Projection of Permanent Forest Loss by Random Forest

Final project for GEOG788P

Models and Methods for Spatial Data Science

Quan Shen

**Introduction**

As the anthropogenic carbon emissions have caused climate change, the U.S. Climate Alliance (USCA), a coalition of 24 states committed to achieving the emissions reductions outlined in the 2015 Paris Agreement, is interested in better science to improve the forest carbon components of greenhouse gases inventories. The forest has been a significant carbon sink and a carbon source at the same time (Hurtt et al. 2002; van der Werf et al. 2009). A clear understanding of the forest carbon plays an indispensable roll in reducing carbon emissions.

My faculty advisor, Dr. George Hurtt and his team are using the Ecosystem Demography model (ED), which simulates the physiological processes of individual vegetation (Hurtt et al. 2002), to map and project future forest carbon stocks in Maryland.

However, a flaw of the model is that once initialized, it assumes that vegetation can always regrow on their original location. Nevertheless, the permanent forest loss exists. For example, there are new impervious surfaces every year for urban sprawl (Johnson 2001), and trees cannot grow on those anymore. The accuracy of the model would increase if it takes permanent forest loss into consideration. Therefore, my goal of this project is to use data science methods to project permanent forest loss in Maryland for ED, to increase the model accuracy.

**Literature review**

Although there are a number of papers on forest loss, I only found one on projecting forest loss. Thompson et al. (2017) used a cellular land change model, landcover maps generated from 30m Landsat data by Olofsson et al. (2016), and a few other datasets containing information from 1990-2010 to project forest loss in New England from 2010-2060. Thompson et al. (2017) concluded that distance to developed land, distance to road, and population density are the three strongest predictors of forest loss. The pressure of forest loss from other land cover types has been confirmed a few studies (Curtis et al. 2018; Drummond & Loveland 2010; Johnson 2001).

Even though the work from Thompson et al. (2017) is not the same from my goal of projecting permanent forest loss in Maryland, their work provided some important insights for me. I could use the strongest predictors directly to project my permanent forest loss.

**Data**

Since remote sensing raster data have easy access, a wide range of related products, as well as abundant processing tools and algorithms (Liu 2015), I would like to use raster data for this project. I used landcover datasets from National Land Cover Database (NLCD) to determine the forest loss and to assess the influence of neighboring landcover to forest loss. The distance to developed land and distance to road are implied in landcover datasets. NLCD datasets contains the landcover information for the continental US. The raster datasets are generated from Landsat data and thus have the 30m resolution. The pixel value is the encoded landcover type. Table 1 describes the legend. NLCD publishes a new landcover dataset every two to three years. For this project I used 2004, 2006, 2008, 2011, 2013, and 2016, six landcover datasets.

