Development of a Spatially Explicit Surface Coal Mining Predictive Model

Project Final Report

December 15, 2013

Submitted to:

Brad Kreps

The Nature Conservancy, Clinch Valley Program Director

146 E. Main St., Abingdon, VA 24210.

bkreps@tnc.org

276-676-2209

Submitted by:

Michael P. Strager*, Associate Professor of Resource Management

Jacquelyn M. Strager, Research Coordinator, Natural Resource Analysis Center

Wesley Burnett, Assistant Professor of Energy and Resource Economics

Aaron E. Maxwell, Research Analyst, Natural Resource Analysis Center

*project contact

2004 Agricultural Science Building

Division of Resource Management

Morgantown, WV 26506-6108

304-293-6463

mstrager@wvu.edu

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

The Appalachian region of the eastern United States continues to be an important source of fossil fuel for energy demands within the region and beyond. Despite the volatile nature of the coal industry sector, Appalachian coal mining remains an important factor in the regional economy (Thompson et al., 2001), as well as a significant influence on the natural environment. The Appalachian region produced 336 million short tons of coal in 2012, or just under 1/3 of total U.S. production, with production down slightly in the region compared with previous years (U.S. Energy Information Administration 2013b). Within the region, surface production of coal accounted for two thirds of total production, while underground mining contributed about one third of total production (U.S. Energy Information Administration 2013b). Regional coal resources include primarily steam coal used in electric power generation, and (to a lesser extent) metallurgical coal used in industrial processes. Coal production is shifting within the region, as demand for cleaner-burning, lower sulfur coal has risen due to increased environmental regulation.

The overall future of Appalachian coal resource extraction is increasingly uncertain. There is a complex, dynamic relationship between the price of coal, the price of competing resources (in particular natural gas), and potential greenhouse gas emission reduction policies which may reduce the demand for coal. Coal is subject to increased competition from natural gas as a source of energy for electricity generation, and may be equaled or surpassed by natural gas in the near future depending on oil and gas prices, greenhouse gas related policies, coal production costs, and other factors (U.S. Energy Information Administration 2013a).

Mountaintop removal coal mining has been identified as the main source of land use change across the central portion of the Appalachian region (Saylor 2008). The environmental impacts of mountaintop surface mining include impacts on biodiversity, hydrology, human health, and water quality (Palmer et al. 2012). Recent work has also quantified the current spatial environmental impact of mountaintop removal mining by relating areal extent of surface coal mining activities to coal production (Lutz et al. 2013). This study combines varying estimates of surface coal mine production (EIA 2012) with spatially explicit predictive modeling to map potential future surface mining footprints on the landscape through the year 2035.

The goal of this project was to create a spatially explicit 1km^2 grid cell model for the Appalachian Landscape Conservation Cooperative (Figure 1) predicting where surface coal mining is likely to occur in in a projected future time period, under two different scenarios. To accomplish this goal we combined GIS spatial analysis, a Random Forests predictive model, and future mining buildout scenarios.

This report provides a detailed methodology of our approach and discussion of our results. The report has three main sections.

Section 1. Predictor variable selection. A fundamental first step to our project was to select those landscape variables which can be used to effectively predict the locations of future surface mining. Critical to our spatial model were the: general geographic extent of coal, physical properties of the coal resource, and infrastructure related predictor variables (mainly related to transportation/delivery of coal). Additional spatial inputs used in the model included active surface mine permits and exclusion areas. The exclusions are areas where surface mining will not occur due to incompatible land uses such

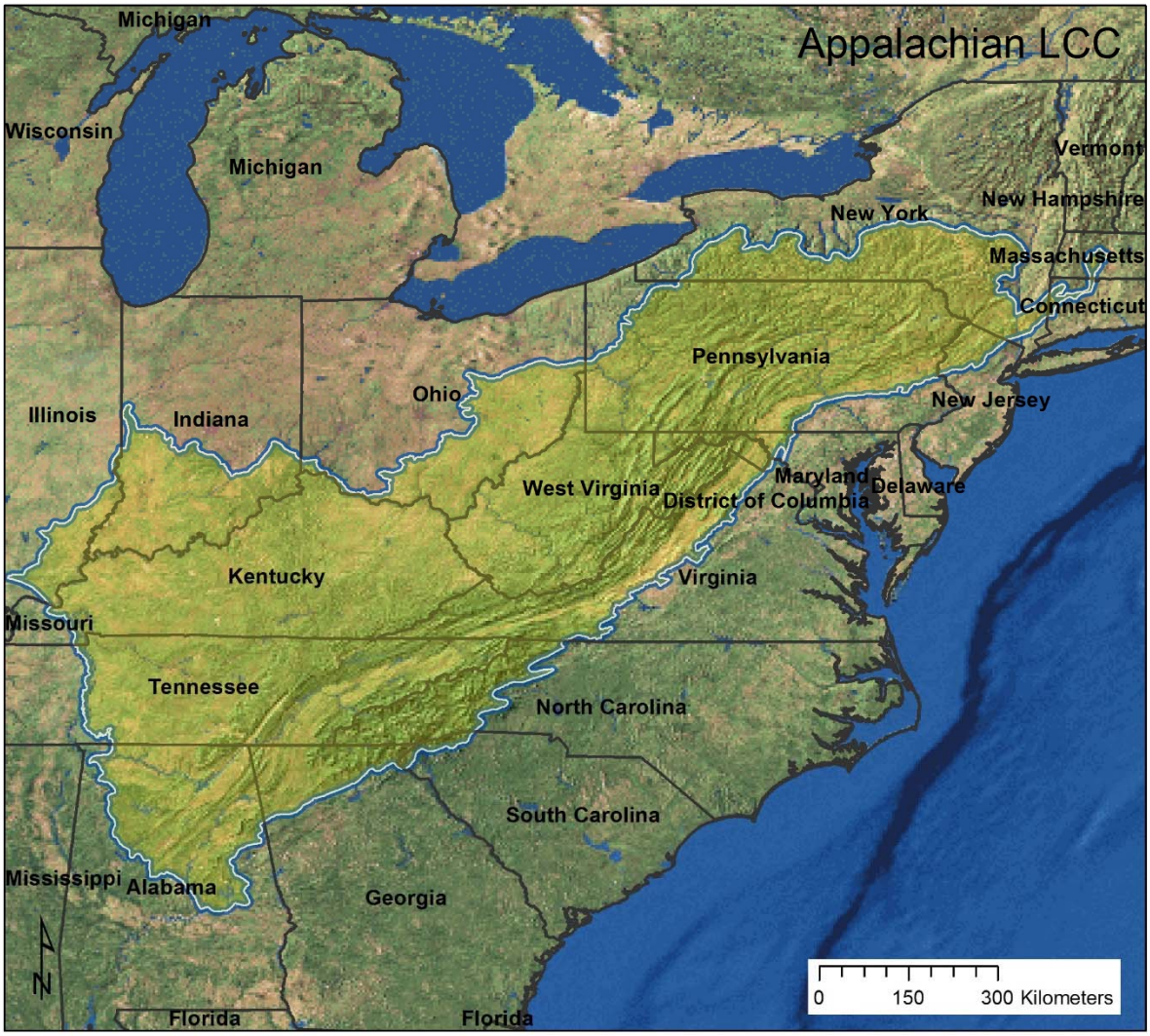
as developed urban or residential land, water, permanent public and private conservation areas, and areas which have been extensively surface mined and reclaimed in the last 10 years.

Section 2. Random Forests predictive model and results. We provide a brief overview of why Random Forests was chosen for this study as well as methods for integrating predictor variables into Random Forests and model parameters.

Section 3. Predictive mapping: Future surface mining footprint. Here we describe how we sequentially allocated regional predictions of coal production to the highest probability Random Forests 1km^2 cells. Our final output includes mapped layers showing the locations of probable surface mining activity under low and high coal production scenarios for the year 2035.

All digital and spatial data layers used in this study are available for distribution as well as the Random Forests code parameters for R.

Figure 1. Appalachian Landscape Conservation Cooperative (LCC) boundary.



2. METHODS

2.1 PREDICTOR VARIABLE SELECTION

Variables used to predict the likelihood of surface mining in this study included general properties (Energy Information Administration coal supply region, Environmental Protection Agency mountaintop removal region), physical properties of the coal resource (coal geology type, sulfur content, ash content, and BTUs), and infrastructure related predictors (network distance to existing coal fired power plants, network distance to intermodal transportation facilities, network distance to inland ports, distance to rail, human population density). All grids were calculated with a cell size of 1km^2 using ESRI ArcGIS 10.1 software, with analysis extent limited to the Appalachian LCC. For distance grids (distance to power plants, distance to railroads etc.) distances were calculated to features outside the Appalachian LCC prior to limiting grids to the study area boundary.

Predictor variables:

Energy Information Administration coal supply region

The U.S. Department of Energy Administration (EIA) has published maps for generalized coal supply regions of the United States. The Appalachian LCC includes four distinct coal supply regions: Northern Appalachian, Central Appalachian, Southern Appalachian, and Eastern Interior/Illinois (portion) (Figure 2).

Environmental Protection Agency mountaintop removal region

The U.S. Environmental Protection Agency has designated portions of West Virginia, Kentucky, Virginia, and Tennessee as the mountaintop removal/valley fill mining region (Figure 3) (U.S. Environmental Protection Agency 2005). Project reviewers suggested that the mountaintop removal region be used as a categorical predictor variable as a surrogate for areas appropriate for future surface mining due to favorable overburden and seam thickness within this region.

Coal geology type

Generalized coal field boundaries were derived from a map of coal fields of the United States at a 1:5,000,000 scale (U.S. Geological Survey 2008). Generalized coal fields include areas with known coal-bearing geology, and were used to limit the extent of predicted future mining probability within the study area (future mining was limited to areas within mapped coal fields).

Within this coal field boundary, we also obtained state level geologic maps from datasets compiled by USGS for U.S. states (U.S. Geological Survey 2013). Next, the generalized state level geologic maps were classified into geologic units containing coal, and those without coal. Finally, the geologic units containing coal were further cross-referenced into 17 different geological units region wide based on generalized lithology and formation provided in general state level geological labeling among different states reference materials. The cross referencing process was necessary due to inconsistencies and labeling among the different states. This was completed using a chronostratigraphic correlation chart (Ruppert et al. 2010) (Figure 4). Formations were grouped based on geologic age so as to produce the 17 categories of similar lithology that are not impacted by state boundaries. The results of this correlation were assessed by visualizing the data at state borders (Figure 5).

Sulfur percentage of coal

The sulfur content of coal is one aspect of coal quality. Coal contains varying amounts of sulfur, and when coal is burned, the sulfur (combined with oxygen) will form sulfur dioxide, a greenhouse gas. Restrictions on sulfur dioxide emissions from power plants have made the relative sulfur content of coal an important consideration in the economic viability of different coal resources (with low sulfur coal generally being more desirable). The percentage of sulfur content in the coal (Figure 6) was interpolated using borehole data (x, y, z) from the USGS Coal Quality database (Bragg et al. 1997). Prior to interpolation, borehole data were limited to samples taken at the surface (underground or deep mine samples were excluded). Underground and borehole samples (excluded) were identified by sample depth values and/or descriptive text in the comments field in the sample database. Surface samples were also identified by values in the comments field indicating samples were taken at road cuts, pits, and strip mines. "The data source of borehole locations contains a comprehensive analyses of more than 13,000 samples of coal and associated rocks from every major coal-bearing basin and coal bed in the U.S. The data in the coal quality database represent analyses of the coal as it exists in the ground. The data commonly are presented on an as-received whole-coal basis." (Bragg et al. 1997). While different coal seams may be encountered with each of the borehole sites, an overall sulfur percentage is assumed for each site. Boreholes were targeting different coal beds, and the assumption was made that the coal seam being sampled at that location would provide a representative estimate of sulfur content for that area. Or, the coal seam being sampled at the borehole would be the coal seam being mined in that area.

The interpolation process for sulfur, ash, and BTU followed standard geostatistical kriging steps outlined by Johnston (2001). They included first exploring the data for normality, examining trends and the semiovariogram, and testing model output runs until a satisfactory root mean squared error and mean standardized error from the cross validation prediction errors were found.

For sulfur, an ordinary kriging model was applied and anisotropy examined to account for directional influences. This was useful especially since the coal geology follows unique ridge and topographical synclines. A total of ten lags were applied with a size of 20,000 to best fit the distribution of the input point locations. The search neighborhood was standard sized with a maximum of 5 neighbors. The results for sulfur cross validation indicated an accurate predicted surface with a root-mean-square standardized prediction error of 1.009 (a value closer to 1.0 is preferred).

Ash content of coal (ash yield)

Ash content of coal is also related to relative coal quality. Ash content is related to the portion of coal that remains after combustion. Ash yield was also obtained from the USGS Coal Quality database (Bragg et al. 1997) and was also interpolated using methods similar to those used for sulfur content. Ash content is shown in Figure 7.

For ash, again ordinary kriging was applied with anisotropy examined for the directional influences which indicated an improved fit with an angle of 44.6 and 45 tolerance. The lags used were different for ash – 12 total with a lag size of 12,000. The search neighborhood was standard sized with a maximum of 5 neighbors as with sulfur. The results for sulfur cross validation indicated an accurate predicted surface with a root-mean-square standardized prediction error of 1.001.

BTU content

British thermal unit (Btu) content of coal is related to the amount of energy provided by a given amount of coal. Btu content of coal per lb. was derived from the USGS Coal Quality database (Bragg et al. 1997) using methods similar to ash and sulfur content. Btu content is shown in Figure 8.

For the BTU interpolation, a simple kriging model was applied with a log score transformation to make the variances more constant throughout the study area and bring the data closer to being normally distributed. Anisotropy was applied to account for direction in the semivariogram and covariance. The preferred angle was 32 with a 21.4 degree tolerance. Twelve lags with a size of 15,000 was found to fit the model best with the averaged data points. Again here, the standard neighborhood search was used with a maximum of 5 neighbors. The fit for BTU was not as well as ash and sulfur with a root-mean-square standardized error of 0.887.

Distance to coal fired power plants

Existing coal fired power plants were identified using information published by the U.S. Energy Information Administration, based on form EIA-860 “Annual Electric Generator Report” (U.S. Energy Information Administration 2011a). The locations were determined using latitude/longitude coordinates provided by SourceWatch (2013) and shapefiles provided by Energy Information Administration (U.S. Energy Information Administration 2012a). We identified a total of 318 existing power plants as of 2011. We then removed a total of 92 of these plants that are scheduled for closure between 2013 and 2020, according to news reports and accounts compiled by Source Watch (2013). An additional 25 new coal fired facilities (including power plants, cogeneration facilities, coal to liquids plants) were added to the final dataset that are proposed, planned, in permitting, or under construction for this area as noted by Source Watch, (2013), the Sierra Club (2013), and National Energy Technology Laboratory (2012). For our final predictor variable, we calculated distance along a highway network (ESRI 2012a) to 251 coal fired power plant facilities (226 existing, 25 new). Distance along the highway network was initially calculated along 1km^2 cells along the actual highways, and was then extrapolated out to cover all cells within the Appalachian LCC using an inverse distance weighted interpolator (Figure 9).

Distance to intermodal transportation facilities

Intermodal transportation facilities are locations where freight may be transferred between different modes of transportation (i.e. truck to barge, truck to rail, etc.). Intermodal facility point locations were obtained from the National Transportation Atlas Database, and were then limited to all facilities except ports, which were mapped separately (Bureau of Transportation Statistics 2011) (Figure 10). Distance to intermodal facilities was mapped along the highway network, then extrapolated out to all cells within the Appalachian LCC.

