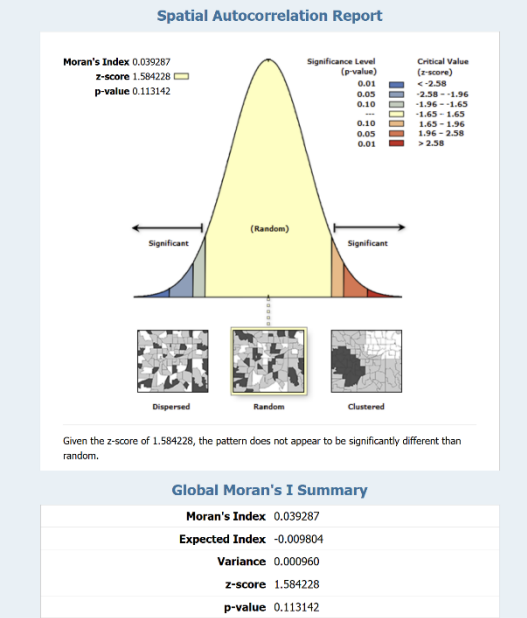


Figure 8. *The Spatial Autocorrelation Report*

3.3 Ordinary Least Squares (OLS)

OLS was run to determine the model's explanatory variables relationship to the dependent variable. OLS is a linear based regression model that is often the starting place of regression techniques. In this test, the Adjusted R^2 and the R^2 will be reviewed. The OLS regression models will be run on population of the census tract, percent agricultural area, percent environmental area, and percent urban area. Below are the results for the OLS regression model. The letter *n* represents the number of observations.

Table 4. *OLS Summaries for Training and Testing Data*

Data	**Model**	***n***	**R^2**	**Adjusted R^2**
Training Population Census Tract Training Percent Agricultural Training Percent Environmental Training Percent Urban	OLS	69	0.286	0.242
Testing Population Census Tract Testing Percent Agricultural Testing Percent Environmental Testing Percent Urban	OLS	34	0.220	0.113

For the OLS model, this was an underfit model due to there being poor training and testing data. The training data had an R^2 of 0.286 and testing data had an R^2 of 0.233. The GWR model was run to see if there was a better fit.

3.4 Geographic Weighed Regression (GWR)

GWR was run to determine the model's explanatory variables relationship to the dependent variable. GWR is a local form of linear regression model that is used to model spatial variables. In this test, the Adjusted R^2 and the R^2 will be reviewed. The GWR regression models will be run on population of the census tract, percent agricultural area, percent environmental area, and percent urban area. Below are the results for the GWR regression model. The letter *n* represents the number of observations.

Table 5. *GWR Summaries for Training and Testing Data*

Data	**Model**	***n***	**R^2**	**Adjusted R^2**
Training Population Census Tract Training Percent Agricultural Training Percent Environmental Training Percent Urban	GWR	69	0.453	0.314
Testing Population Census Tract	GWR	34	0.1776	0.0476

For the GWR model, only training data sets could be run due to multicollinearity on the testing data and the low number of observations. The training data had an R^2 of 0.314. More than likely, the data for the testing would have come back low, making it an underfit model. No assumptions can be made.

3.5 Forest-based and Boosted Classification and Regression (RFR)

RFR is a machine learning tool that generates a model based in two supervised machine leaning methods: a rando forest algorithm and XGBoost algorithm. The RFR for this model was conducted in three stages. Each stage holds a different level of trees conducted: 100 trees, 50 trees, and 25 trees. This was determined based on the presets of the RFR model. The more trees that withstand a test, the more complex the model is.

Table 6. *RFR Summaries for Training Data*

Data	Model	# of Trees	Training R^2	p-value	Testing R^2	p-value
Training Population Census Tract Training Percent Agricultural Training Percent Environmental Training Percent Urban	RFR	100	0.835	0.000	0.099	0.542
Training Population Census Tract Training Percent Agricultural Training Percent Environmental Training Percent Urban	RFR	50	0.851	0.000	0.010	0.177
Training Population Census Tract Training Percent Agricultural Training Percent Environmental Training Percent Urban	RFR	25	0.816	0.035	0.206	0.366

For the RFR model, all variations of the trees 25, 50, and 100, came back as an overfit model. For 100 trees the data the training was R^2 0.835, and p-value was 0.000. The testing was R^2, which was 0.099 and the p-value was 0.543. For 50 trees the data the training was R^2 0.851, and p-value was 0.000. The testing was R^2 was 0.010 and the p-value was 0.177. For 25 trees the training was R^2.816, and the p-value was 0.035. The testing was R^2 is 0.206, and p-value was 0.366. They all had high value for the testing and low R^2 for the training. If the p-value for the testing data would have been lower than 0.05, this would have been considered a well fit model. Due to it having a high p-value, it was considered an overfit model.

3.6 Errors

Errors within the data occurred when the GWR was run. There were too many variables (percent environmental and percent urban) that contained multicollinearity. To fix this, the GWR was run for the testing data with population. Although this does not necessarily fix the issue, it puts something into perspective when running the data. The model was still considered as an underfit model.