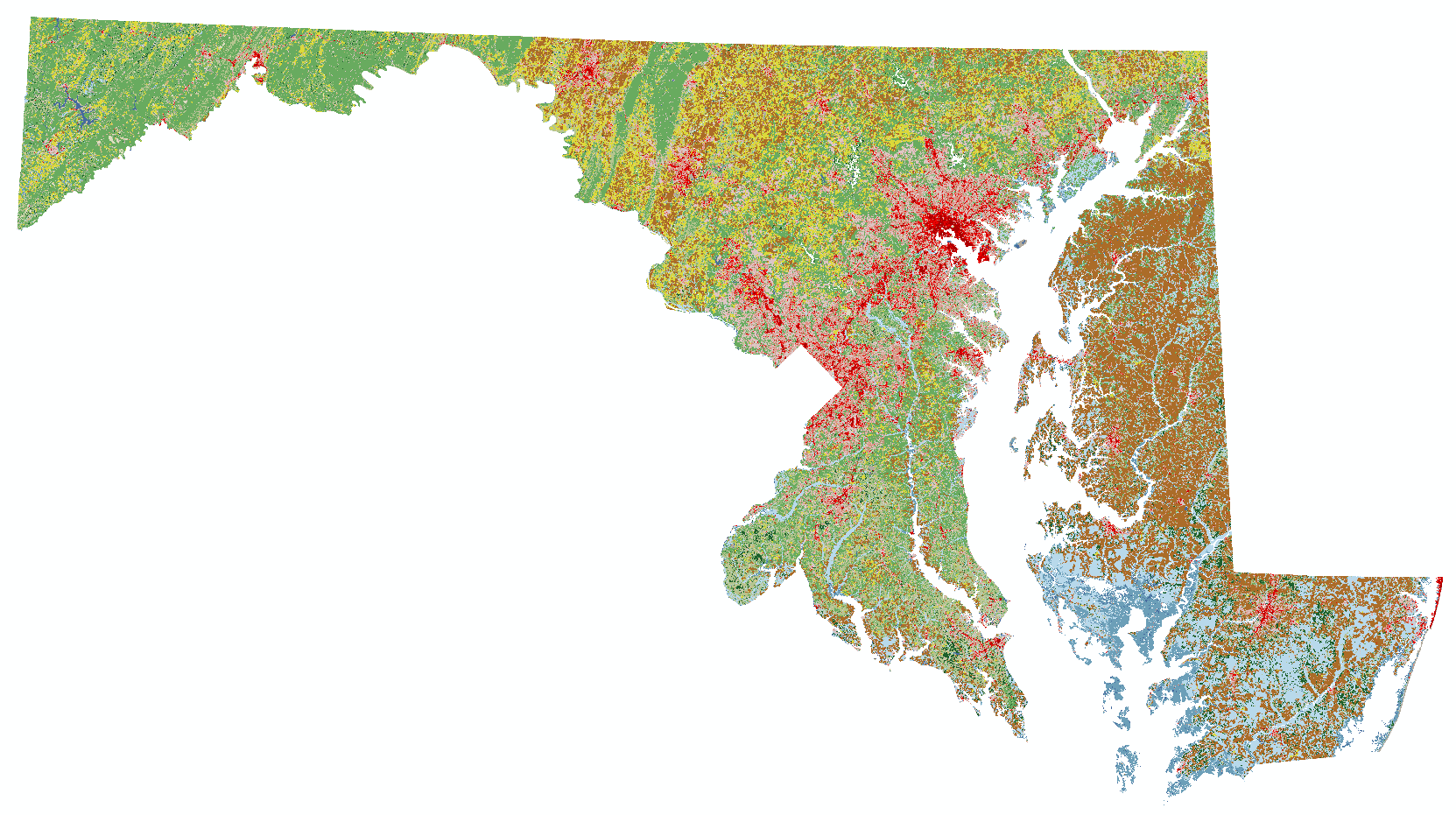
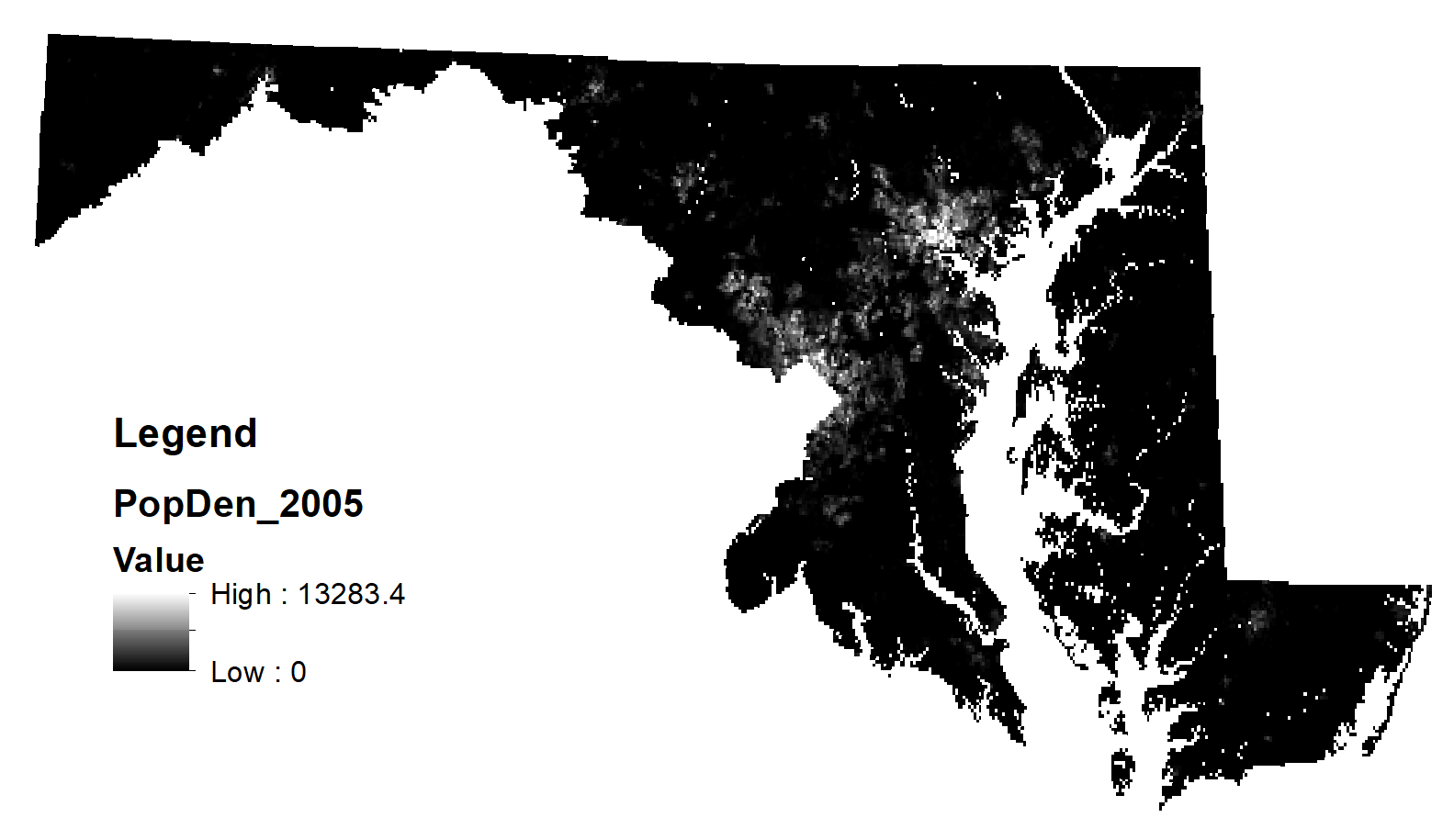
|  |  |  |
| --- | --- | --- |
| Landcover Code | Landcover Description | Landcover Category |
| 11 | Open water | Water |
| 12 | Perennial Ice/Snow |
| 21 | Developed, Open Space | Developed |
| 22 | Developed, Low Intensity |
| 23 | Developed, Medium Intensity |
| 24 | Developed, High Intensity |
| 31 | Barren Land (Rock/Sand/Clay) | Barren |
| 41 | Deciduous Forest | Forest |
| 42 | Evergreen Forest |
| 43 | Mixed Forest |
| 51 | Dwarf Scrub | Shrubland |
| 52 | Shrub/Scrub |
| 71 | Grassland/Herbaceous | Herbaceous |
| 72 | Sedge/Herbaceous |
| 73 | Lichens |
| 74 | Moss |
| 81 | Pasture/Hay | Cultivated |
| 82 | Cultivated Crops |
| 90 | Woody Wetlands | Wetlands |
| 95 | Emergent Herbaceous Wetlands |