Distance to inland ports

Inland ports were also obtained from the National Transportation Atlas Database (Bureau of Transportation Statistics 2011) and were limited to those ports handling coal and coal related commodities (Figure 11). Distance to ports was mapped along the highway network, then extrapolated out to all cells within the Appalachian LCC.

Distance to rail

According to U.S. Energy Information Administration domestic coal distribution statistics, 56% of coal produced by the ten coal-producing states in the study area was distributed using rail in 2011 (U.S. Energy Information Administration 2012a). In addition, a total of 29% of coal distributed domestically was moved by river (barges), with a total of 13% was transported by truck. This implies that proximity to rail, river, and trucking related loading facilities may be an asset in location of mining activity. We found that 20% (10/49) of randomly selected surface mine permits in a 5 county area in WV have existing loading facilities which enable coal to be placed on rail cars for distribution. Figures 12a and 12b show examples of facilities which are adjacent to the rail lines for loading coal. Mining related facilities (for loading coal onto rail cars) are not necessarily limited to locations at end points of rail lines. Mine loading facilities can also be found at any point along rail lines, not just at the end points or at spurs. Mapping distance to existing rails captures more potential locations for access to rail lines from coal mining permit locations, rather than limiting the rail feature dataset to endpoints only of existing railroads.

Locations of railroads were acquired from the Bureau of Transportation Statistics U.S. National Transportation Atlas railroads layer, at the 1:100,000 map scale (ESRI 2012). Distance to nearest rail line was mapped as Euclidean straight line distance across the Appalachian LCC (not limited to distance along network) (Figure 13).

Population density

Population density was calculated across the study area using 2010 Census block group data, and was then converted to raster format, 1km^2 cell size (ESRI 2012) (Figure 14). Generalized land cover from the National Land Cover Dataset for 2006 was also considered as an input, but was omitted from the final model.

Figure 2. U.S. Energy Information Administration (EIA) coal supply regions used in this project.

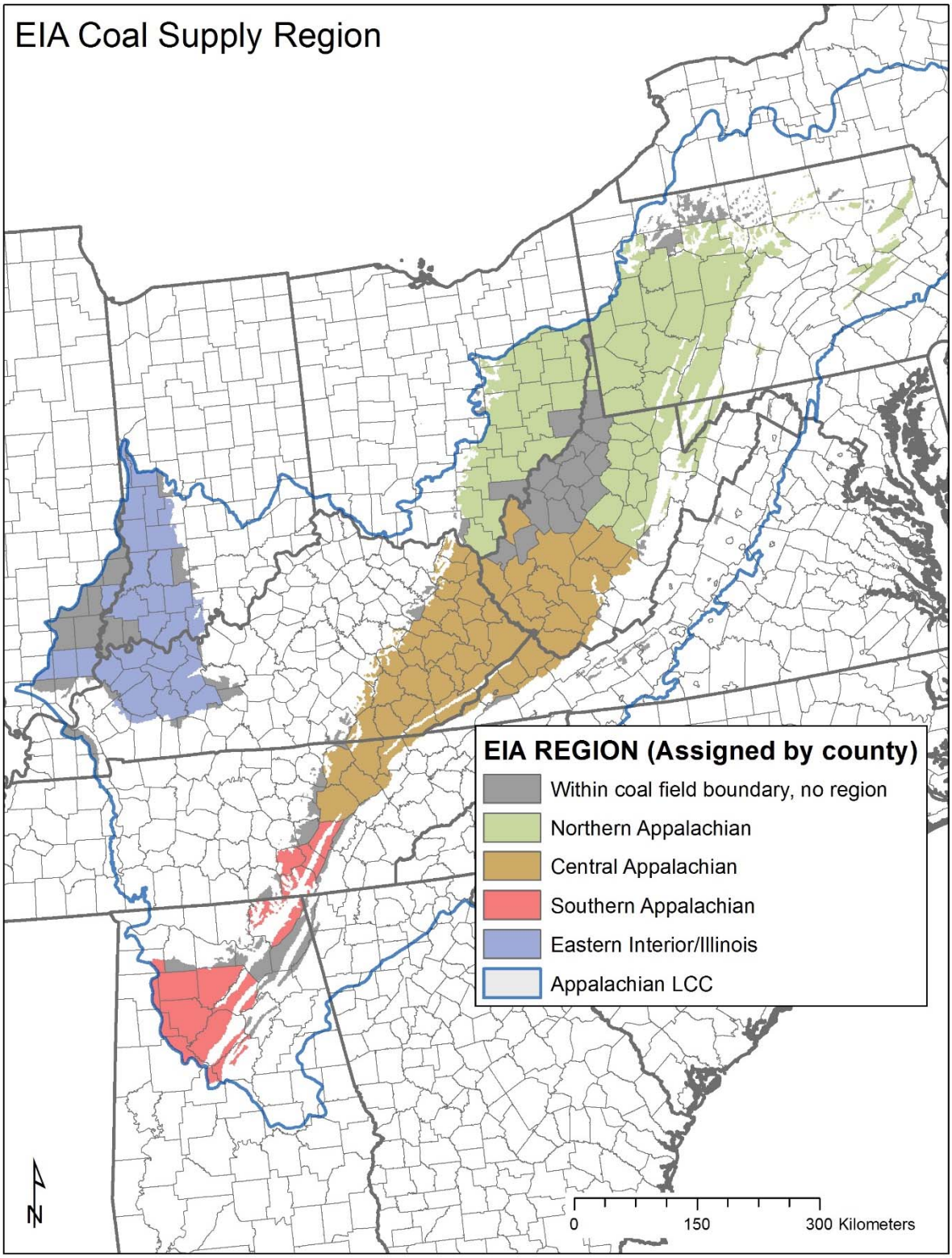


Figure 3. U.S. Environmental Protection Agency mountaintop removal/valley fill region (U.S. EPA 2005).

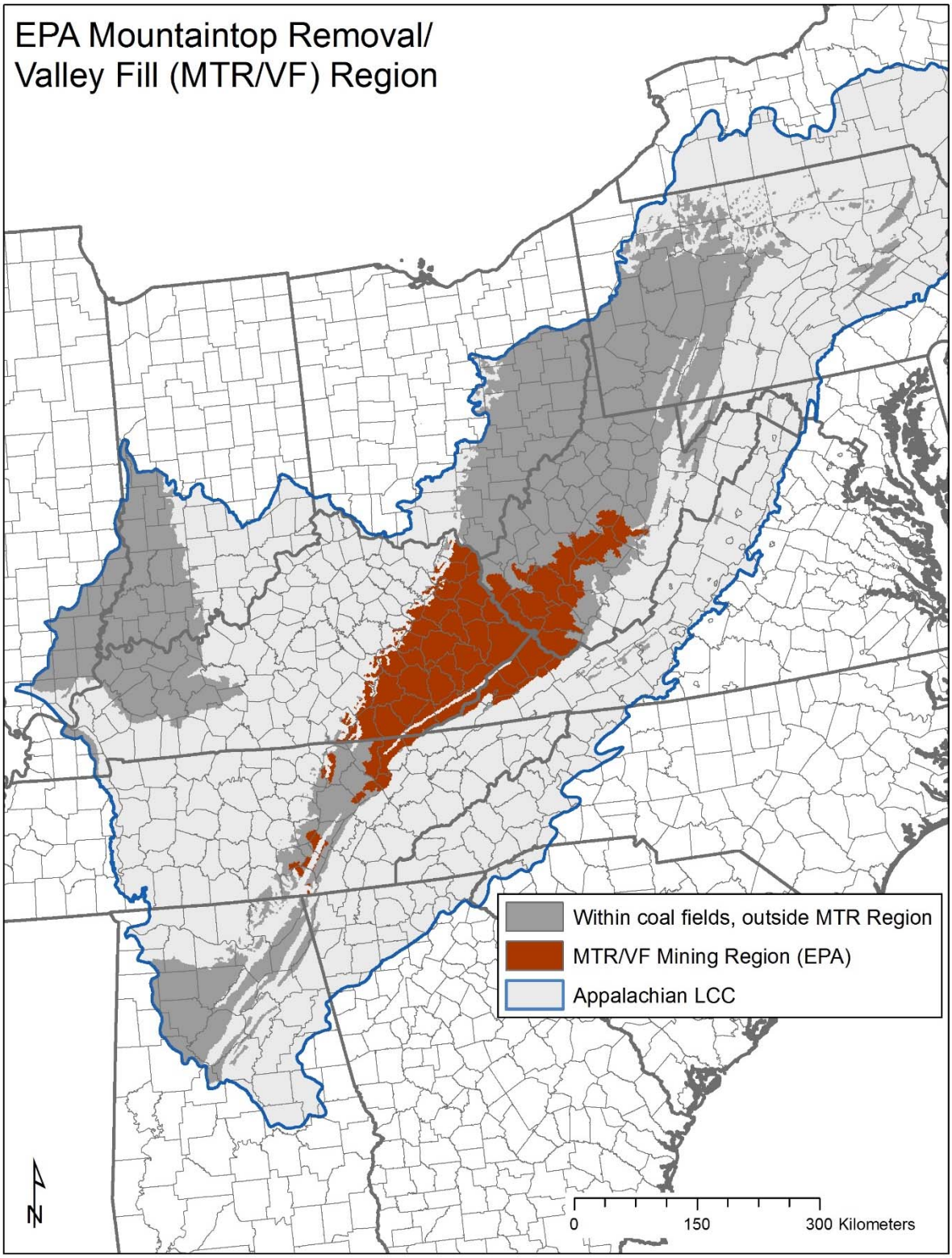


Figure 4. Chronostratigraphic correlation chart for cross referencing state level geological maps. (Ruppert et al. 2010).

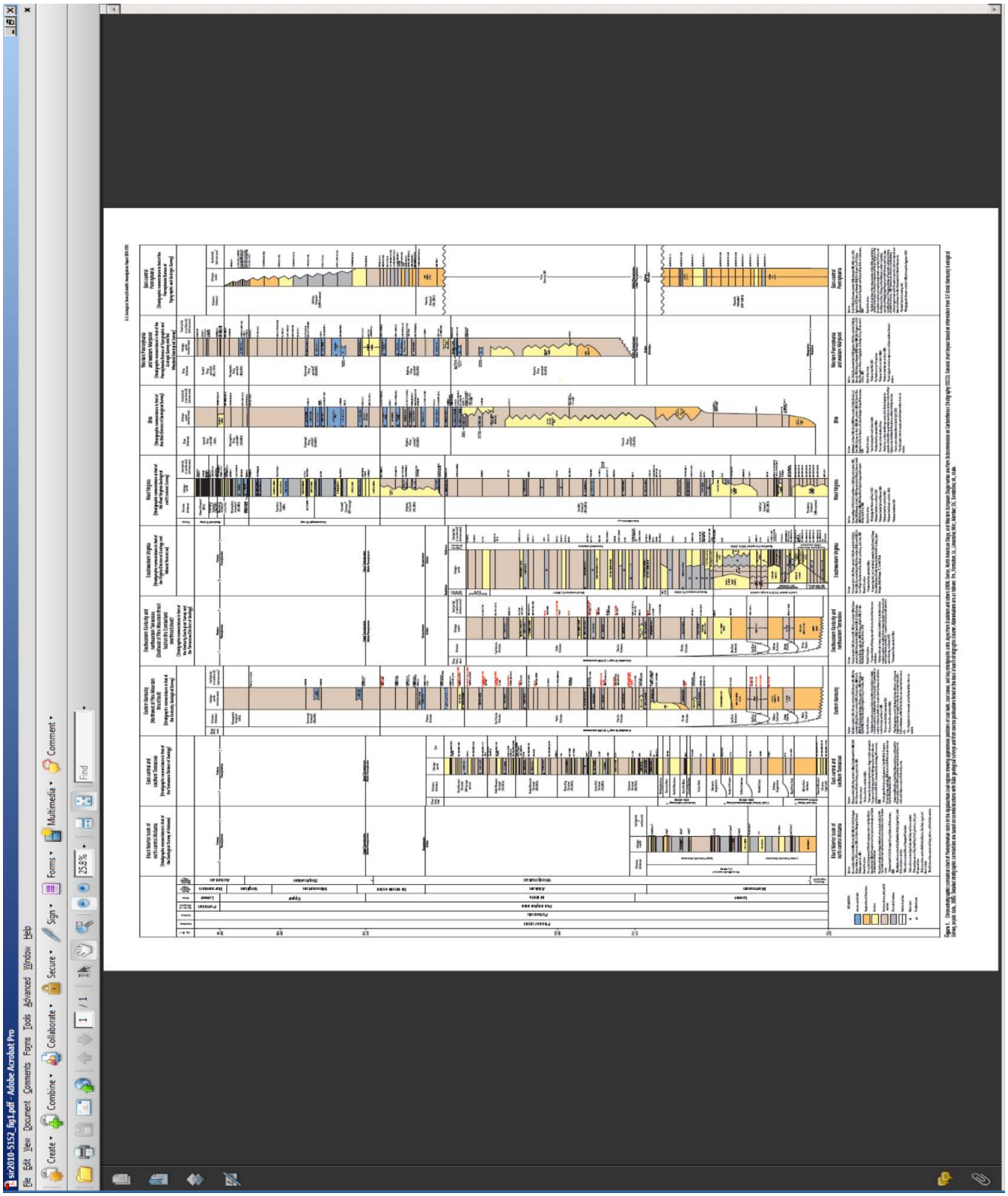


Figure 5. Coal-bearing geological units, as grouped for purposes of this study.

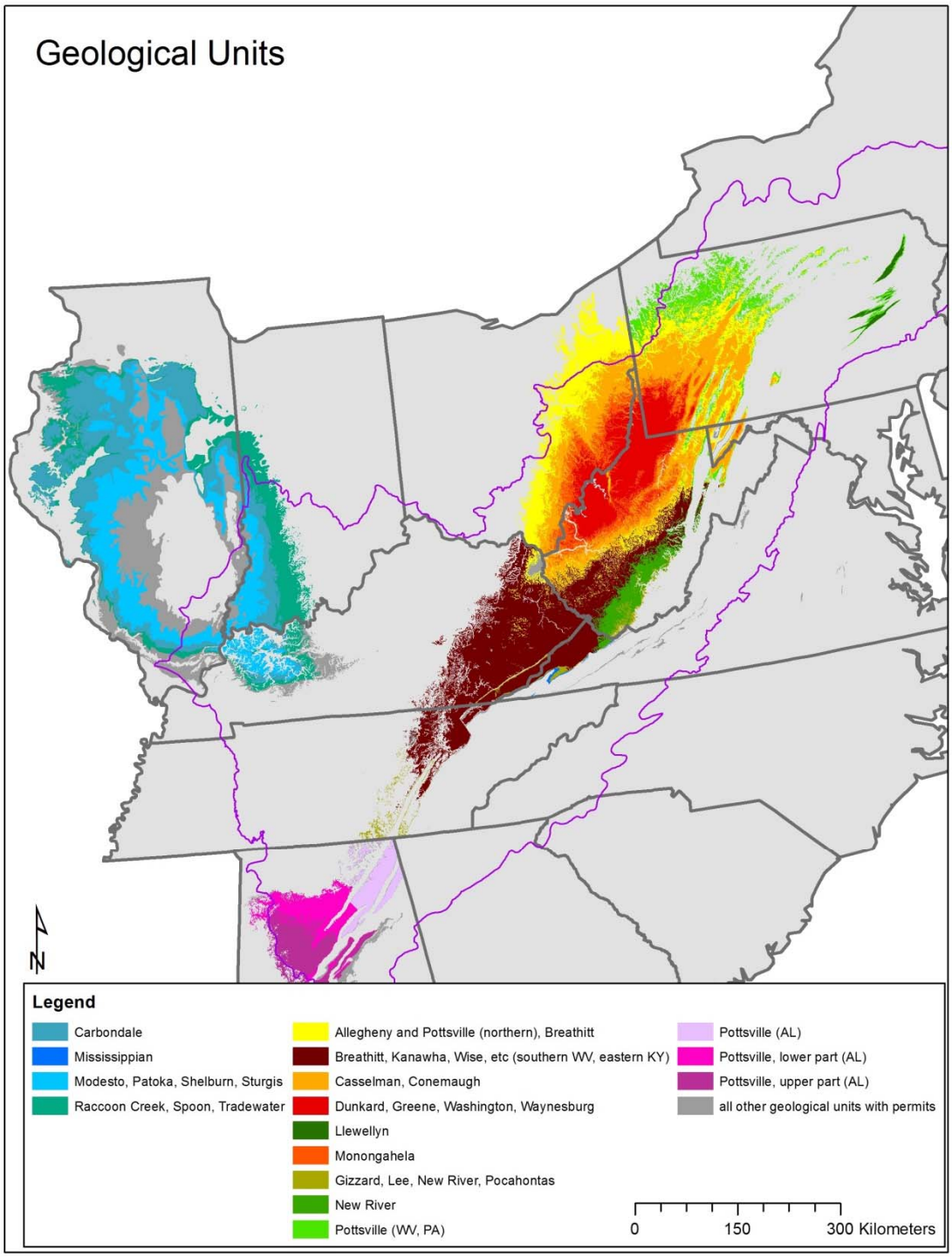


Figure 6. Data interpolated from USGS COALQUAL database points: Sulfur content of coal.