Several issues with the aggregation of data occurred. Due to the limited time of this research project, where to aggregate the data could not be fixed. Aggregation of data was paired with the census tracts. This was held with 214 polygons. Out of the IBI points data, holding 356 points of data, 103 points of accurate IBI points data split by testing and training where aggregated into

the 214 polygons points of the census tract. This means that the census tracts were not a good way to split the data given that there were several hundreds of points of data missing.

There are two ways in which the data could be more efficiently aggregated. First is using zonal statics for the mean of the IBI data should have been calculated and then placed into the census tracts. Although this would not have been so inclusive, the data would have been better summed up within each census tract. Though, this could also cause errors. Alternatively, the use of Thiessen polygons could have been used to interpolate the IBI variable of interest over PGCo. Th conversion of the Thiessen polygons from raster to polygons could have been interpolated. Then use zonal statistics to calculate the mean value of the Theissen polygons for each census tract.

4 Conclusion

This project investigated MBSS conducted by MDDNR. It looks at PGCo's IBI scores of streams in the entire county instead of smaller stream segments of 75-meters. It involved regression models for populations of the census tracts, percent canopy cover, percent impervious surfaces, percent agricultural areas, percent environmental areas, and percent urban areas. It used OLS and GWR which had underfit results. It used a machine learning tool, RFR, which had an overfit model. None of the data portrayed a well fit model. Current data on this topic of IBI has never been experimented with the findings of this paper, in PGCo accounting for the whole county rather than the 75-meter stream segments. The data displayed in this report has been manipulated but not to the best of abilities. Due to time limitations, this was all that could be completed.

In the future, data collection should be done on the whole state of Maryland. Since there are 23 counties in the state, this would have created much more data for the OLS, GWR, and RFR calculations. It was avoided since the canopy cover and impervious surfaces would have had to been collected for each of the counties within the state of Maryland. This would have served as a lot of data for ArcGIS Pro to process. If it took an hour and a half to process the data from PGCo, from raster to polygons, this more than likely would have taken a whole day to convert the raster to polygon for the impervious surfaces and canopy cover.

References

[1] Alberts, J. M., Beaulieu, J. J., & Buffam, I. (2017). Watershed Land Use and Seasonal Variation Constrain the Influence of Riparian Canopy Cover on Stream Ecosystem Metabolism. Ecosystems, 20(3), 553–567. https://doi.org/10.1007/s10021-016-0040-9

[2] Arnold Jr., C. L., & and Gibbons, C. J. (1996). Impervious Surface Coverage: The Emergence of a Key Environmental Indicator. Journal of the American Planning Association, 62(2), 243–258. https://doi.org/10.1080/01944369608975688

[3] Bongaarts, J. (2009). Human population growth and the demographic transition. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1532), 2985–2990. https://doi.org/10.1098/rstb.2009.0137

[4] Cross Section Data. (n.d.). Retrieved March 29, 2025, from https://www.hec.usace.army.mil/confluence/rasdocs/rasum/6.5/entering-and-editinggeometric-data/cross-section-data

[5] Dale, V. H. (1997). The Relationship Between Land-Use Change and Climate Change. *Ecological Applications*, *7*(3), 753–769. https://doi.org/10.1890/1051-0761(1997)007[0753:TRBLUC]2.0.CO;2

[6] Hession, W. C., Pizzuto, J. E., Johnson, T. E., & Horwitz, R. J. (2003). Influence of bank vegetation on channel morphology in rural and urban watersheds. Geology, 31(2), 147–150. https://doi.org/10.1130/0091-7613(2003)031<0147:IOBVOC>2.0.CO;2

[7] Klauda, R., Kazyak, P., Stranko, S., Southerland, M., Roth, N., & Chaillou, J. (1998). Maryland Biological Stream Survey: A State Agency Program to Assess the Impact of Anthropogenic Stresses on Stream Habitat Quality and Biota. Environmental Monitoring and Assessment, 51(1/2), 299–316. https://doi.org/10.1023/A:1005903822990

[8] Lessard, J. L., Cheng, M. S., Akinbubola, C., Stribling, J. B., & Leppo, E. W. (2012). Relationships Among Land-Use, In-Stream Stressors, and Biological Condition in Prince George's County, MD. 1–10. https://doi.org/10.1061/40856(200)419

[9] Metre, P. C. V., Waite, I. R., Qi, S., Mahler, B., Terando, A., Wieczorek, M., Meador, M., Bradley, P., Journey, C., Schmidt, T., & Carlisle, D. (2019). Projected urban growth in the southeastern USA puts small streams at risk. *PLOS ONE*, *14*(10), e0222714. https://doi.org/10.1371/journal.pone.0222714

[10] Meyer, W. B., & Turner, B. L. (1992). Human Population Growth and Global Land-Use/Cover Change. Annual Review of Ecology and Systematics, 23, 39–61. https://www.jstor.org/stable/209721

[11] Moglen, G. E. (2009). Hydrology and Impervious Areas. Journal of Hydrologic Engineering, 14(4), 303–304. https://doi.org/10.1061/(ASCE)1084-0699(2009)14:4(303)