Table 1. Legend of NLCD Landcover datasets.

Source: <https://www.mrlc.gov/data/legends/national-land-cover-database-2016-nlcd2016-legend>

The population density datasets used in this project are from NASA's Earth Observing System Data and Information System. They describe global population density, and the resolution is 1km. The pixel value is the population density of the pixel size. I used 2005, 2010, and 2015, three years of data.

As the two kinds of datasets are in different projections, different resolutions, and cover different extent, I used ArcGIS to project, clip, and resample them, resulting 30m resolution raster files in UTM zone 18 North projection, covering the area of Maryland. Figure 1a and 1b show the preprocessed NLCD 2004 and population density 2005 files respectively.

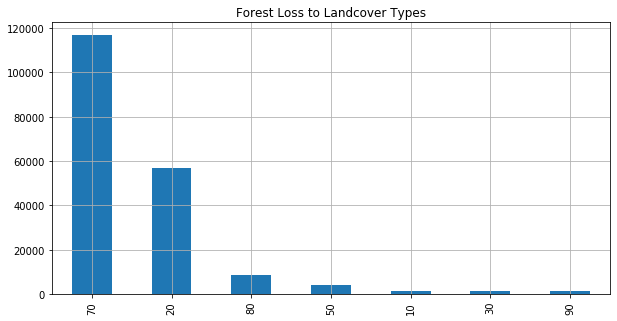
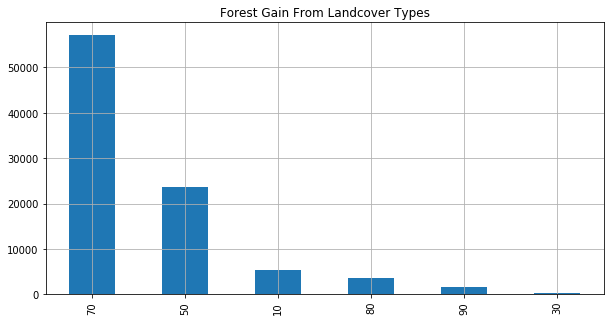
 

a. NLCD 2004 b. Population Density 2005

Figure 1. Raster files with the same extent, resolution, and projection after preprocessing

**Method**

To project the permanent forest loss, I first identified which landcover change indicates permanent forest loss. Using the six NLCD datasets, I was able to track the landcover change in 2004 to 2016. For easier calculation and better visualization, I binned the landcover codes by 10. For example, the landcover code for developed area is 20 and forest is 40, only keeping the main categories. If a pixel was forest in 2004 but changed to non-forest later and stayed non-forest until 2016, the pixel was classified as semi-permanent loss and the landcover type after forest was recorded. It is semi-permanent loss because the pixel still has a chance to become forest in the future. When a pixel experienced forest loss but became forest again by 2016, it is non-permanent loss and the landcover type before the forest was recorded. Landcover types that only have forest lost to but no forest gain from indicate that the conversion from forest to them is irreversible. Therefore, the permanent forest loss could be identified by comparing the landcover types that forest lost to and forest gained from.

a. Landcover types that forest loss to b. Landcover types that forest gain from

Figure 2. The x-axis is the binned landcover types, while the y-axis is the frequency of each type