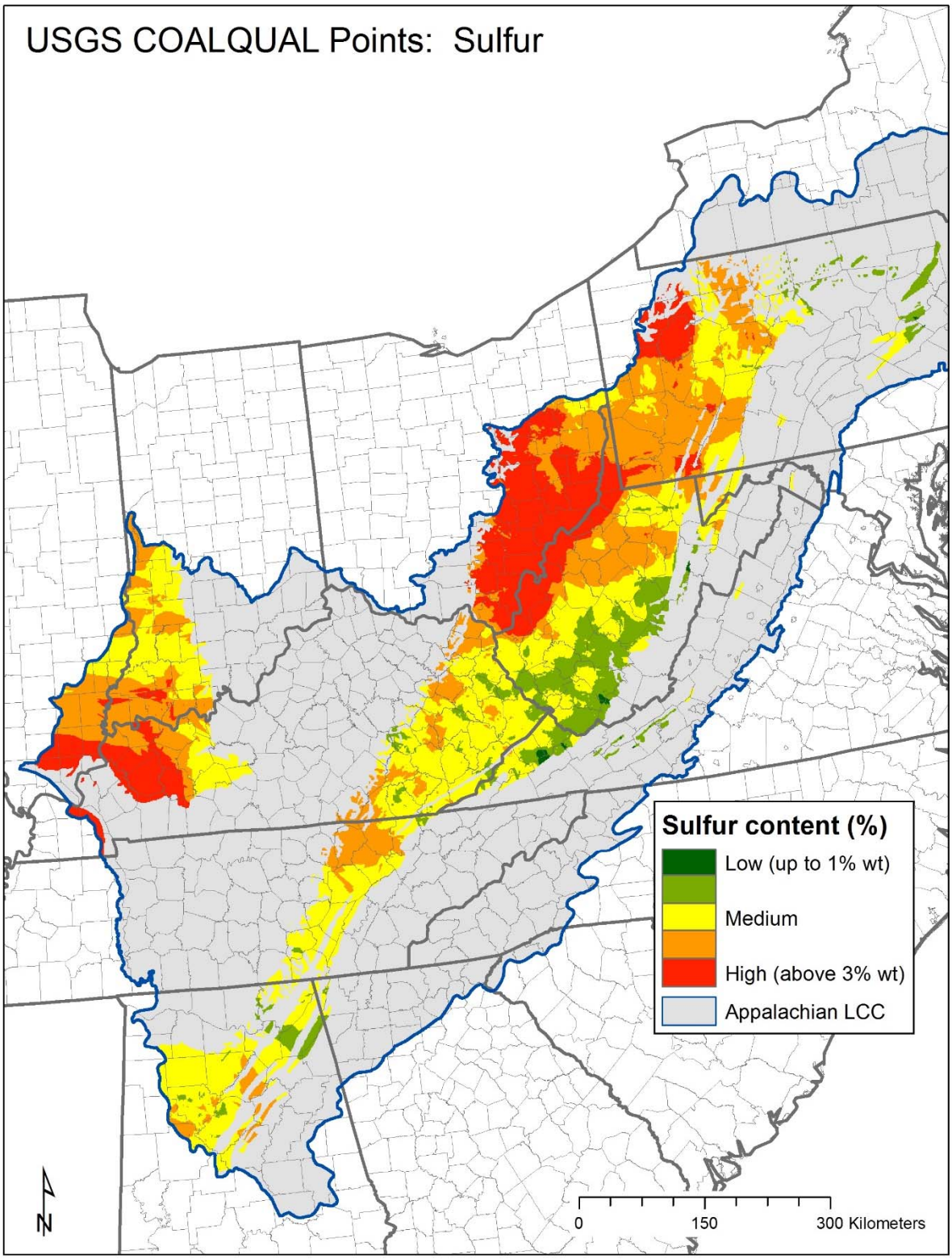


Figure 7. Data interpolated from USGS COALQUAL database points: Ash yield of coal.

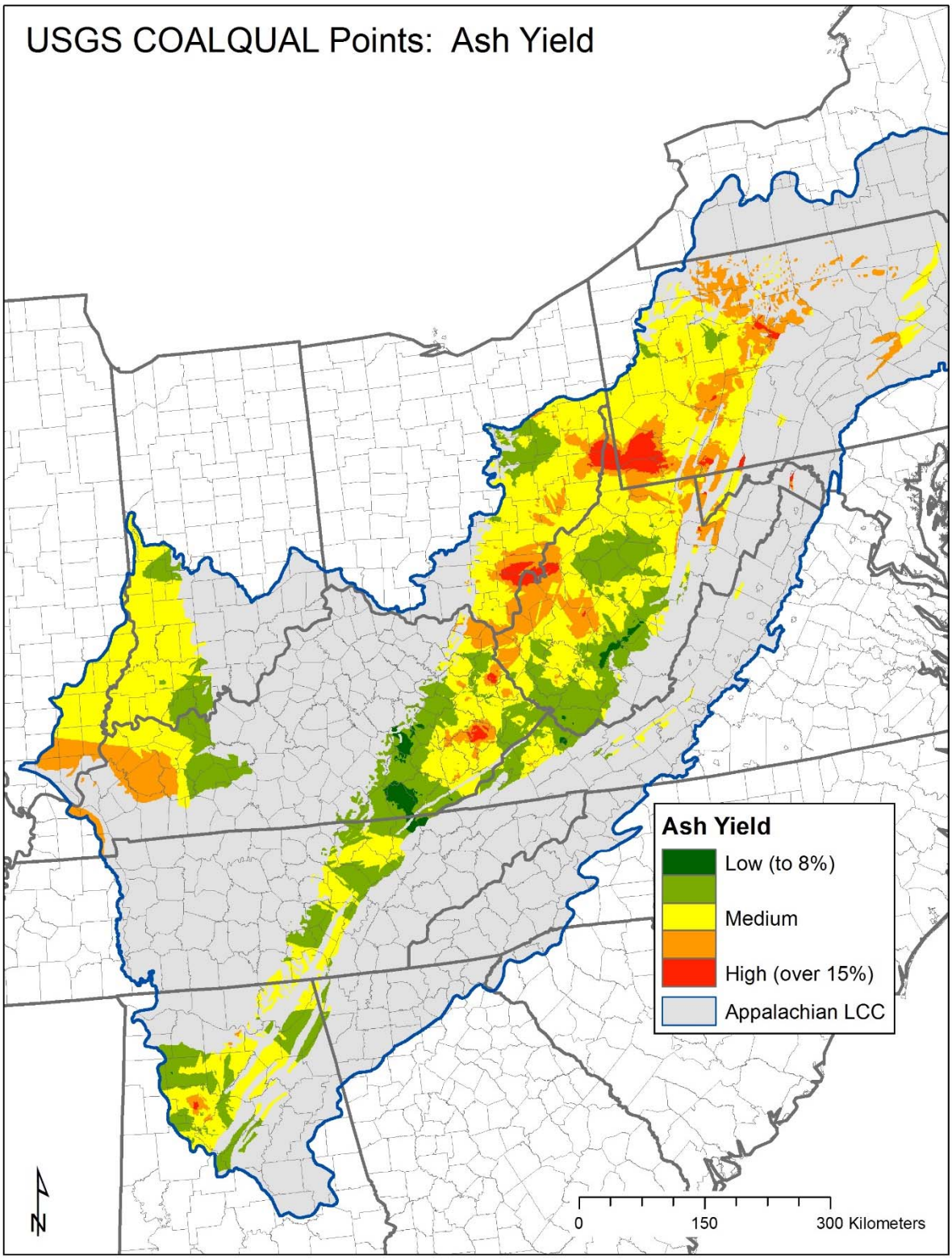


Figure 8. Data interpolated from USGS COALQUAL database points: BTU content of coal.

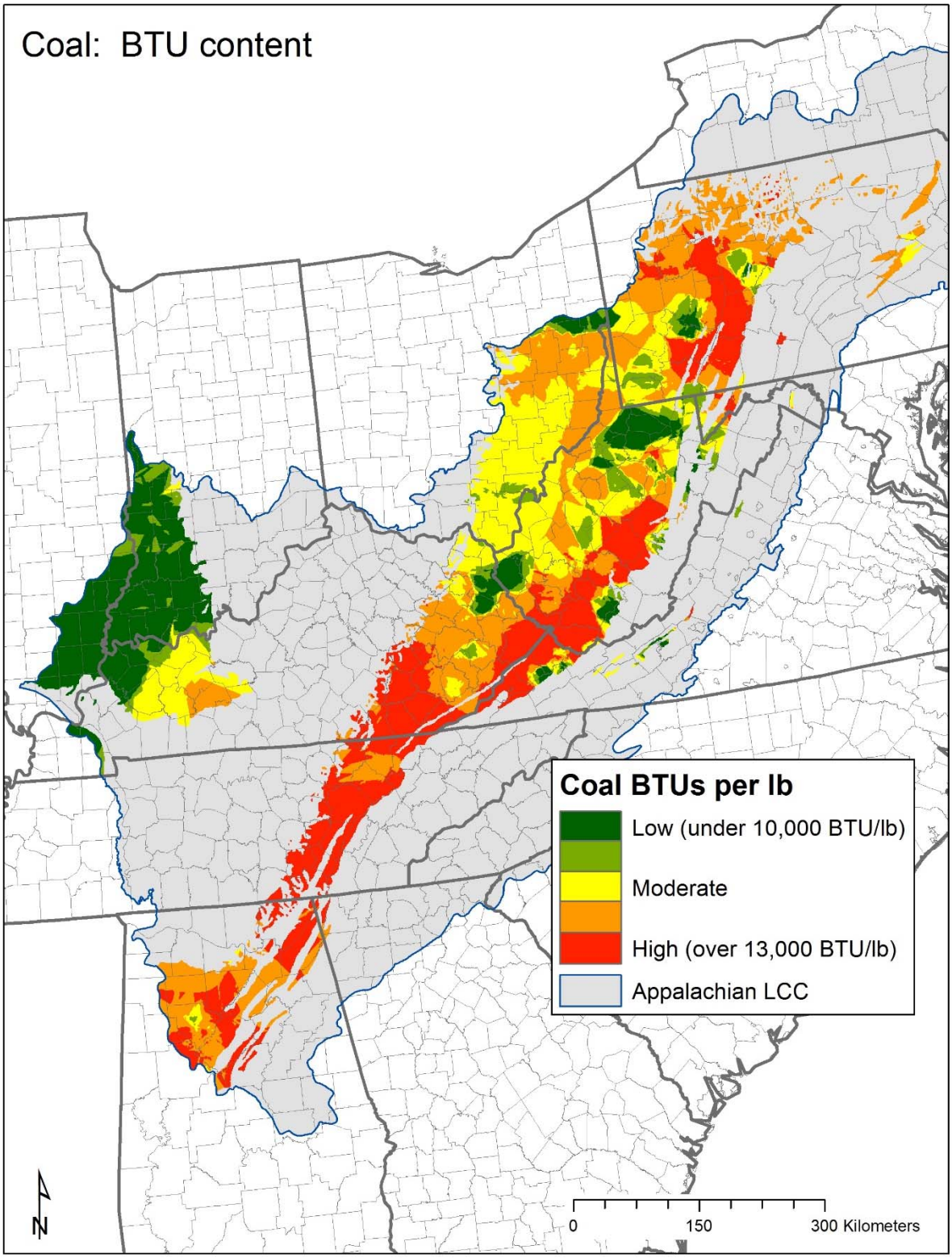


Figure 9. Power plant locations (coal fired), and distance to power plants along road network.

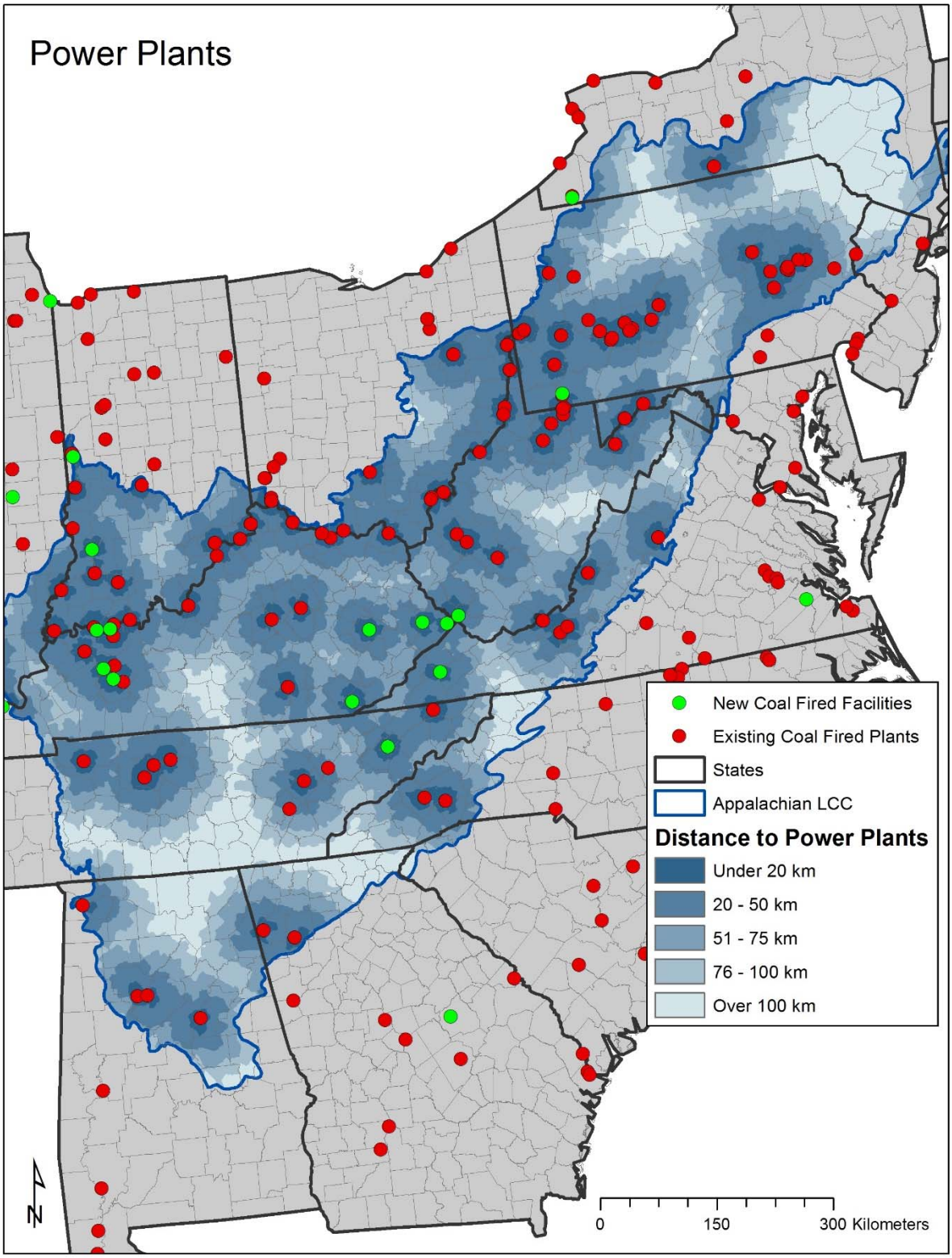


Figure 10. Intermodal transportation facilities, and distance to facilities along road network.