[12] Prince George's County, Maryland. (2025). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Prince_George%27s_County,_Maryland&oldid=1285740946

[13] *Prince George's County, MD population by year, race, & more*. (2025, May 6). USAFacts. https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/state/maryland/county/prince-georges-county/

[14] Pumhirunroj, B., Littidej, P., Boonmars, T., Bootyothee, K., Artchayasawat, A., Khamphilung, P., & Slack, D. (2023). Machine-Learning-Based Forest Classification and Regression (FCR) for Spatial Prediction of Liver Fluke Opisthorchis viverrini (OV) Infection in Small Sub-Watersheds. *ISPRS International Journal of Geo-Information*, *12*(12), 503. https://doi.org/10.3390/ijgi12120503

[15] Riparian zone. (2025). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Riparian_zone&oldid=1270927044

[16] River morphology. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=River_morphology&oldid=1237480481

[17] Rosgen, D. (n.d.). A STREAM CHANNEL STABILITY ASSESSMENT METODOLOGY. https://wildlandhydrology.com/resources/docs/Assessment/Rosgen_2001_Channel_Stability.pdf

[18] Sheehan, K. R. (2011). *Exploration of Stream Habitat Spatial Modeling; Using Geographically Weighted Regression, Ordinary Least Squares Regression, and Natural Neighbor Interpolation to Model Depth, Flow, and Benthic Substrate in Streams* [Ph.D., West Virginia University]. https://www.proquest.com/docview/1113338594/abstract/9F375419629A4413PQ/1

[19] Springer, D. (1999). Ecosystem metabolism. In: Environmental Geology. Encyclopedia of Earth Science. https://doi.org/10.1007/1-4020-4494-1_99

[20] Stribling, J. B. (2024). Landscape changes and watershed erosion in Prince George's County, Maryland. River Research and Applications, 40(7), 1314–1342. https://doi.org/10.1002/rra.4292

[21] U.S. Census Bureau. (n.d.). *Prince George's County, ... - Census Bureau Profiles Results*. Retrieved May 6, 2025, from https://data.census.gov/profile?q=Prince+George%27s+County,+Maryland+Tolowa

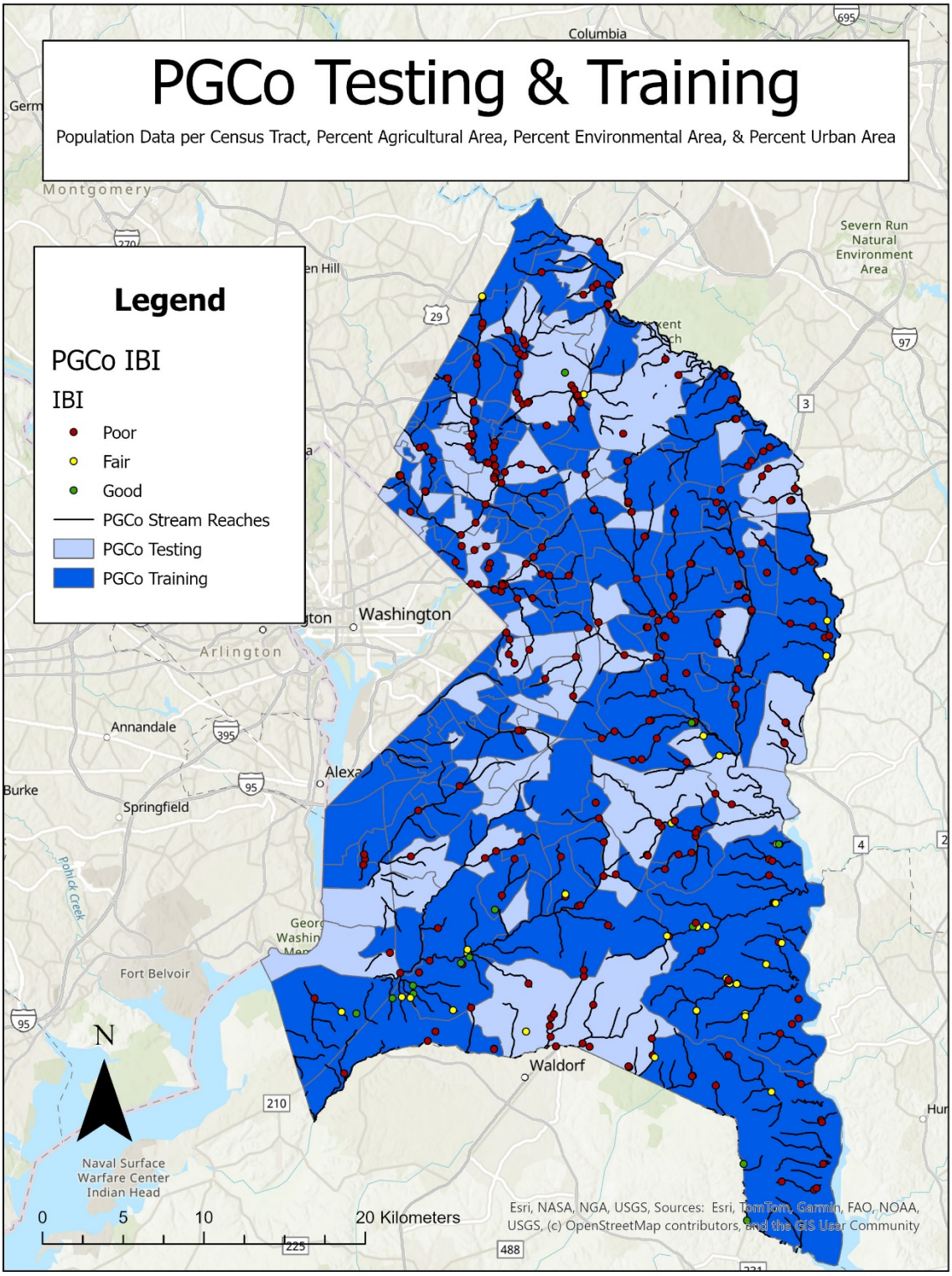
[22] Virta, A. (n.d.). *A County with Rich History: Prince George's County History*. Prince George's County Historical Society. https://www.pghistory.org/PG/PG300/history.html