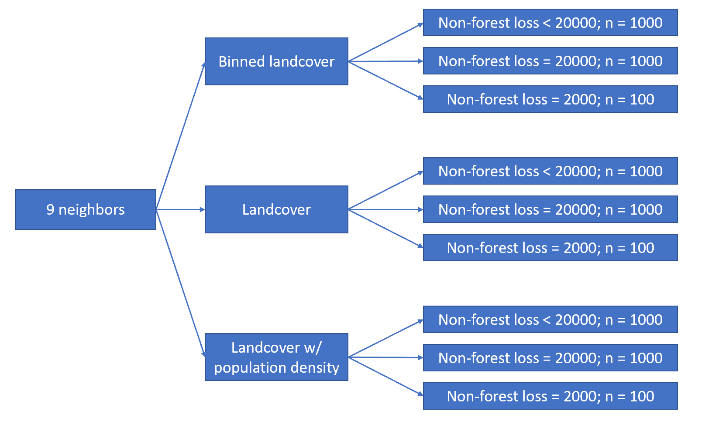
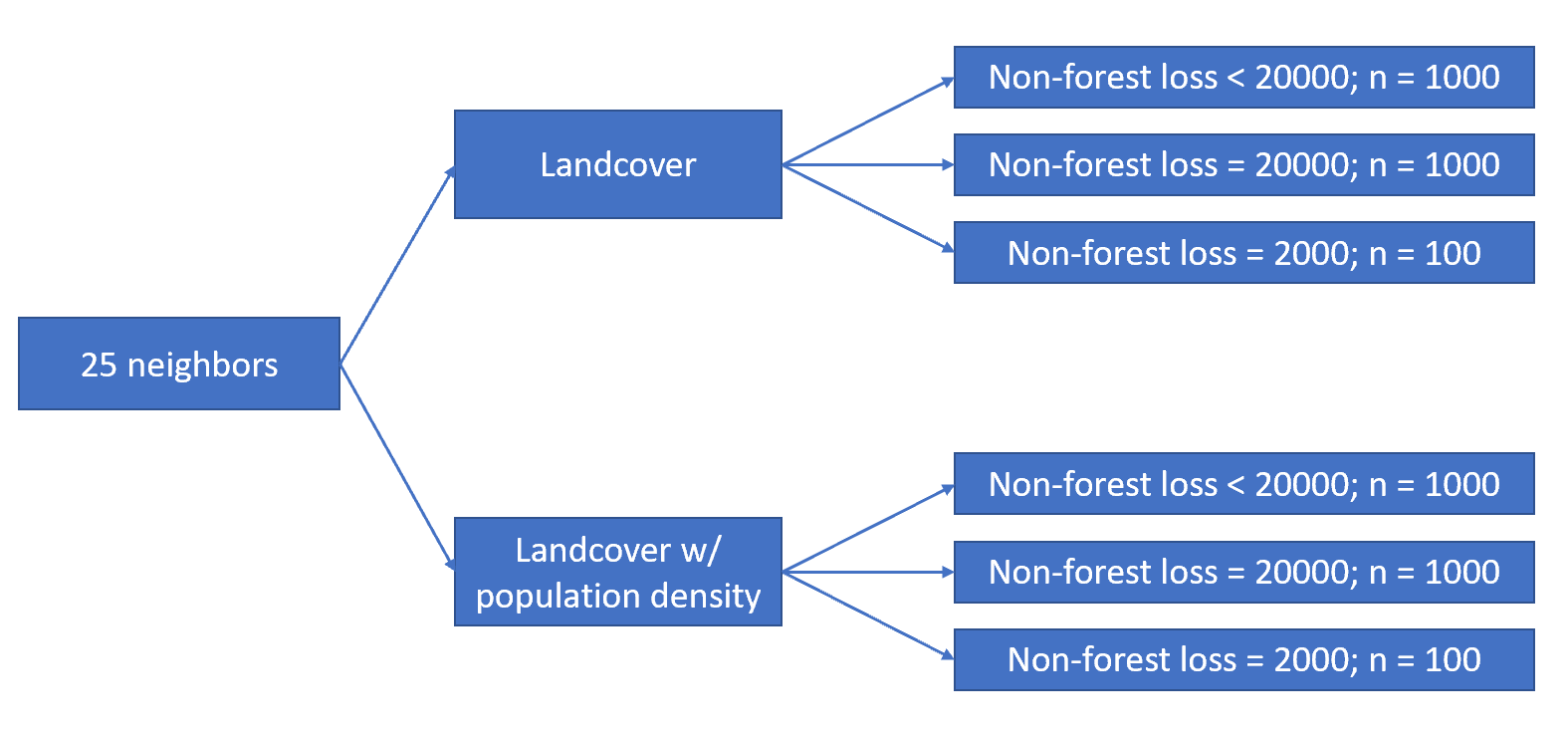
The landcover types that forest lost to and gained from are shown in Figure 2a and 2b above. All the landcover codes in 2a are in 2b except for “20”, which represents developed land. Thus, we could conclude that forest lost to developed land is permanent. The conclusion is reasonable because trees cannot grow on impervious surfaces.