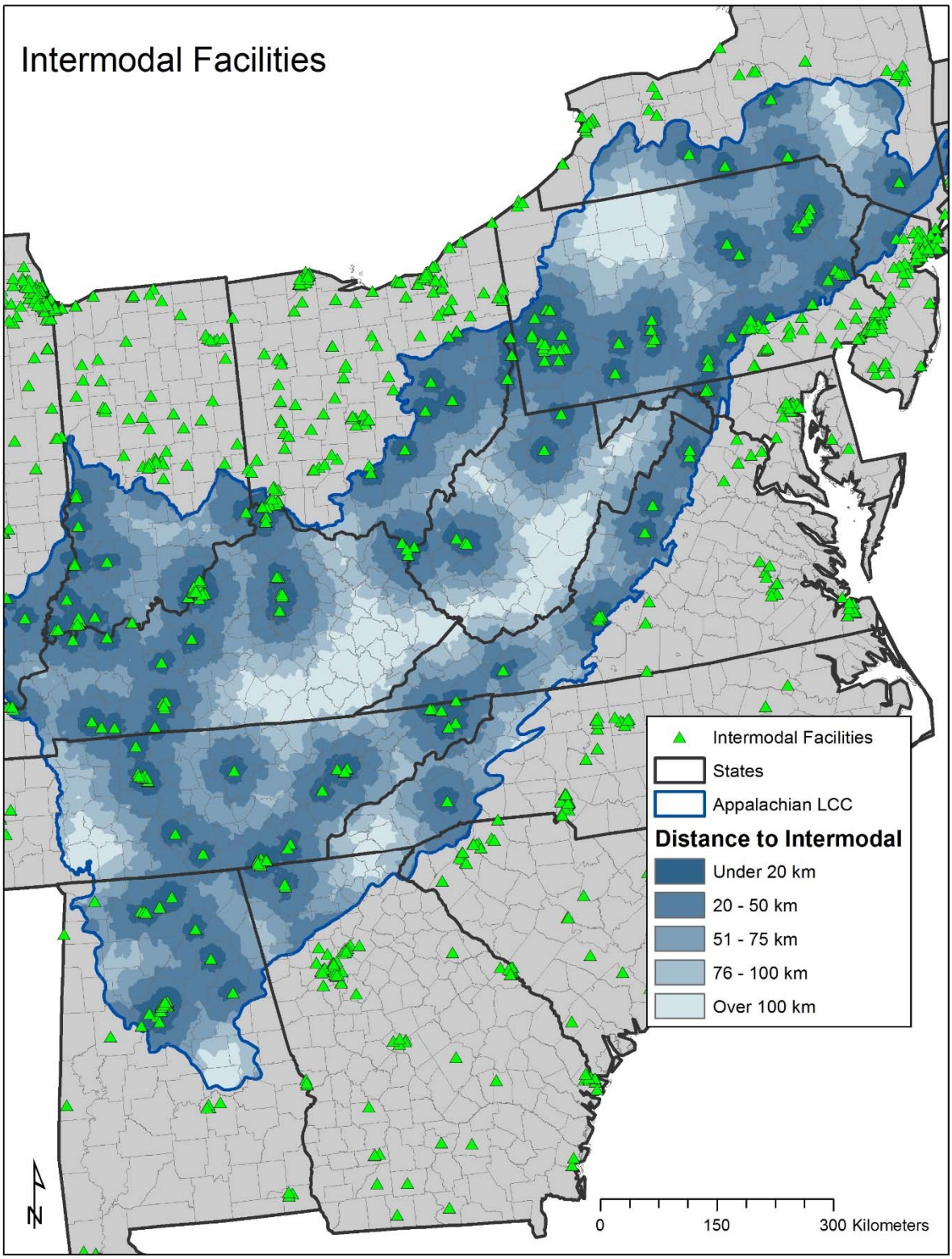


Figure 11. Inland ports, and distance to ports along road network.

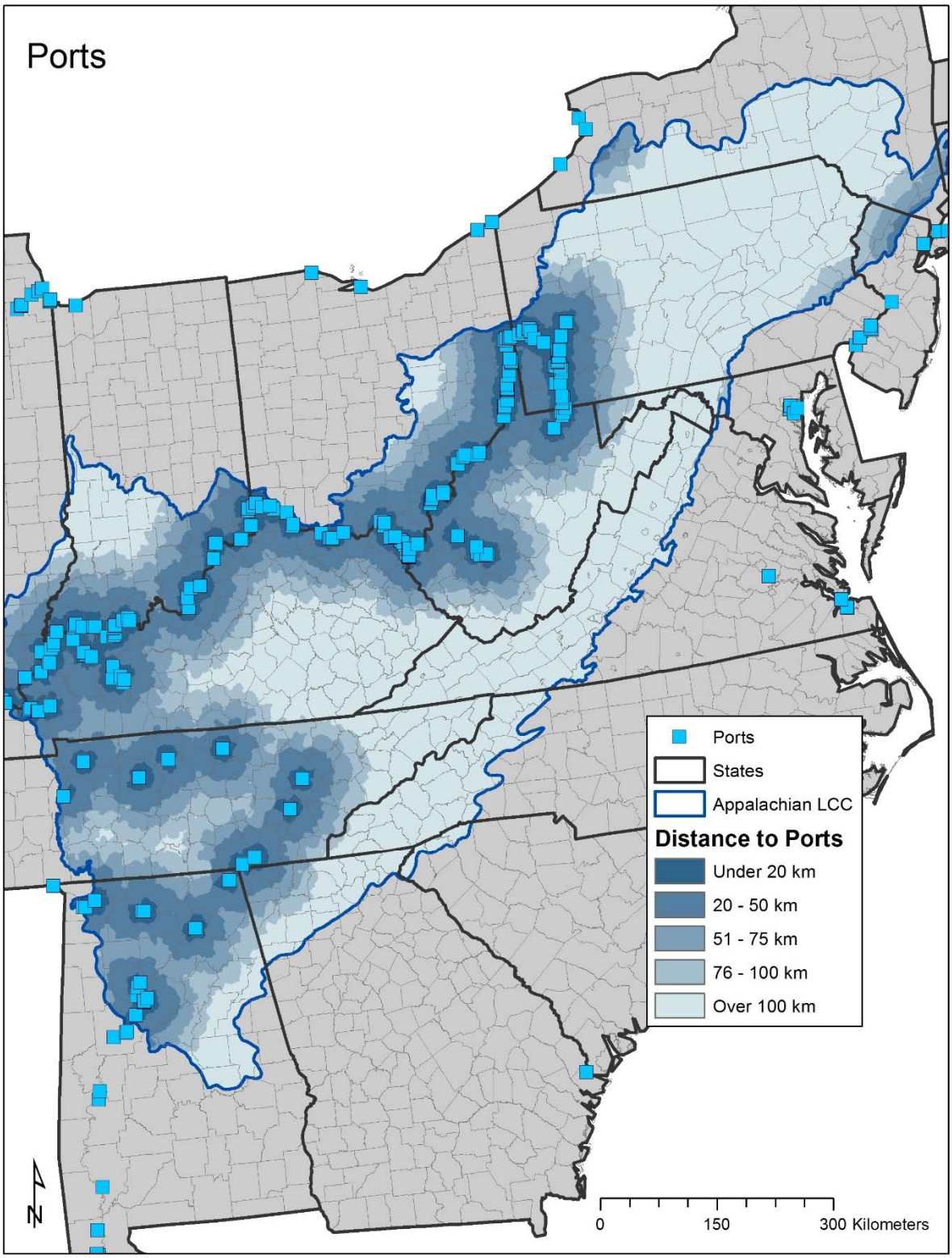
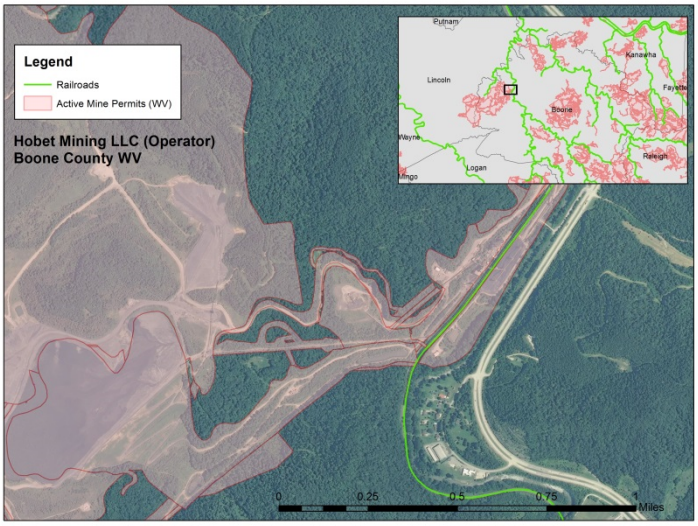


Figure 12 a. Bandmill no 1. Mine, Boone County, WV.

Figure 12b. Hobet mine, on site rail loading facility, Boone County, WV.



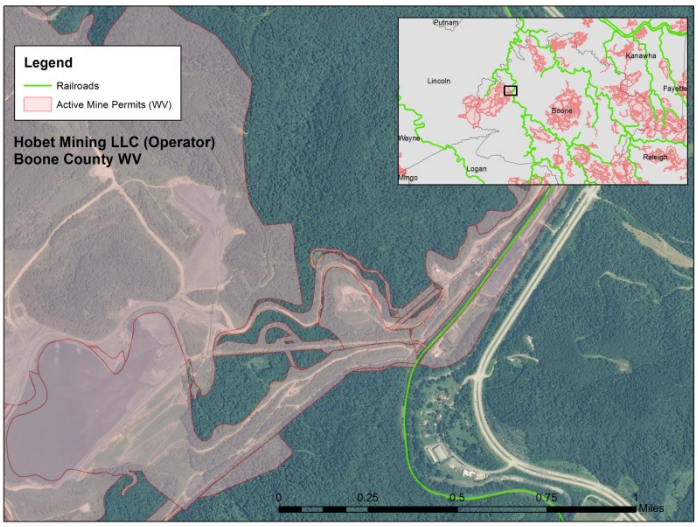


Figure 13. Railroads, and distance to railroad (Euclidean distance).

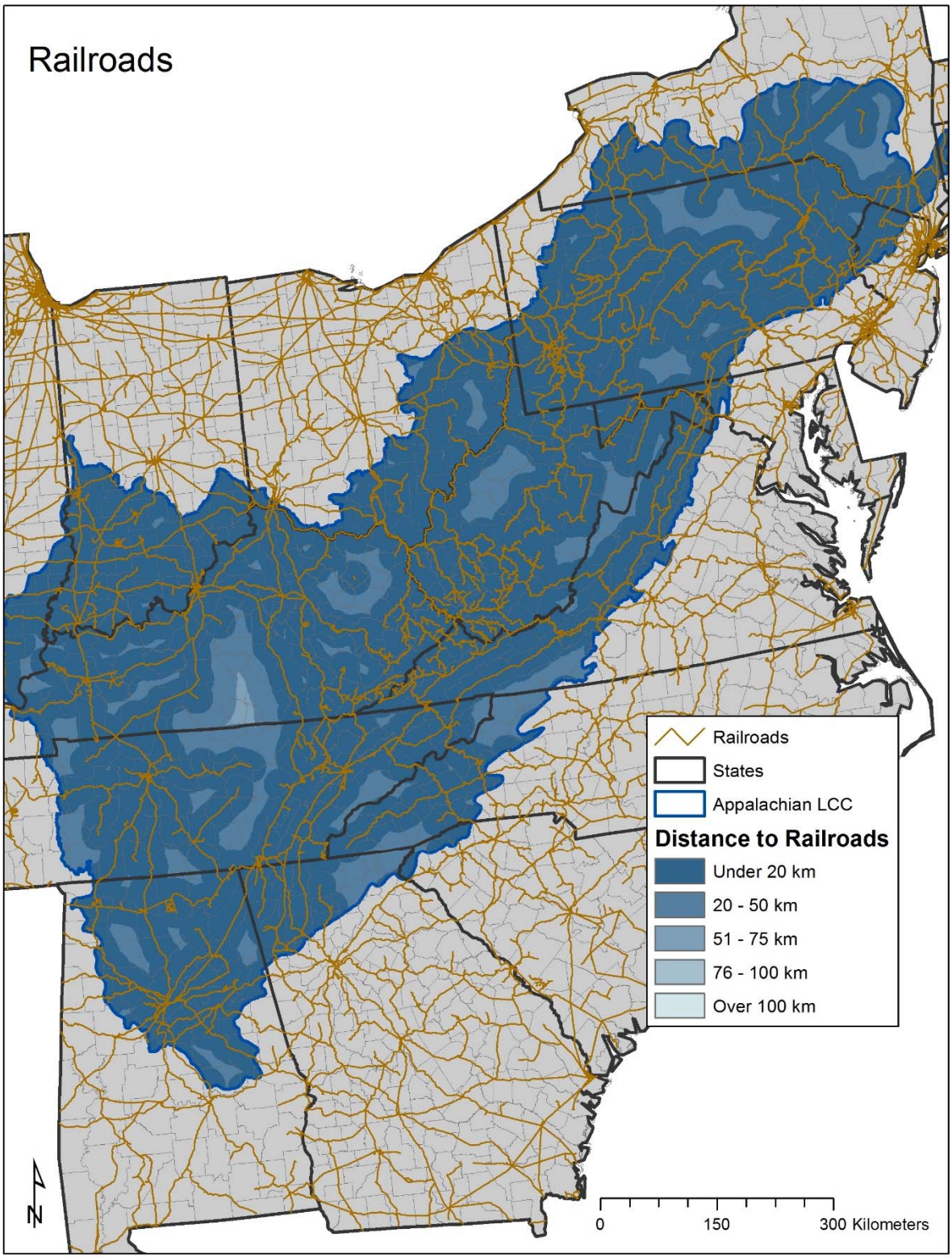


Figure 14. Population density (persons per square mile, 2010).