Appendix A. Acronyms

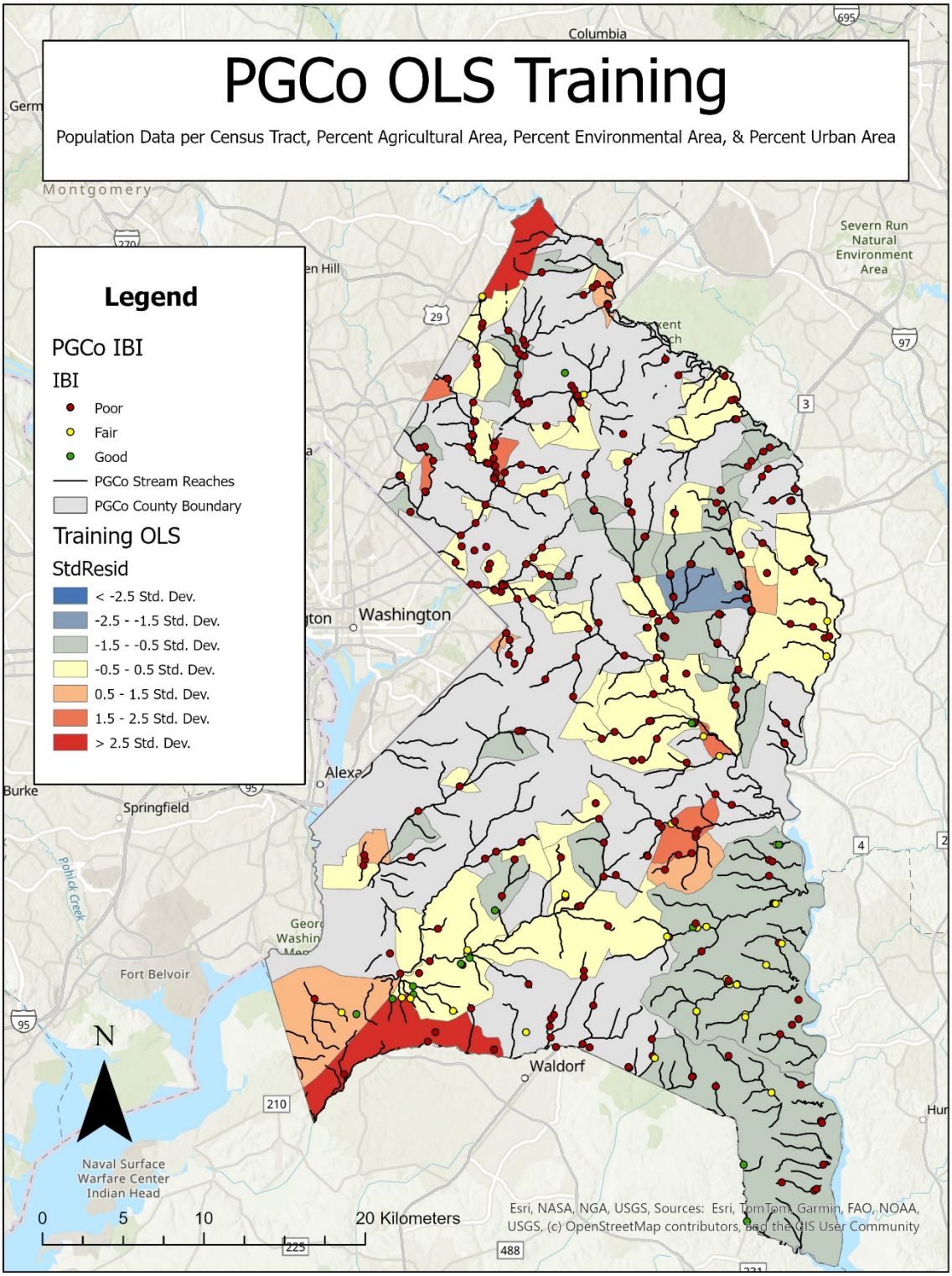
AOI	Area of Interest
GDB	Geodatabase
GWR	Geographic Weighted Regression
IBI	Indices of Biological Integrity
MBSS	Maryland Biological Stream Survey
MLS	Multi-Linear Regression
OLS	Ordinary Least Squares
PGCo	Prince George's County
RFR	Forest-based and Boosted Classification and Regression
VIF	Variance Inflation Factor