Next, I used NLCD 2004 and 2006 files to create a binary raster file which illustrate the spatial location of forest to developed land conversion, the permanent forest loss, in the two years by python. Pixels of permanent forest loss have value 1, while others have value 0. Another four loss files were created using the same method. However, there are only three files among the five files containing permanent forest loss pixels. They are the loss in 2005-2006, 2009-2011, and 2014-2016.

Then I applied both non-spatial and spatial regressions to determine whether landcover types and population density have influences on the permanent loss. The permanent forest loss is the dependent variable, and the landcover type and population density are the independent variables. I used Ordinary Least Squares (OLS) model for the non-spatial regression, and spatial lag model and spatial error model for the spatial regressions. A major advantage of spatial models is that they include the spatial effects by adding a spatial weight matrix to the calculation, and thus considering spatial dependence and spatial heterogeneity (Haining 2013; Rey et al.). The spatial lag model calculates the spatial dependance by including the spatially lagged dependent variable into the regression, to assess the influence of the dependent variable to its neighboring area (Haining 2013; Rey et al.). The spatial error model, on the other hand, includes a spatially lagged error term to estimate the spatial heterogeneity (Rey et al.). To build a dataset that contains the landcover, population density, and forest loss variables, I convert the raster files into shapefiles and then spatially joined them together. The spatial weight matrix was defined by the queen contiguity. The variables were standardized.

To further improve the results, I applied a Geographically Weighted Regression (GWR) to the dataset with the three variables. Different from the previous spatial regressions, GWR can better capture the spatial heterogeneity by constructing local spatial relationships, assigning larger weights to closer features and smaller weights to further features (Fotheringham et al. 2017; Oshan et al. 2019). I also tried to apply a Multiscale Geographically Weighted Regression (MGWR), which is more advanced than GWR by allowing each variable to have a different spatial scale (Fotheringham et al. 2017; Oshan et al. 2019). But unfortunately, my laptop was not able to handle it due to the limited memory space.

Finally, I used random forest technique on the dataset to project the permanent forest loss by population density and neighboring landcover types. However, due to the limited memory space of my laptop, I had to reduce my study area and only focused on Prince George’s County in MD. The training data are the 2006, 2011, and 2016 datasets. The forest loss file between 2005-2006 was paired with NLCD 2004 and population density 2005 raster files; forest loss 2009-2011 was paired with NLCD 2008 and population density 2010; and forest loss 2014-2016 was with NLCD 2013 and population density 2015. The test data were NLCD 2016 and population density 2015. The goal was to project the permanent forest loss between 2017-2019. I repeatedly ran the model on different scenarios of the training data with various parameters to get the best result (Fig. 3). In the first scenario, I defined the landcover neighborhood as 9 neighboring pixels, and then used the forest loss raster files with three different data: the binned NLCD data only, the original NLCD data only, and the combination of the original NLCD data and the population density data. For each situation, I further modified the length of the training data and the number of trees in the random forest model. With the same number of the permanent forest loss pixels, the training data had either less than 20,000 or exactly 20,000 pixels of other landcover types. The model has either 100 or 1,000 trees. Based on the results from the first scenario, I built the second scenario in which the neighborhood was defined by 25 neighbors. For this scenario I only kept the datasets which had good performance in the last one. Figure 3a and 3b below illustrate the different random forest runs. After running the model, I validated the results using the forest loss between 2017-2019 in the Global Forest Watch (GFW) dataset describing forest loss by year (Hansen et al. 2013).