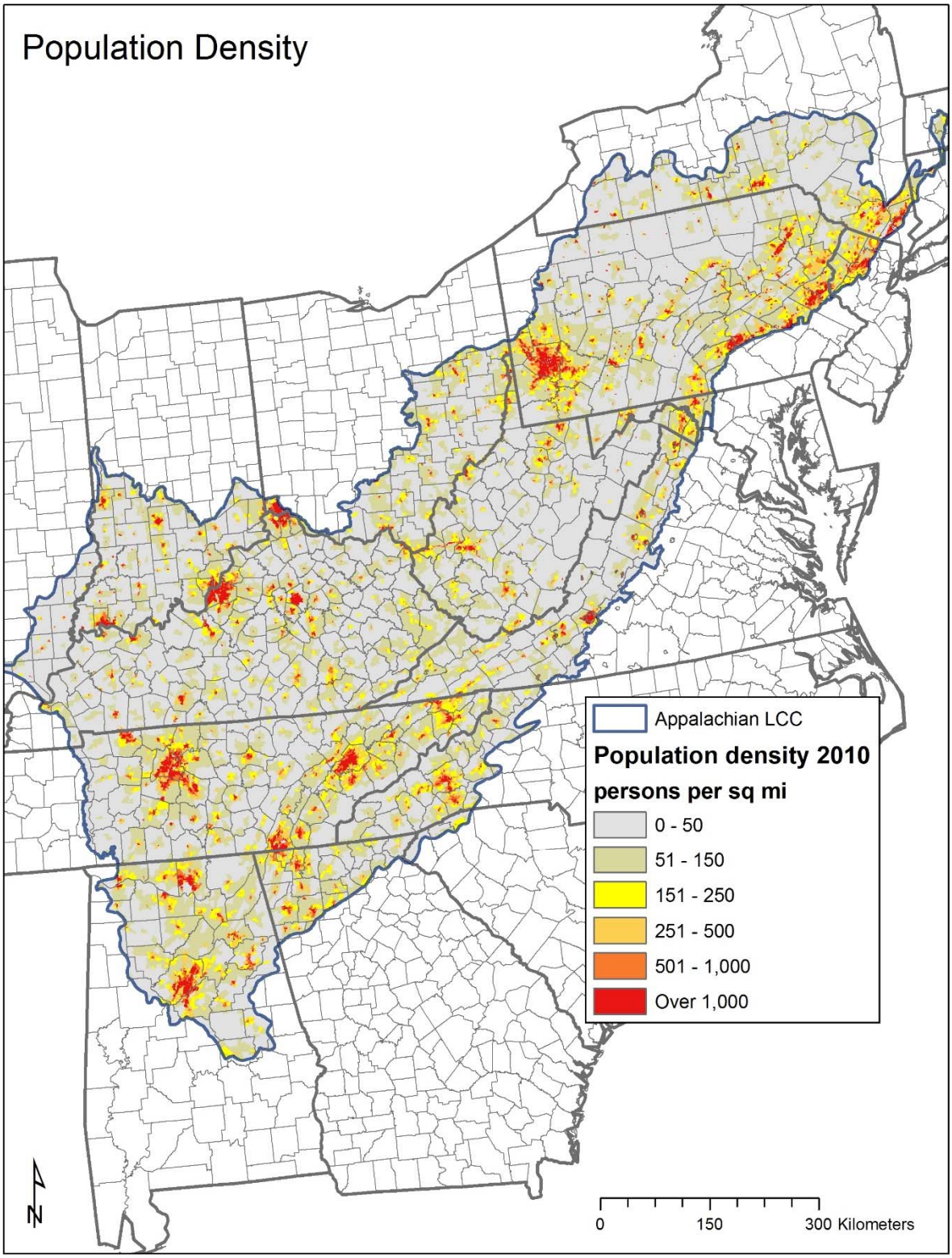
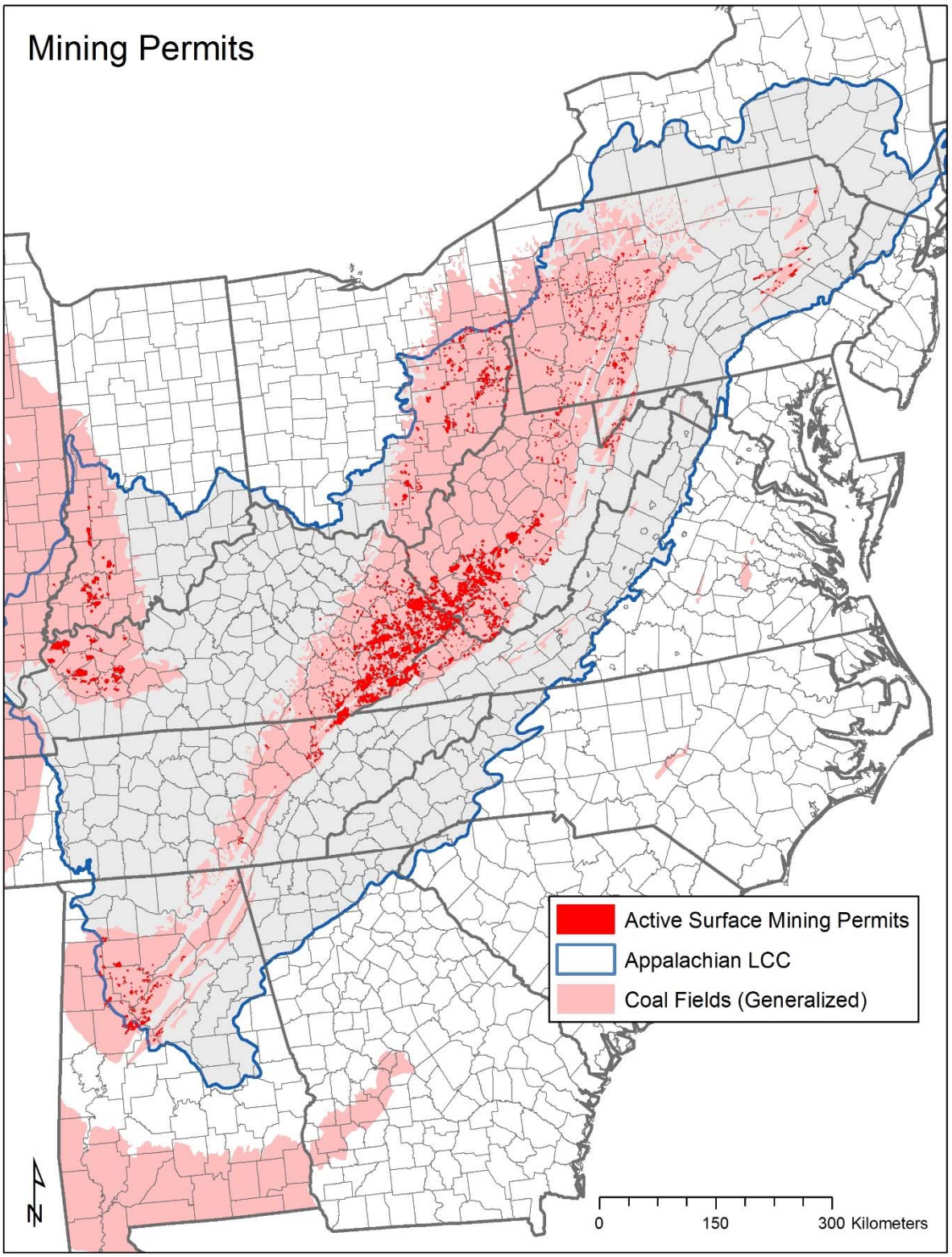


Figure 15. Active surface mining permits from state agency datasets.



Additional spatial data development

Other considerations

Other factors considered to be important for new surface mining activity included past and existing mining, stripping ratios (overburden, coal bed thickness), coal reserves remaining, surface ownership patterns, and coal quality as related to market demand. Each of these factors were specifically mentioned by reviewers in various stages of this project, and were also mentioned in the Environmental Impact Statement for mountaintop mining in the Appalachian region (U.S. Environmental Protection Agency 2005). Ultimately, these factors were not included (directly) in the final modeling process, after investigation of available datasets and data quality. Location and extent of past mining were not uniformly available for the entire study area, as mining datasets from individual states varied greatly in quality. Data related to stripping ratios (overburden, seam thickness) were available for some coal seams (U.S. Geological Survey 2000) and states in the study region (Illinois (Illinois State Geological Survey 2012); Indiana (Indiana Geological Survey 2000); West Virginia (West Virginia Geological and Economic Survey 2013); Virginia (Virginia Tech 1999) but not others. Remaining coal reserves are available on a county-by-county basis for some states (see West Virginia Coal Association 2012 for example) or on a regional level from the U.S. Energy Information Administration, but reserve data are not consistently published at a detailed enough spatial scale for the region in order to be included in the project. The limitation of the model to surface mining activity only (rather than surface and underground combined) also caused some difficulty in obtaining suitable datasets.

For surface land ownership patterns, it has been suggested that the differing nature of land ownership among states may be related to surface mining – specifically that surface mines of eastern Kentucky are characterized by smaller land owners, while surface mines in neighboring southwestern West Virginia are more likely to be owned by larger corporate land owners (U.S. Environmental Protection Agency 2005). Based on a quick cross reference with existing permit data, we did not find this to exist as the average permit size in Kentucky was larger than the average permit size for West Virginia. In any case, land ownership data for such a large study region is nearly impossible to assemble, particularly in light of the relatively coarse spatial scale of this work (1km^2 cell size). We also did not have access to adequate mineral rights data for the entire study region, another important consideration.

Active surface mine permit locations

The previously listed independent variables were analyzed with the dependent variable of location of active surface mine permits. The centroids of each permit were calculated for the model runs. Surface mining permit locations were obtained from individual state agencies (Table 1) for the ten coal-producing states within the study area (Figure 15). Mining permits were further limited to active surface mining permits only by excluding underground mines and permits associated with inactive or historical mines. In certain states, if permit status (active/inactive) was not indicated, permits were limited to those with dates from the year 2000 to the present only, in an attempt to limit analysis to current, active mines.

Table 1. Data sources and extent: active surface coal mine permits, by state.

State	Permit Data Source	Active Surface Polygons (n)	Active Surface Permit Area (km^2)
Alabama	Alabama Surface Mining Commission (2013b)	1073	257.8
Illinois	Illinois State Geological Survey (2012)	18	0.1
Indiana	Indiana Department of Natural Resources, Division of Reclamation (2013)	33	331.8
Kentucky	Kentucky Division of Mine Permits (2012)	1736	2594.5
Maryland	Maryland Department of the Environment (2012)	13	6.3
Ohio	Ohio Department of Natural Resources (2013b)	309	423.2
Pennsylvania	Pennsylvania Department of Environmental Protection (2012)	626	405.6
Tennessee	Office of Surface Mining Reclamation and Enforcement (2013)	99	85.4
Virginia	Virginia Department of Mines Minerals and Energy (2013)	90	1.6
West Virginia	WV Department of Environmental Protection (2013)	1524	1022.3

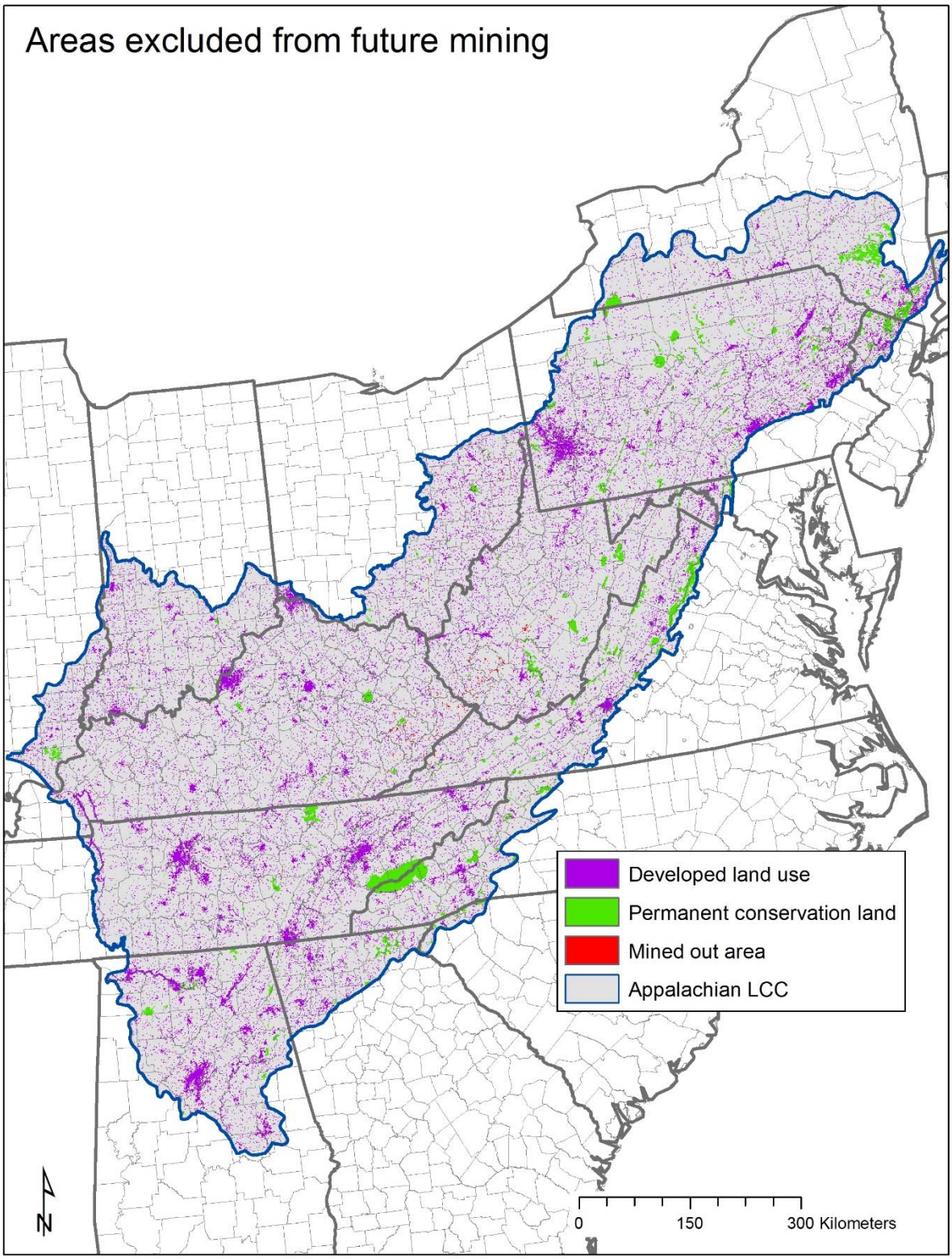
Exclusion areas

In addition to the above mentioned predictor variables and surface mine permit locations we also integrated spatial data sets as "exclusions" or areas where surface coal mining could not occur. Areas excluded from our predictive modeling of future surface mining include permanent conservation lands and areas with existing land uses that are not conducive to mining activities (urban and developed lands, water) based on the 2006 National Land Cover Dataset (Fry et al., 2011). For purposes of this work, we considered permanent conservation lands to be (in most cases) lands compiled in the Conservation Biology Institute's Protected Areas Database (Conservation Biology Institute 2012) with Gap Analysis status 1 or 2. Some adjustments were necessary for erroneously classified areas. Conservation lands with Gap Analysis status 1 and 2 (U.S. Geological Survey 2011) generally indicate areas with permanent protection from land use conversion and/or management plans designed to limit disturbance and may include national parks, national wildlife refuges, state parks and preserves, and U.S. Forest Service wilderness areas (among others), although further assessment of outstanding mineral leases on these tracts may result in their re-inclusion in the area where mining may occur. In all, 57,185 km^2 throughout

the study area (9.6%) was excluded due to land use restrictions, while 14,366 km^2 of the study area was excluded due to presence of conservation lands (2.4%) (Figure 16).