Appendix B. Maps of Models

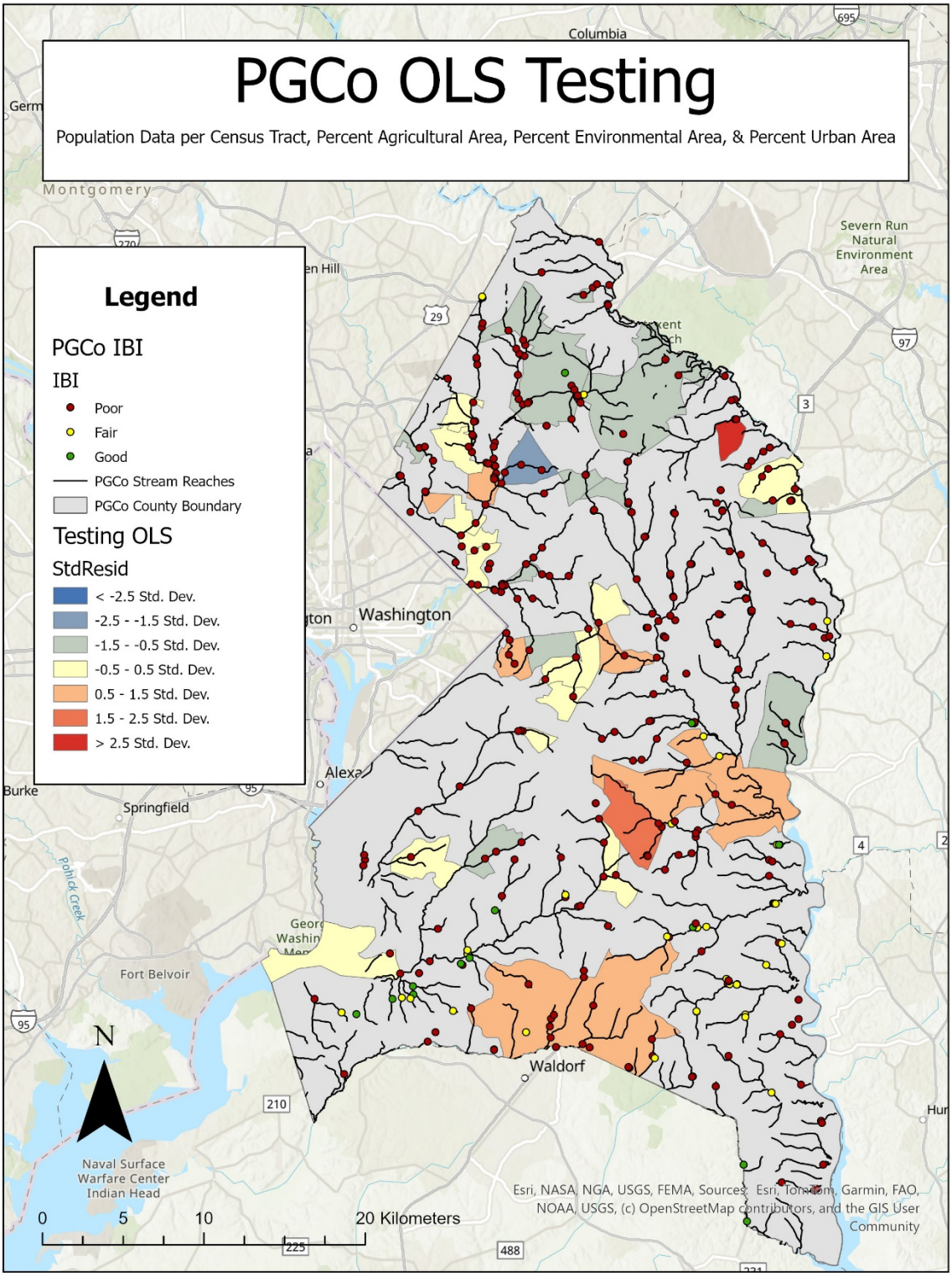
1. Testing and Training Data



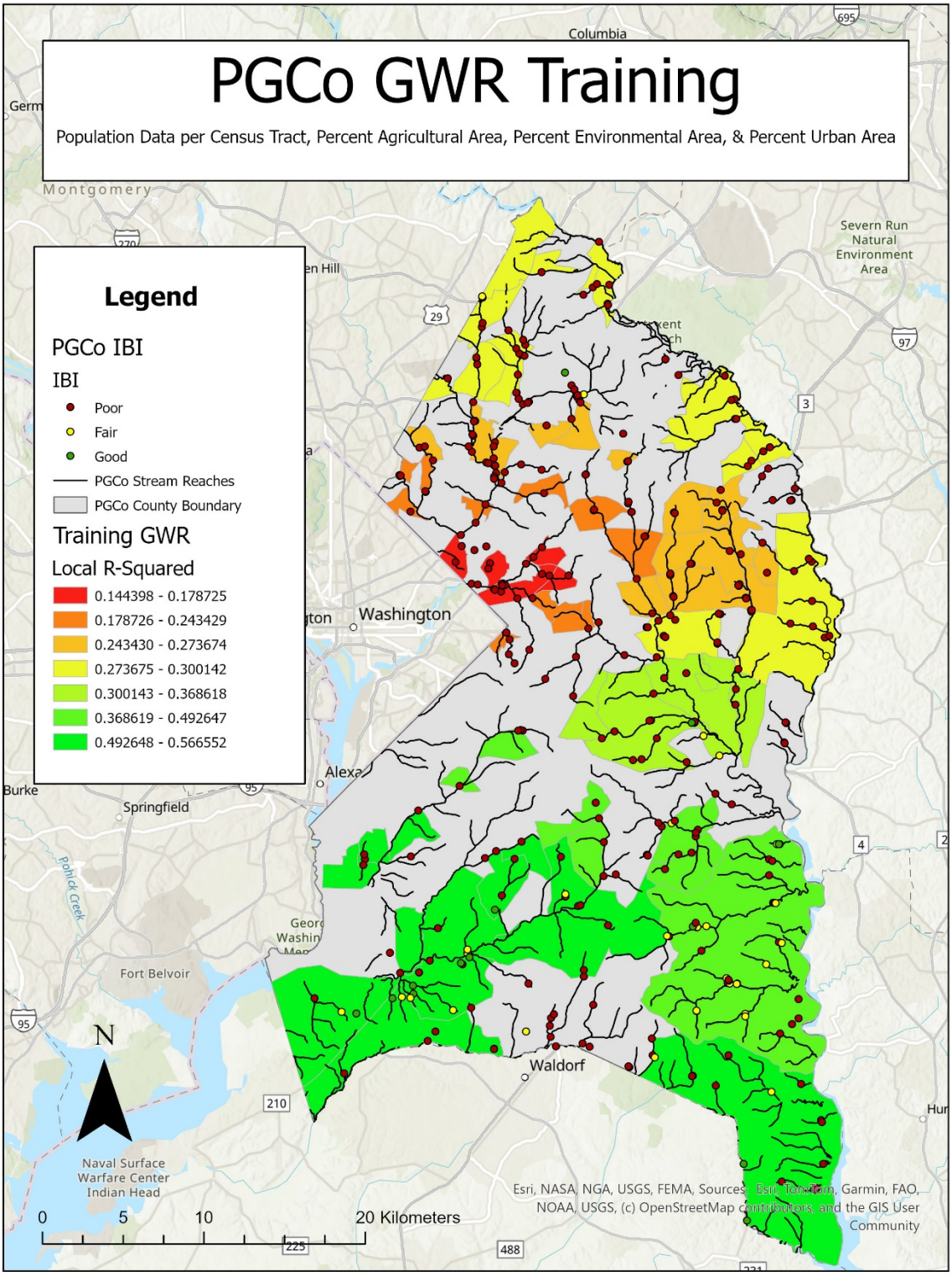
2. Training