a. The first scenario b. The second scenario

Figure 3. Random forest runs

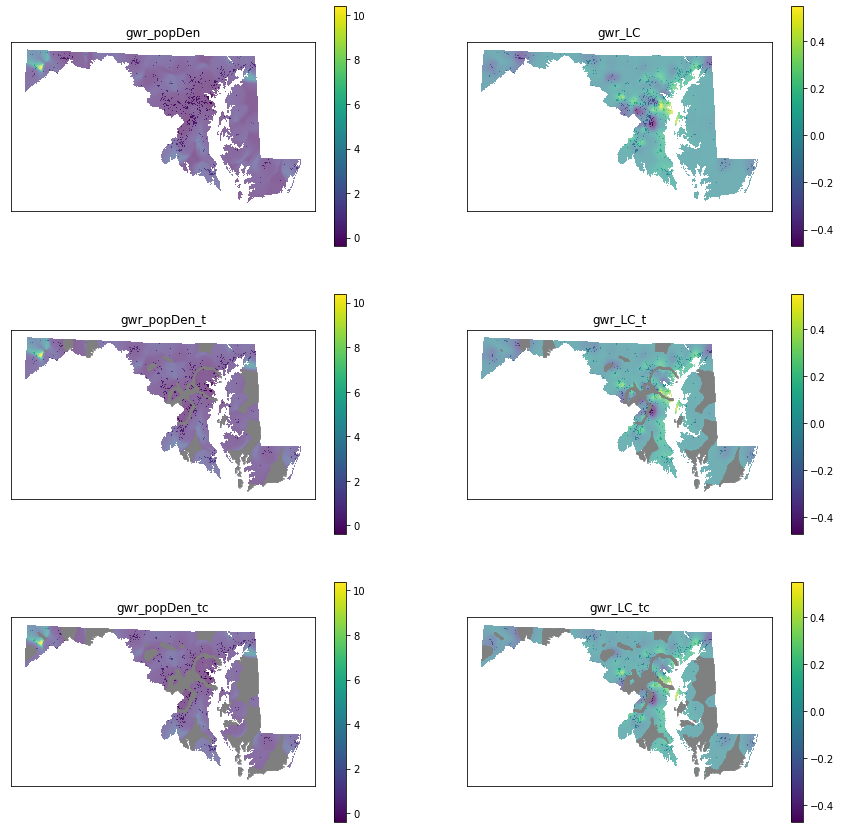
**Results**

Table 2 below shows the results from the non-spatial and spatial regressions. All the variables are statistically significant. The landcover has slightly negative effects on the forest loss, implying that a land with a smaller landcover code might be more likely to have permanent forest loss, while the codes for developed lands are 20’s, rather small codes. The population density has positive effects on the forest loss, as areas of higher population density might need more other types of land use than forest (Thompson et al. 2017). The spatial lag model exhibits large spatial dependence of the forest loss. The forest loss of a location has strong positive effects on its neighboring areas. The spatial error model reveals large spatial heterogeneity of forest loss, indicating that there are omitted variables contributing to forest loss. However, the R2 values of the OLS and the spatial error models are very low, so the two models fail to explain most variation of the forest loss. The spatial lag model has a much higher R2 value, and thus produces the most accurate estimations among the three.

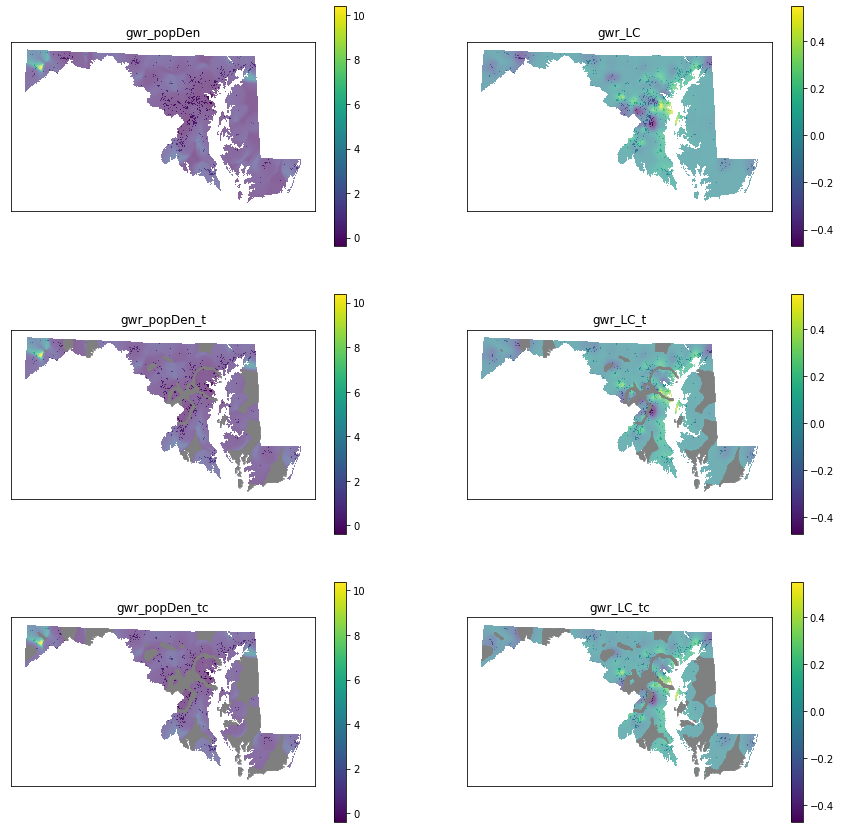
|  |  |  |  |
| --- | --- | --- | --- |
|  | OLS | Spatial lag | Spatial error |
| Intercept | 0 | -0.018 | -0.082 |
| Landcover | -0.051 | -0.009 | -0.0105 |
| Population density | 0.274 | 0.033 | 0.064 |
| ρ |  | 0.890 |  |
| λ |  |  | 0.802 |
| R2 | 0.0783 | 0.6147 | 0.0783 |

Table 2. Estimations from OLS, spatial lag, and spatial error models

Figure 4 is the GWR results of the landcover and the population density variables. The two graphs of Fig. 4a contain all local estimations without filtering out the statistically insignificant ones. The graphs of Fig. 4b masks out statistically insignificant values at the 5% significance level. The R2 value for the GWR is 0.4.



a. General results of GWR



b. Statistically significant results at the 5% significance level

Figure 4. Results of GWR on population density (left) and landcover types (right)

In Figure 4 we could see that the GWR successfully shows the location-specific spatial heterogeneity of the two variables, especially the landcover. Some neighborhoods of landcover types have negative effects on the forest loss and some have positive effects. This confirms the spatial dependence of the forest loss detected in the spatial lag model. Together, the models suggest that the neighboring landcover types have influences on forest loss. The population density has positive effects with larger magnitude than the landcover, which is the same as the results of OLS, spatial lag, and spatial error models above.