We also identified areas with an extensive recent history of surface mining for exclusion from future mining, as we are assuming these areas to be “mined out”, meaning they will not be surface mined again in the future. Mined out areas were identified as cells within current active surface mine permits (Figure 15) that were classified as Barren land cover in the 2006 National Land Cover Dataset (Fry et al. 2011). This method ensured we were capturing large contiguous areas of previous surface mining, and not newly opened mines (since we were using 2006 land cover). By using this method, we excluded mining on a total of 567 km2, or 12% of the area contained within active surface mine permits.

Figure 16. Exclusion areas for modeling process – conservation lands, land use restrictions, past mining.



2. 2 RANDOM FORESTS PREDICTIVE MODEL

Background

The machine learning algorithm Random Forests was used to produce a probability score for each of the 1km^2 grid cells within the Appalachian LCC containing coal. A higher probability score indicates a greater likelihood of future mining. This algorithm offers many advantages: it does not require any assumption of data distribution, it can handle categorical predictor variable and predictor variables with different scale, it runs efficiently on large datasets, it is robust to outliers and noise, it estimates the importance of the predictor variables in the model, and it only requires two user-defined parameters (Cutler et al. 2007) (Lawrence et al. 2006), (Peters et al. 2007), (Pino-Mejías et al. 2010), (Prasad et al. 2006).

Random Forests functions as an ensemble (multiple) of decision trees. A decision tree is based on a hierarchical concept in which decision rules are applied to segment the data using recursive partitioning into more homogeneous subsets, producing rules that define classes. A single decision tree does have shortcomings; for example, changes in the training data can induce a high variance in the classification and result in low classification accuracies. Also, they can over-fit against the calibration data. Ensemble methods, including boosting, bagging, and Random Forests, have been developed to address these shortcomings (Breiman 2001).

Random Forests functions as an ensemble (multiple) decision trees as a means to improve upon the accuracy of a single tree. Instead of using all the training data in each tree, a bootstrap sample (or random subset) of the training data is drawn for each tree. This is known as boosting. Also, it uses only a random subset of the predictor variable in each tree instead of all variables. This is done to decrease the correlation between trees, which decreases the generalization error. Random Forests allows for a group of weak classifiers to function as a strong classifier.

Because Random Forests only uses a subset of the data in each tree (the bootstrap sample), some of the data are withheld. The withheld data are called out-of-bag (OOB) data. In order to assess the importance of a predictor variable in the model, the variable is withheld and the model is rerun without the variable present. The OOB data are classified, and the mean decrease in accuracy (once the variable is removed from the model) for the classification of the OOB data provides an estimate of the importance of that variable.

Random Forests differs from other ensemble methods based on how the ensemble is generated. Also, instead of using all the predictor variables in each tree, only a subset of the variables is used, and the best variable from the subset is selected for splitting the data at each node. This results in a decrease in the strength of a single tree; however, the correlation between trees is reduced. As a result, the randomized predictor variable selection reduces the generalization error (Breiman 2001).

Application of Random Forests in this project

In order to predict the probability of surface mine occurrence in a given pixel, eleven predictor variables were used: population density (floating point raster grid); distance to railways, ports, power plants, and intermodal transportation facilities (integer raster grids); coal-bearing geology type (categorical raster with 17 categories); EPA mountaintop mining removal region (categorical raster); EIA coal supply region (categorical raster); and percent sulfur content, BTU content, ash content (floating point raster grids). All grids were generated at a 1 km^2 raster resolution.

In order to run the Random Forests algorithm, presence (surface mining) and absence (no surface mining) data are required. For presence data surface mine permit centroids were used. Only mines permitted after the year 2000 were considered. In order to create more variability and decrease correlation in the model, we generated five separate sets of absence data. First, random points were generated across the coal extent of the study area, and all random points occurring within a mine permit or within 0.5 miles of a surface mine centroid were removed. Five separate random samples were drawn for the larger set. For each set of training data we used an equal number of presence and absence points (5,165 of each and 10,330 total). The same presence data were used in each set with a different set of absence points.

Each training point was labeled as presence (with an existing surface mine permit) or absence. All of the predictor variables were appended to the points from the raster cell at that location using the software tool Geospatial Modeling Environment (http://www.spatialecology.com/gme/).

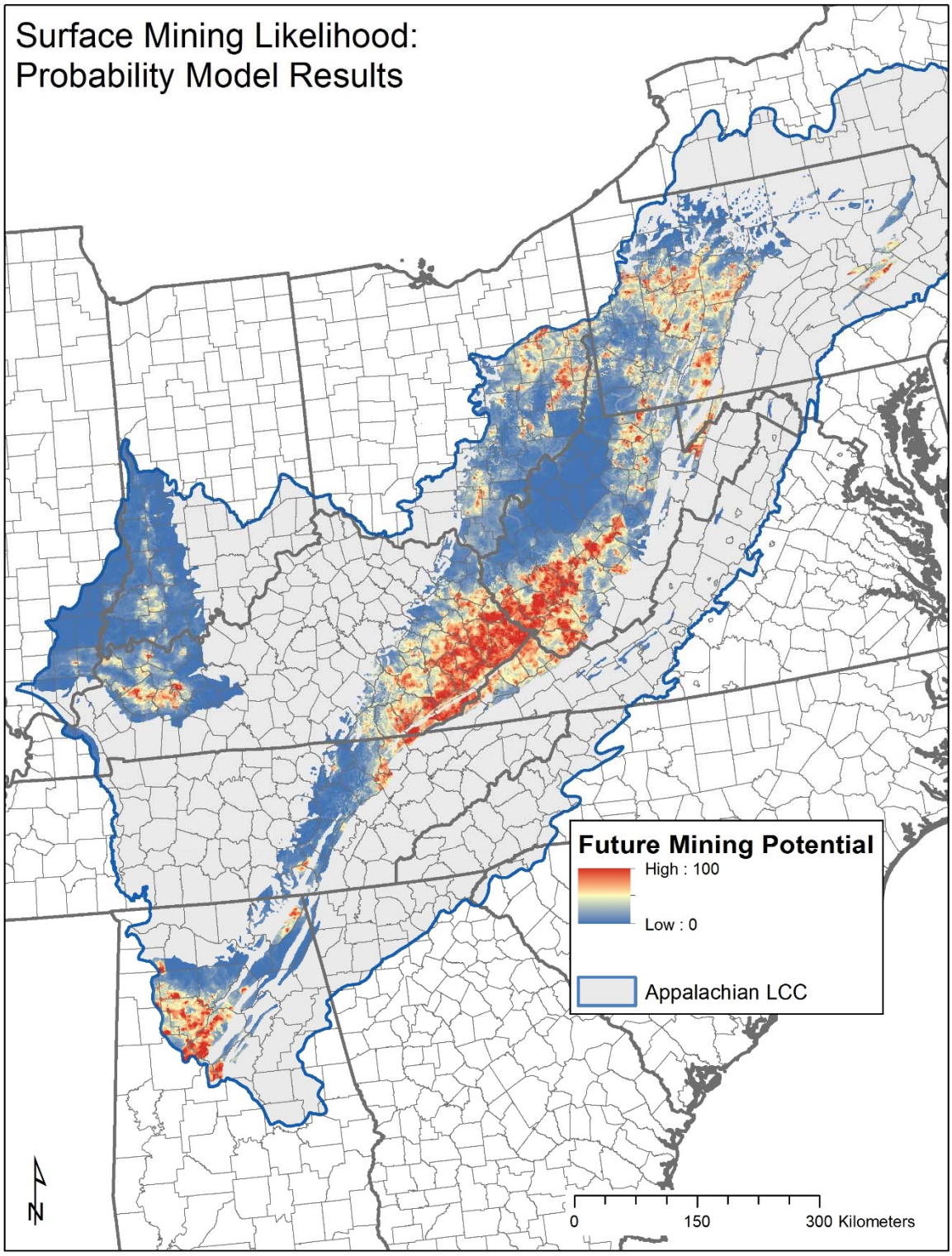
The data were then read into the statistical package R. The Random Forests algorithm was executed using the RandomForest package. A separate model was generated for each of the five training sets. For each model 1,000 trees were generated and the number of randomly selected predictor variables sampled in each tree was set to 3 (the square root of the number of bands, which is the default setting). All five of the models were then combined to produce a model containing 5,000 trees.

Using the final combined model, all 1 km^2 pixels in the coal extent study area were classified using the Random Forests model. The probability of the pixel being a surface mine was reported, and a raster grid output was generated.

The resulting (0 to 100) prediction is shown in Figure 17. As estimated by the out-of-bag mean decrease in accuracy, the coal geology type and the sulfur content were found to be the most important predictor variables in the model, though all variables contributed (Figure 18). In Random Forests modeling, the out-of-bag error estimate is an internal measure of the random Forests “tree” based on samples not used to build a particular classification tree (Breiman 2001). For each training dataset, the out-of-bag error estimate was around 15% and the misclassification of presence and absence points were evenly balanced. Plotting the error rate against the number of trees generated suggests that 1,000 trees per set is more than ample to stabilize the result.

The code used for running Random Forests in the statistical program R and additional output from the Random Forests model are provided in the Appendix section of this report.

Figure 17. Surface mining probability from Random Forest model results. Areas in red have highest probability of new surface mining activity, areas in green have the lowest. Model extent was limited to the known extent of coal in the region (USGS coal fields).



2.3 PREDICTIVE MAPPING: FUTURE SURFACE MINING FOOTPRINT

In order to map future potential surface mining activities on a landscape scale, we used results from the probabilistic Random Forests modeling of surface mine potential along with regional-level estimates of future coal mining production for the years 2012 through 2035.

Regional coal production estimates for the four EIA coal supply regions (Northern, Central and Southern Appalachians, Eastern Interior/Illinois) were obtained using various coal production scenarios from the EIA's Annual Energy Outlook (U.S. Energy Information Administration 2013a), Table 68 Annual Energy Outlook coal production by region and type. Values were obtained for two different EIA economic/coal production scenarios for comparison: a low coal production scenario and a high coal production scenario. The low coal production scenario ("GHG25+low gas") predicts the lowest future coal production of any of EIA's 28 total scenarios, due to very restrictive greenhouse gas emissions policies and low prices for competing resources of oil and gas. The high coal production scenario ("low coal cost") predicts the highest coal production due to lower costs for coal mining wages, transportation, and mine equipment (leading to increased coal production).

EIA coal production estimates provide total production estimates only (surface and underground combined). We converted future production projections to surface projections only by multiplying each production total by the percentage surface according to the following regional figures (based on 2010-2011 production data in the Annual Energy Outlook): Northern Appalachians: 20.08% surface, Central Appalachians: 48.68% surface, Southern Appalachians: 40.06% surface, Eastern Interior/Illinois: 30.72% surface. Surface mining production estimates from the year 2012 through the year 2035 were then summed to produce a total cumulative surface coal production value for each region.

In order to estimate surface area impacted by coal mining activities, we required a numeric relationship between surface mine production amounts and a corresponding area disturbed. We initially proposed using current active surface mine permit data along with recent production statistics in order to derive a production to area ratio. However, single mines may produce coal for extended periods of time, and this method would not adequately capture the entire life cycle of a mine. In addition, mapped mine permit polygons may include areas that are not actually disturbed during surface mining, so the actual disturbed area may be much smaller than mapped permit area. A recent government study concluded that mapped mine permits do not offer an accurate way to estimate area disturbed by surface mining, based on current permit database and mapping methods used in WV and KY (GAO 2009). Instead, Lutz et al. (2013) developed a regression model to estimate tons of coal produced per unit areal disturbance for 47 counties in southern WV and eastern KY. The model was based on total area of surface mining disturbance from 1985-2005 (at 5 year time intervals), compared with surface coal production statistics for corresponding time period. Lutz et al. (2013) estimated that 1 ton of coal equates to 0.87m2 of surface disturbance. For the current study, this figure was converted to 1.15 million tons of coal produced per square kilometer of surface land disturbance.

Future surface mining scenarios analyzed included low coal production and high coal production models (US Energy Information Administration 2013a) for the years 2012-2035. For each scenario, we created a new map layer showing potential locations for future surface mining activities on a cell-by-cell basis using a 1 km^2 grid for the study area. Using the figure of 1,150,000 short tons per km^2, we allocated future mining production on a cell-by-cell basis within each EIA region first to those cells with the highest future mining probability, then continuing to cells with lower future mining probability, until the

total amount of future production for a particular scenario and region was allocated. Prior to allocation, adjacent cells with identical mining probability values were grouped together to ensure that contiguous areas of high mining probability were preserved in the results (rather than assigning "new" mining to single cells). Cells containing urban or built up land, water, conservation lands, and centroids of existing mining permits were excluded (masked out) prior to build-out analysis as described earlier in this report.

3. RESULTS

3.1 RANDOM FORESTS MODEL (PROBABILITY OF FUTURE SURFACE COAL MINING)

The final Random Forests model scenario included a total of ten predictor variables: the continuous variables of coal ash content, coal BTU content, distance to railroads, distance to power plants (along road network), distance to ports (along road network), distance to intermodal transportation facilities (along road network), population density, and the categorical variables of EIA region, EPA Mountaintop Removal mining region, and coal geology type. We considered removing low-performing variables from the final model, based on variable contribution to the overall result. However, alternative models with fewer variables did not perform as well as the full model, producing higher classification error rates. Model significance was tested vs. randomly generated models and was found to be significant with a p-value of 0.0101.

The final output of the Random Forests model is a pixel based probability of future surface mining presence. The final probability values were re-scaled from 0 to 100 (Figure 17). Results indicate that the highest probability of future surface mining is found in the Central Appalachian region, particularly throughout southwestern West Virginia and eastern Kentucky. Other pockets of higher probability are found in western Kentucky, central Alabama, and to a lesser extent, north central West Virginia and the bituminous coal region of Pennsylvania and Ohio.

The total area within each EIA coal supply region with relatively high probability (over 90) is listed in Table 2. The Central Appalachian region has the highest percentage of high probability areas for the four regions, while the Northern Appalachian and Easter Interior/Illinois regions have a very small percentage of their area within high probability – 90% or higher (Figure 19). Note that while the Northern, Central and Southern Appalachian regions lie completely within the current study boundary (Appalachian LCC), the Eastern Interior / Illinois coal supply region also includes production in portions of western and central Illinois and Mississippi that are not included in the Appalachian LCC study area for this project. Based on the most recent available coal production statistics from 2011 (U.S. Energy Information Administration 2012a), there are a total of six counties in the Eastern Interior/Illinois region that produce coal but are located outside of the project study region. For 2011, these six counties accounted for 11.7% of the total surface coal production for the Eastern Interior/Illinois region (so approximately 11-12% of coal production in this region will not be accounted for in our model results and projections).

Figure 18. Importance of predictor variables measured as out-of-bag mean decrease in accuracy.