OLS: Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area



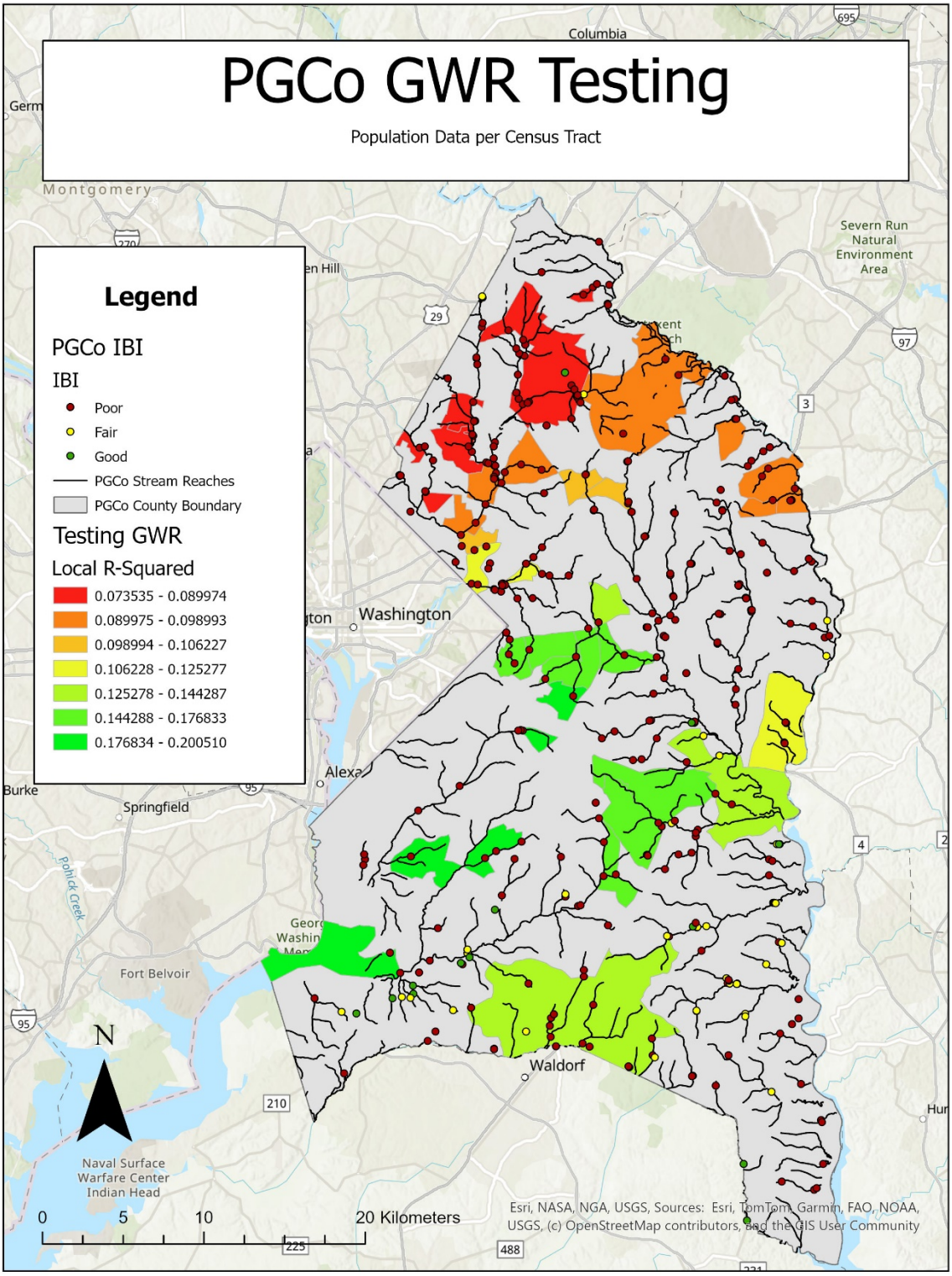
3. Testing OLS: Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area