Table 3 illustrates the random forest results on the first scenarios, in which each neighborhood includes 9 pixels. The first two rows describe the training data and the decision trees in the random forest. The next row shows the number of accurately projected forest loss. The “Projected” row is the total number of permanent forest loss projected by the model, while the “GFW” row is the reference data, containing the number of forest loss detected in 2017-2019 by GFW. The last two rows are accuracies calculated using the three rows right above them.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Binned landcover | | | Original landcover | | | Original landcover with population density | | |
| # of non-forest-loss | <20000 | =20000 | =20000 | <20000 | =20000 | =20000 | <20000 | =20000 | =20000 |
| # of trees | 1000 | 1000 | 100 | 1000 | 1000 | 100 | 1000 | 1000 | 100 |
| Overlap | 940 | 297 | 269 | 1207 | 700 | 643 | 1079 | 1616 | 1775 |
| Projected | 6159 | 3573 | 3509 | 6580 | 4748 | 4476 | 6003 | 8056 | 8346 |
| GFW | 7617 | 7617 | 7617 | 7617 | 7617 | 7617 | 7617 | 7617 | 7617 |
| User’s accuracy | 15.26% | 8.3% | 7.67% | 18.34% | 14.74% | 14.36% | 17.98% | 20.06% | 21.27% |
| Producer’s accuracy | 12.3% | 3.9% | 3.53% | 15.8% | 9.2% | 8.4% | 14.16% | 21.2% | 23.3% |

Table 3. Random forest results on the first scenario: neighborhood = 9 pixels

In Table 3, the accuracies of the random forest model results are not very satisfying. The models using binned landcover data produced the least accurate results, and the models using both the original landcover data and the population density data produced the most accurate results, with the user’s accuracy of 21.27% and the producer’s accuracy of 23.3%. Among the models using both data, the model including 20,000 non-forest-loss landcover pixels and 100 trees has the highest accuracies. However, among the other models, the ones with less than 20000 non-forest-loss pixels and 1,000 trees produced better results.

I would like to check if a larger neighborhood could output better outcomes. Since the models with the binned landcover data did not have a good performance, I removed them in my second scenario, in which a neighborhood has 25 pixels. Table 4 presents the model results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Original landcover | | | Original landcover with population density | | |
| # of non-forest loss | <20000 | =20000 | =20000 | <20000 | =20000 | =20000 |
| # of trees | 1000 | 1000 | 100 | 1000 | 1000 | 100 |
| Overlap | 1454 | 700 | 698 | 111 | 35 | 35 |
| Projected | 7220 | 4581 | 4468 | 1221 | 813 | 733 |
| GFW | 7617 | 7617 | 7617 | 7617 | 7617 | 7617 |
| User’s accuracy | 20.14% | 15.28% | 15.62% | 9.09% | 4.3% | 4.77% |
| Producer’s accuracy | 19.09% | 9.2% | 9.16% | 1.46% | 0.46% | 0.46% |

Table 4. Random forest results on the first scenario: neighborhood = 25 pixels

Unexpectedly, the accuracies of the models using both original landcover the population density are lower than the models only using the landcover data, and the lowest in both scenarios. The models using only the landcover data in the second scenario are only slightly better than those using the same data in the first scenario. In this scenario, the two models with less than 20000 non-forest-loss samples and 1,000 trees produced better results. However, more evidence is needed to assess the influences of the number of non-forest-loss samples and the number of trees on the model accuracy.

**Discussion**

To determine the causes of the unsatisfactory projection results, I compared the loss files created from NLCD datasets to the GFW forest loss dataset. The comparison is in Table 5. Although the user’s accuracy is much higher than the projection results, the producer’s accuracy remains low. Therefore, the GFW dataset is not suitable as the reference data. Besides the statistics shown in the table, another reason to not use GWF is that it records all kinds of forest loss, including both non-permanent and permanent loss. However, the models only project permanent forest loss. The ideal reference data should be the permanent loss file made from NLCD 2016 and the next NLCD landcover dataset. Since NLCD publishes new landcover dataset in every two to three years, the next landcover dataset is expected to be available soon.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Loss & GFW 05-06 | Loss & GFW 09-11 | Loss & GFW 14-16 |
| Overlap | 742 | 568 | 756 |
| Loss | 806 | 2363 | 795 |
| GFW | 11998 | 4255 | 5809 |
| User’s accuracy | 92.05% | 24.04% | 95.1% |
| Producer’s accuracy | 6.18% | 13.3% | 13% |