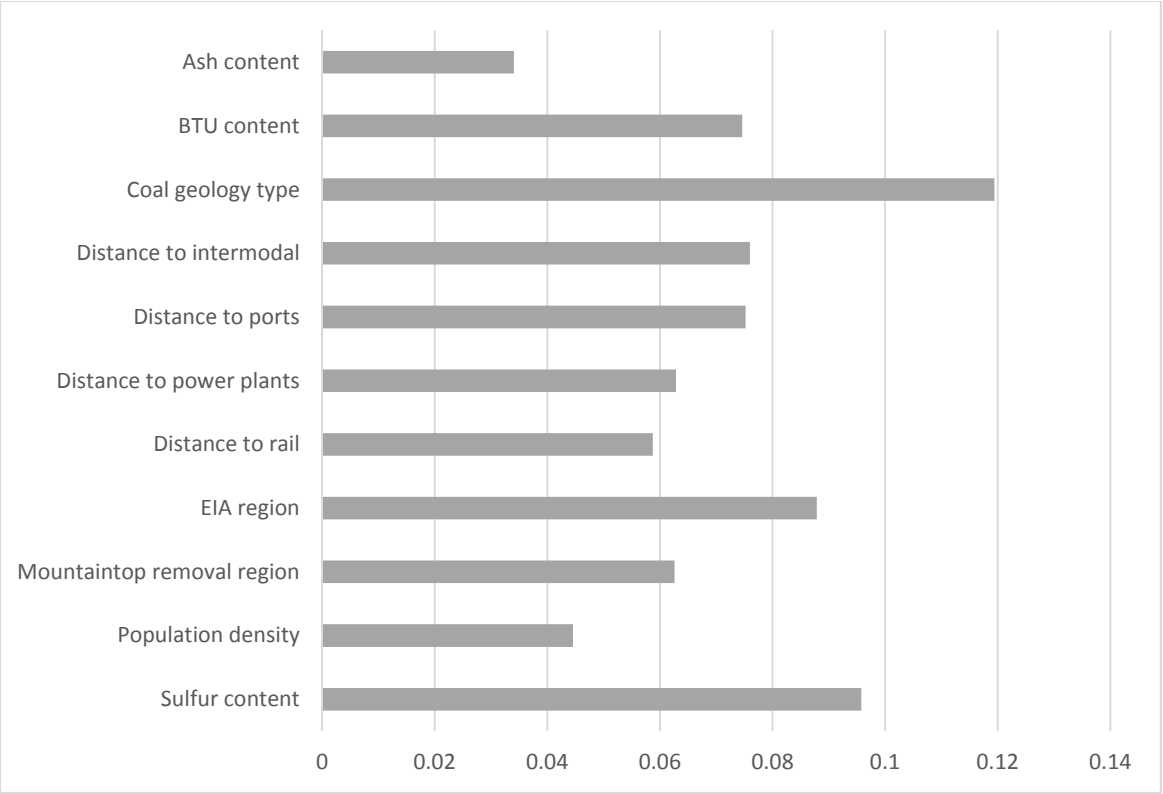


Figure 19. Random forest result – high probability areas (90%+) with high likelihood of future surface mining.

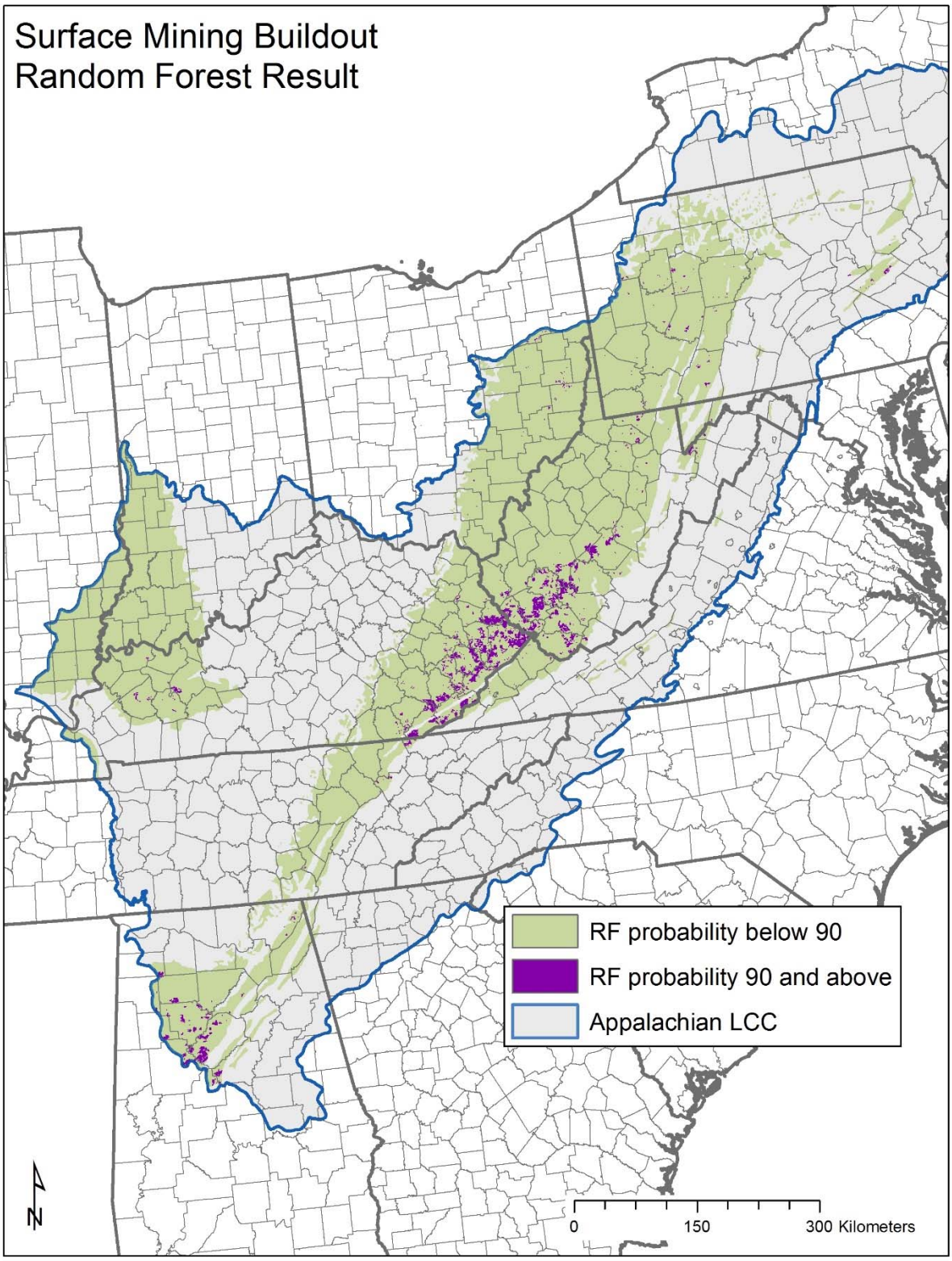


Figure 20. Low coal production scenario: Future mining footprint for low coal production model through 2035 (based on EIA GHG25+low gas price scenario).

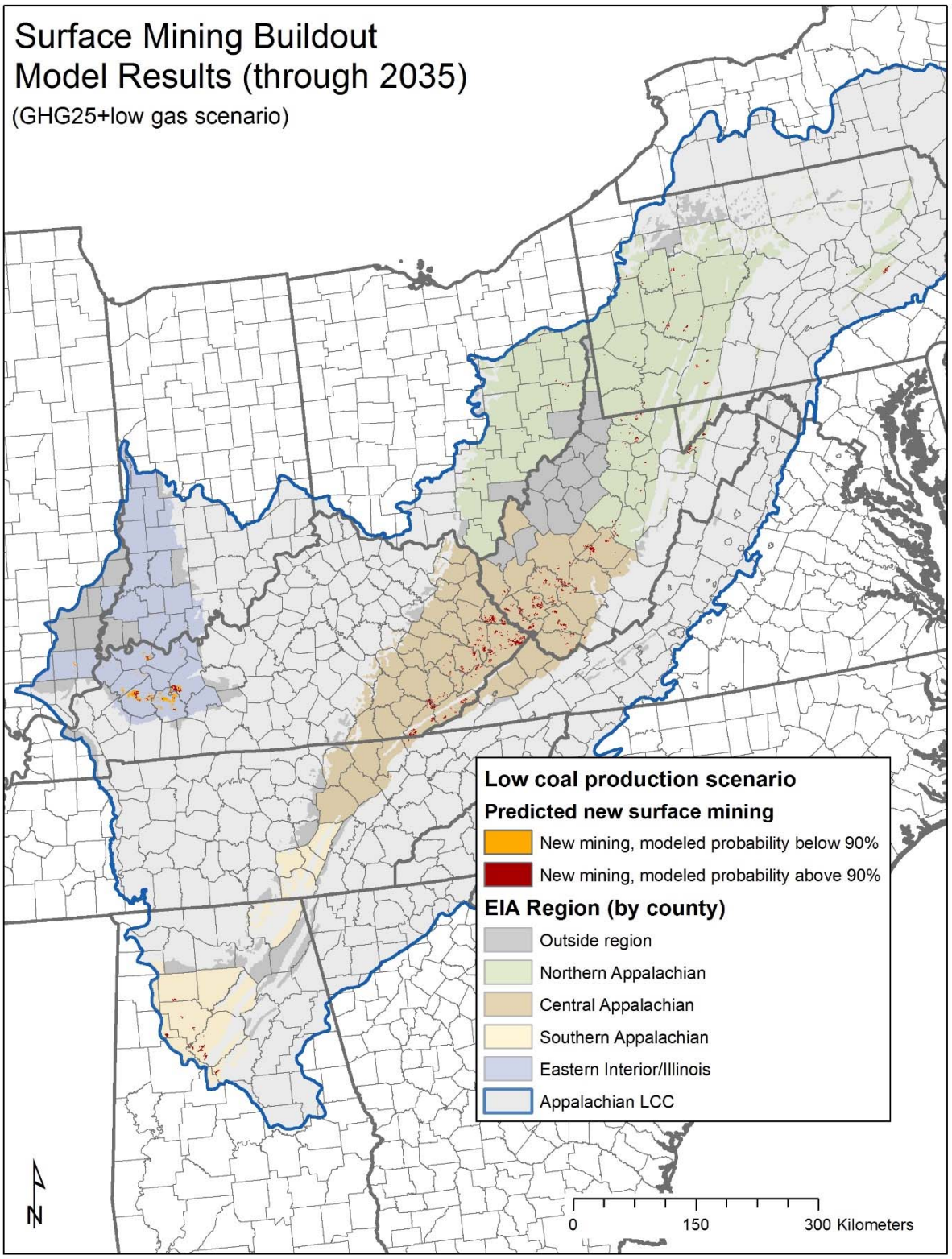


Figure 21. High coal production scenario: Future mining footprint for coal production through 2035 (based on EIA low coal production cost scenario).

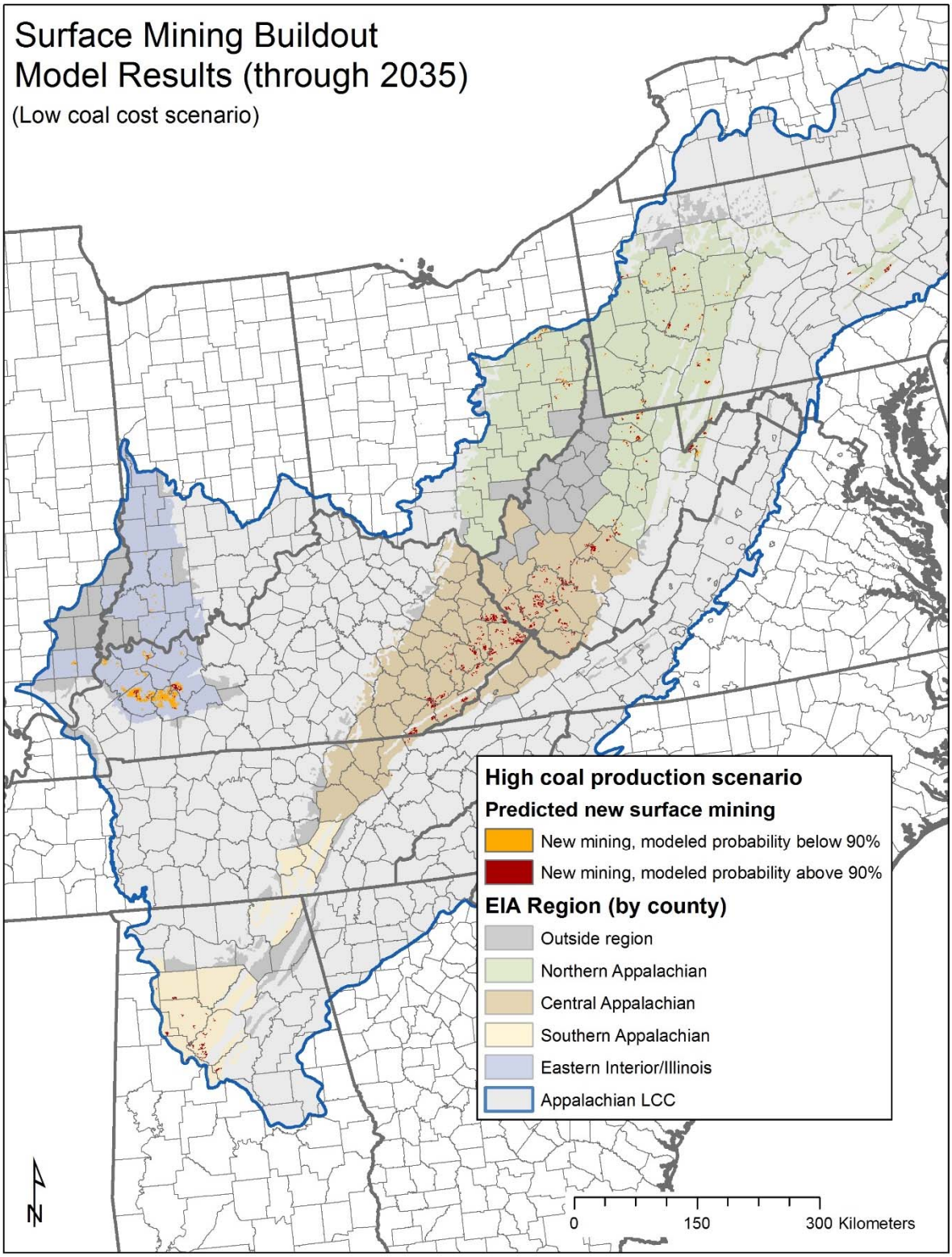


Table 2. EIA coal supply regions, with area of relatively high (90% or higher) probability of future surface coal mining, based on Random Forests model results.

Region Name	Area (km2)	Area < 90 (km2)	Area >= 90 (km2)	Percent < 90	Percent >= 90
Northern Appalachian	68,852	68,385	467	99.32	0.68
Central Appalachian	53,788	49,368	4,420	91.78	8.22
Southern Appalachian	14,455	13,635	820	94.33	5.67
Eastern Interior/Illinois	28,147	28,048	99	99.65	0.35

3.2 PREDICTIVE MAPPING: FUTURE SURFACE MINING FOOTPRINT

Results for future surface mining footprint by the year 2035 are shown in Figures 20 (low coal production model: GHG25+low gas) and Figure 21 (high coal production model: low coal cost). Total area (km^2) mapped as new surface mining activity is listed by EIA region in Table 3. We also calculated the percentage of the high probability area (defined for this study as areas with Random Forests model probability at 90 or higher) affected by new mining for each region. For the low coal production scenario, all regions except the Eastern Interior/Illinois are predicted to have all new mining footprints located completely within higher probability areas. For the high coal production scenario, only the Central and Southern Appalachian regions are predicted to have all new mining footprints found within higher probability areas.

To meet production estimates for the low coal production scenario, the three Appalachian regions are each predicted to have all new surface mining development limited to high probability modeled areas (defined for this project as Random Forests model results of 90% and above). These highest probability areas (shown in Figure 19) are concentrated in southwestern West Virginia and eastern Kentucky, with a significant portion in Alabama (Southern Appalachian region). However, surface mine footprints within the Eastern Interior/Illinois region may need to extend beyond the highest modeled probability areas in order to meet projected production figures (according to model results, the area required to meet future coal production in this region has a minimum probability score of 68 (Table 3)). Within this region, under the low coal production scenario, new mining is modeled to occur in lower probability areas concentrated within Hopkins, Henderson, Ohio, and Muhlenberg counties in western Kentucky (Figure 20).