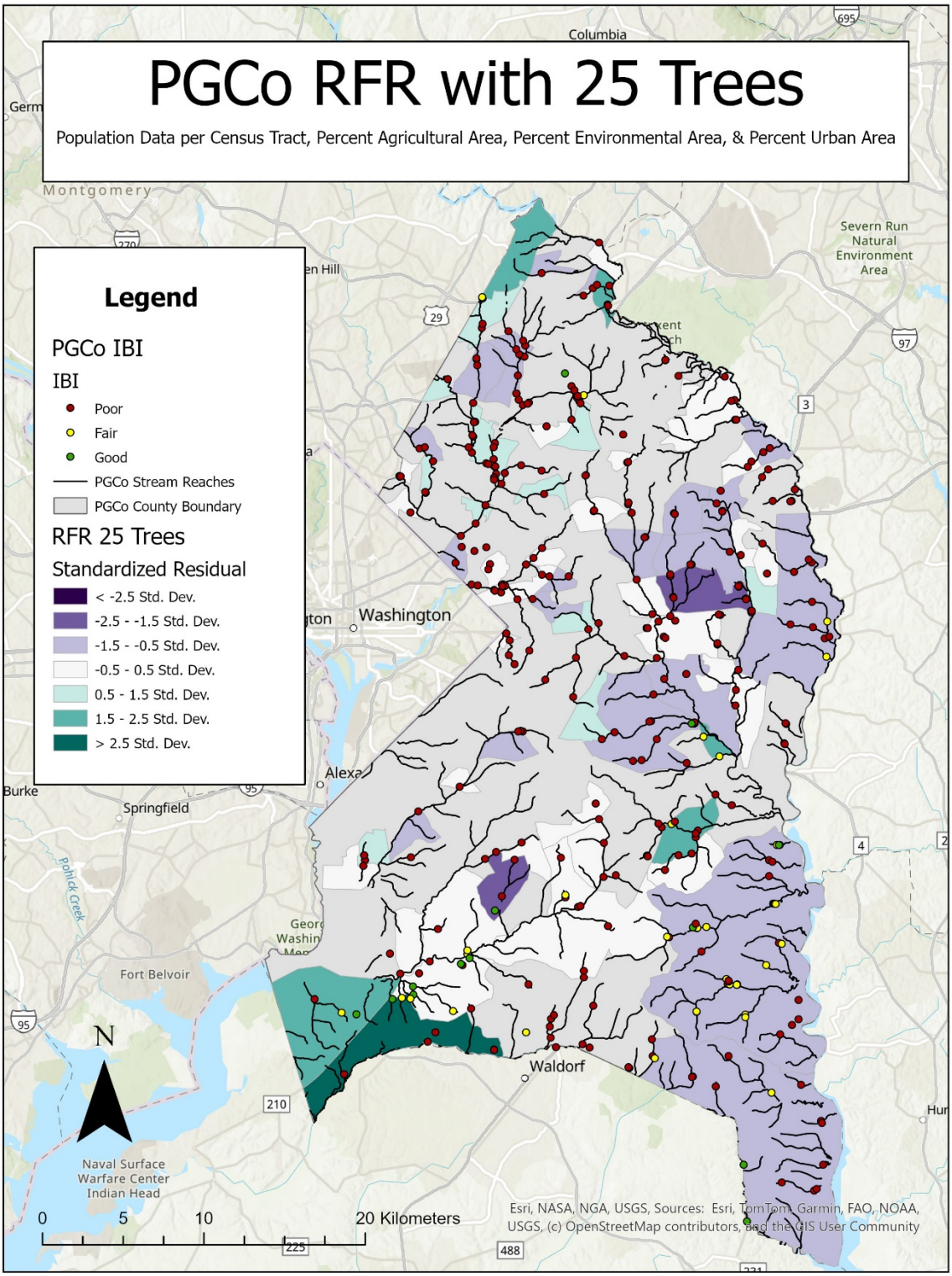
4. Training GWR: Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area