Table 5. Comparison between loss files and GFW

The limitations of the NLCD landcover and the population density datasets could be another cause of the disappointing projections. The NLCD data are categorical data in 30m resolution, meaning that NLCD only captures one landcover type over the 900 sq m area. Thus, the forest change could be neglected in mixed pixels. For example, if there is a pixel classified as developed land since 2006, it had 80% developed land and 20% forest at the beginning and then gradually lost the forest later. Same thing for pixels of 100% forest but lost a portion later. Although the pixel classification does not change, the amount of forest changed but we cannot detect the change using NLCD. Moreover, the resolution of the population density datasets is 1km, even coarser. The coarse resolution does not exhibit the local variation of the population density. For example, if a populated neighborhood covers a half of a pixel and woods cover the other half, the pixel cannot show the difference. Therefore, datasets with higher resolution could help improve the projection. Having more variables could also make a difference.

**Conclusion**

For this project, I used the NLCD landcover datasets and population density datasets to project the future permanent loss. First, I identified the landcover conversion from forest to developed land as permanent forest loss. Then, the permanent loss datasets were created by such landcover conversion.

Next, I used both non-spatial and spatial regression models to determine the effects of landcover and population density on forest loss. The spatial lag model performed better than the OLS and the spatial error model. It suggested that the forest loss has spatial dependence, having positive influences on the neighboring areas. The results of GWR indicated that landcover types had local influences on forest loss. Combining the results from the spatial lag model and the GWR, we could state that the landcover types of neighboring pixels have impacts on forest loss. All the regression models agreed that population density had positive effects on permanent loss.

Lastly, I used random forest multiple times to find the best projection. The projection results showed that more neighbors, non-forest-loss samples, and/or more decision trees do not necessarily produce better results. The best result was generated using both landcover data and population density data, with 9 neighbors, 20000 non-forest-loss samples, and 100 decision trees. However, even the best projection did not have satisfactory accuracies, due to the unsuitability of the reference data as well as the limitations of the input data.

I am looking forward to the new NLCD landcover dataset that will be published in the next year. When it is available, I would use it to create a new permanent forest loss file, which will be the ideal reference data of my projection. With the new dataset, I could also add a new variable which describes the landcover change history of each pixel in the random forest model. Other spatial variables such as slope and aspect will be considered if datasets are available. Besides NLCD, I would like to find other landcover datasets with higher spatial resolution, same for the population density datasets. Hopefully I will get a satisfactory projection result.

**References**

In-class references:

1. Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6), 1247–1265. https://doi.org/10.1080/24694452.2017.1352480
2. Haining, R. (2013). Spatial Data and Statistical Methods: A Chronological Overview. *Handbook of Regional Science*, 1277–1294. <https://doi.org/10.1007/978-3-642-23430-9_71>
3. Liu, P. (2015). A survey of remote-sensing big data. *Frontiers in Environmental Science*, 3. <https://doi.org/10.3389/fenvs.2015.00045>
4. Oshan, T., Li, Z., Kang, W., Wolf, L., & Fotheringham, A. (2019). Mgwr: a python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *Isprs International Journal of Geo-Information*, 8(6), 269–269. <https://doi.org/10.3390/ijgi8060269>
5. Rey, S. J., Arribas-Bel, D., & Wolf, L. J. Spatial Regression. *Jupyter Notebook*. <https://geographicdata.science/book/notebooks/11_regression.html>.

Additional references:

1. Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A., & Hansen, M. C. (2018). Classifying drivers of global forest loss. *Science* (New York, N.y.), 361(6407), 1108–1111. https://doi.org/10.1126/science.aau3445
2. Drummond, M. A., & Loveland, T. R. (2010). Land-use pressure and a transition to forest-cover loss in the eastern united states. *Bioscience*, 60(4), 286–298.
3. Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* 342 (15 November): 850–53.
4. Hurtt, G. C., Pacala, S. W., Moorcroft, P. R., Caspersen, J., Shevliakova, E., Houghton, R. A., & Moore, B. (2002). Projecting the future of the u.s. carbon sink. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3), 1389–1394.
5. Johnson, M. P. (2001). Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda. *Environment & Planning* A, 33, 717–736.
6. Olofsson P, Holden CE, Bullock EL, Woodcock CE. (2016). Time series analysis of satellite data reveals continuous deforestation of New England since the 1980s. *Environ. Res. Lett.* 11(6). <https://iopscience.iop.org/article/10.1088/1748-9326/11/6/064002>
7. Thompson, J. R., Plisinski, J. S., Olofsson, P., Holden, C. E., Duveneck, M. J., & Baldwin, R. F. (2017). Forest loss in new england: a projection of recent trends. *Plos One*, 12(12). <https://doi.org/10.1371/journal.pone.0189636>
8. van der Werf, G. R., Morton, D. C., DeFries, R. S., Olivier, J. G. J., Kasibhatla, P. S., Jackson, R. B., … Randerson, J. T. (2009). Co2 emissions from forest loss. *Nature Geoscience*, 2(11), 737–738. <https://doi.org/10.1038/ngeo671>