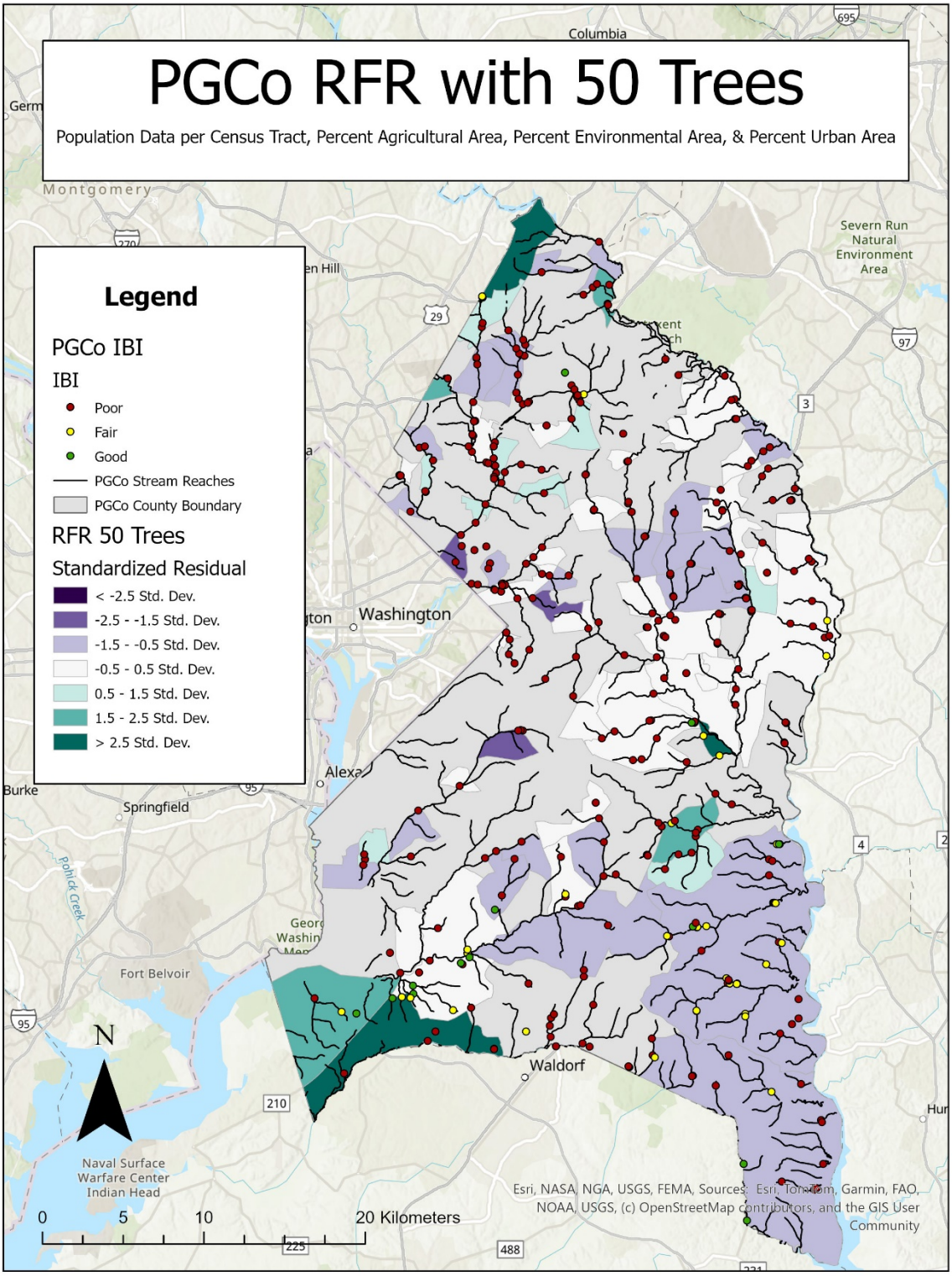
5. Testing GWR: Population Census Tracts ONLY Multicollinearity of Percents of Areas



6. RFR Machine Learning: 25 Trees of Training Data for Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area



7. RFR Machine Learning: 50 Trees of Training Data for Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area



8. RFR Machine Learning: 100 Trees of Training Data for Population Census Tracts and Percentage of Agriculture, Environmental, and Urban Area