In order to meet the future high coal production scenario, the area associated with future coal production for both the Eastern Interior/Illinois and the Northern Appalachian regions exceeds the current high probability area for those regions. In the Northern Appalachian region, in order to meet high coal production predictions, new mining is modeled to extend into areas with a minimum model probability of 85 (Table 3). These areas are found scattered across counties in eastern Ohio, western Pennsylvania, and north central West Virginia (Figure 21). In the Eastern Interior/Illinois region, new mining is modeled to extend into areas with a minimum model probability of 54 (Table 3). Within this region, new mining areas are again concentrated in western Kentucky, with smaller amounts in Illinois (similar to the low coal production scenario).

Table 3. By EIA coal supply region, total area mapped as new surface mining under differing scenarios (low coal production, high coal production). High probability areas are those areas with future surface mine probability from Random Forests model at 90% or greater. Minimum Random Forest probability result for mapped new mining areas is given by region.

Region name	Low coal production scenario			High coal production scenario		
	Mapped new mining (km^2)	Percent of high probability area	Minimum probability score in new mining	Mapped new mining (km^2)	Percent of high probability area	Minimum probability score in new mining
Northern Appalachian	272	58.24	91	776	166.17	85
Central Appalachian	953	21.56	97	1186	26.83	96
Southern Appalachian	103	12.56	99	148	18.05	99
Eastern Interior/Illinois	453	457.58	68	991	1001.01	54

3.3 ASSESSMENT OF RESULTS

In an effort to compare our model results for locations of future surface coal mining activity with established data, we compare our results with three related sources of data: coal seam level data (coal availability/thickness), remaining coal reserves, and newly permitted areas.

Coal seam level data

Data on individual coal seams are available from multiple state geological survey agencies. Mapped coal seam properties include coal seam depth to top of the seam, seam thickness, and overall coal availability. Mapped properties vary by state and seam.

Within Illinois, the Illinois State Geological Survey has mapped coal seam properties for several of the state's prominent coal seams. According to seam-level reports (Treworgy et al. 1999, Treworgy et al. 2000, Korose et al. 2002), most Illinois seams have more limited resources available for surface mining, with the majority of remaining coal being underground. Of mapped coal seams, the Herrin, Danville, and Dekoven-Davis seams all have polygons indicating coal available for surface mining within the Appalachian LCC study area (ISGS 2013). Figure 22 shows surface coal availability for these three seams, overlaid with predicted locations of new surface mining from this project. There is high correspondence between areas of mapped availability of surface coal resources with the results of this study, particularly in Saline County IL. Restrictions to surface coal development listed in the ISGS datasets include depth to coal seams, unfavorable overburden or stripping ratios, land cover restrictions, and mined out areas, among others.

Within Indiana, coal availability data (seam thickness, depth to seam) are available for a small number of coal seams. Figure 23 shows the depth to the Springfield coal seam across southwestern Indiana (Indiana Geological Survey, 2000). The depth to the coal seam increases from east to west. This corresponds in general with the results of the current model, which shows increased likelihood of future surface mining toward the west-central portions of Indiana (Pike, Daviess, Warrick counties) rather than the extreme western border of the state.

USGS data on overburden and seam thickness are available for six major producing seams in the Appalachian (U.S. Geological Survey 2000) and three major seams of the Illinois coal regions (Hatch and Affolter 2002). Of these seams, three Appalachian seams (Pittsburgh, Upper Freeport, and Fire Clay) have some areas with less than 200ft of overburden which may theoretically be available for future surface mining. A comparison of the level of overburden of these three seams with modeled potential areas for future surface mining (model results) is presented in Figure 24. The model tends to predict that future surface mining will be concentrated in areas of lower overburden, particularly for the Pittsburgh seam. Similar results are seen for the three mapped seams in the Illinois region (Baker-Danville, Herrin, and Springfield coals): the model predicts future surface mining to be more prevalent in areas of lower overburden (Figure 25).

Figure 22. Coal availability for Illinois for Danville, Herrin, and Dekoven-Davis seams (surface minable coal) compared with predicted new surface mining from Random Forest model result for high coal production scenario.

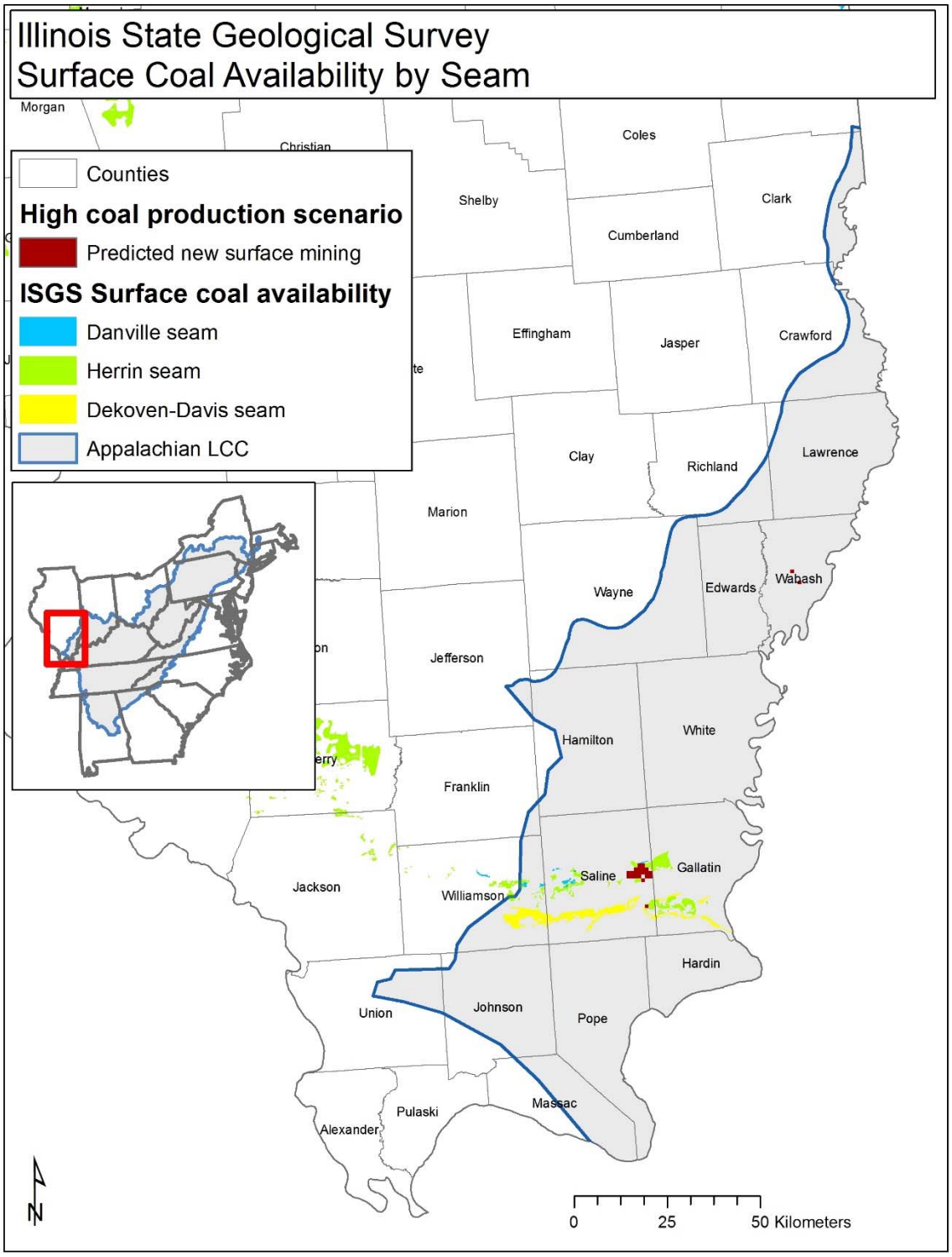


Figure 23. Coal availability for Illinois for Springfield coal seams (depth to coal) compared with predicted new surface mining from Random Forest model result for high coal production scenario

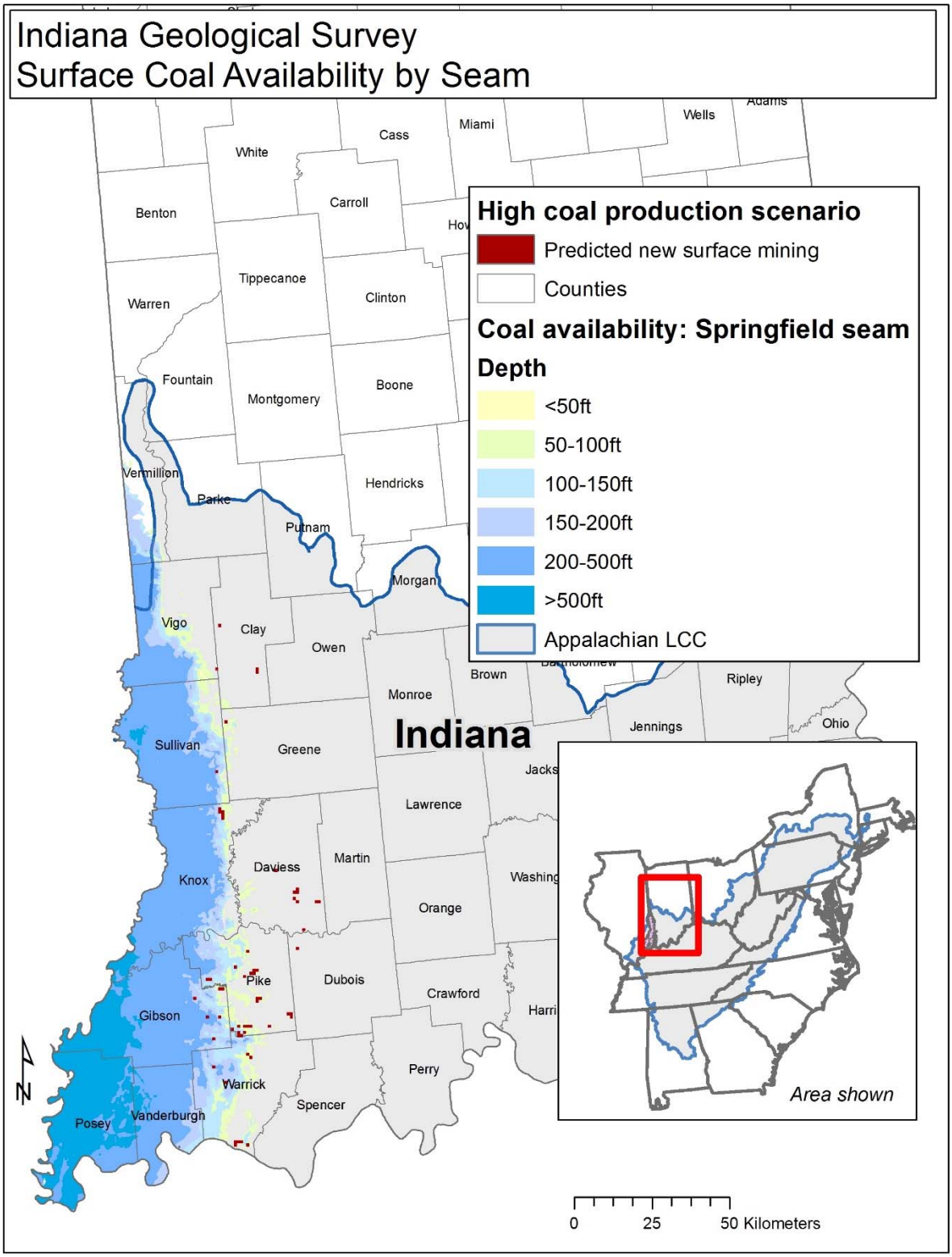
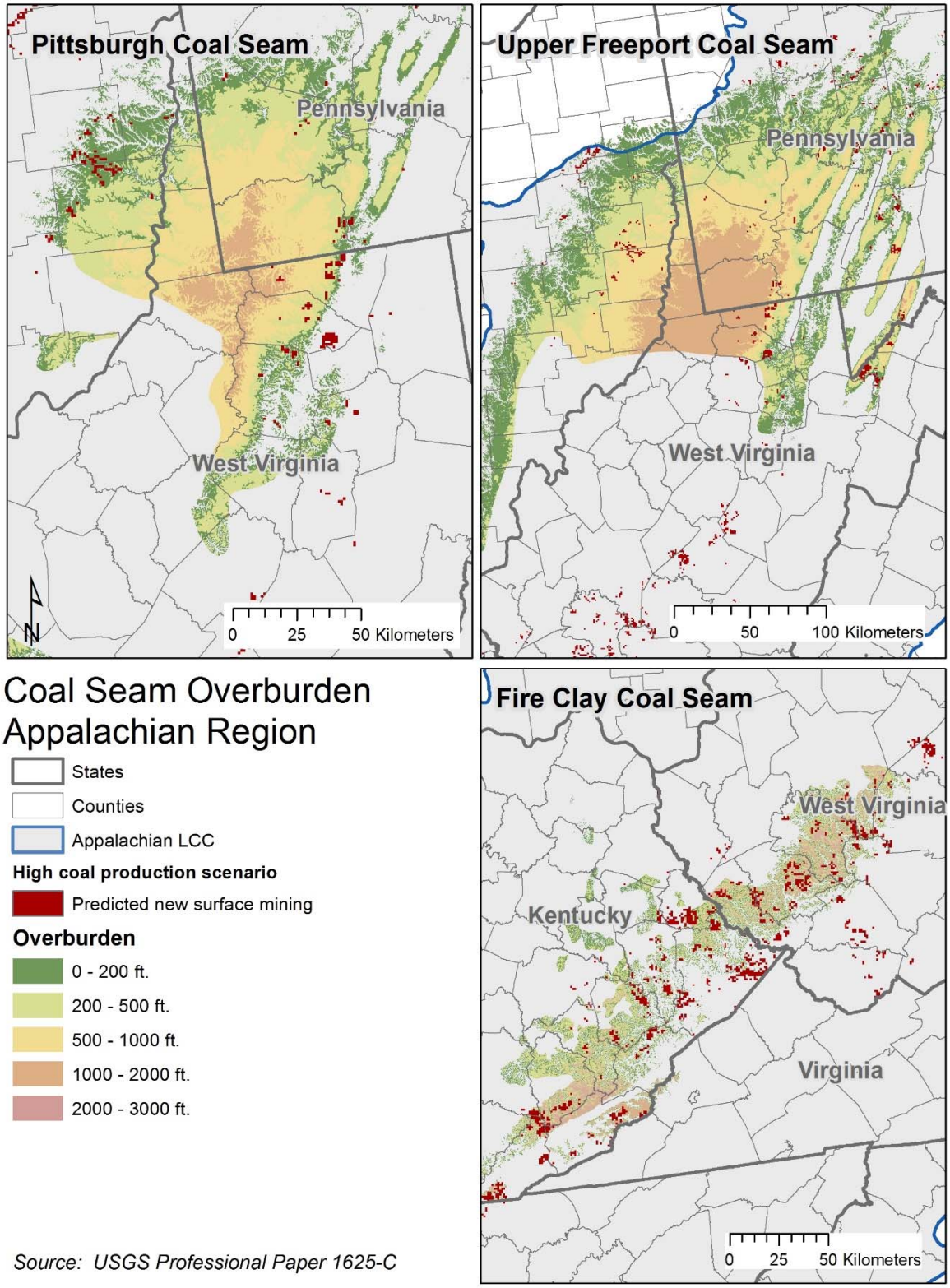
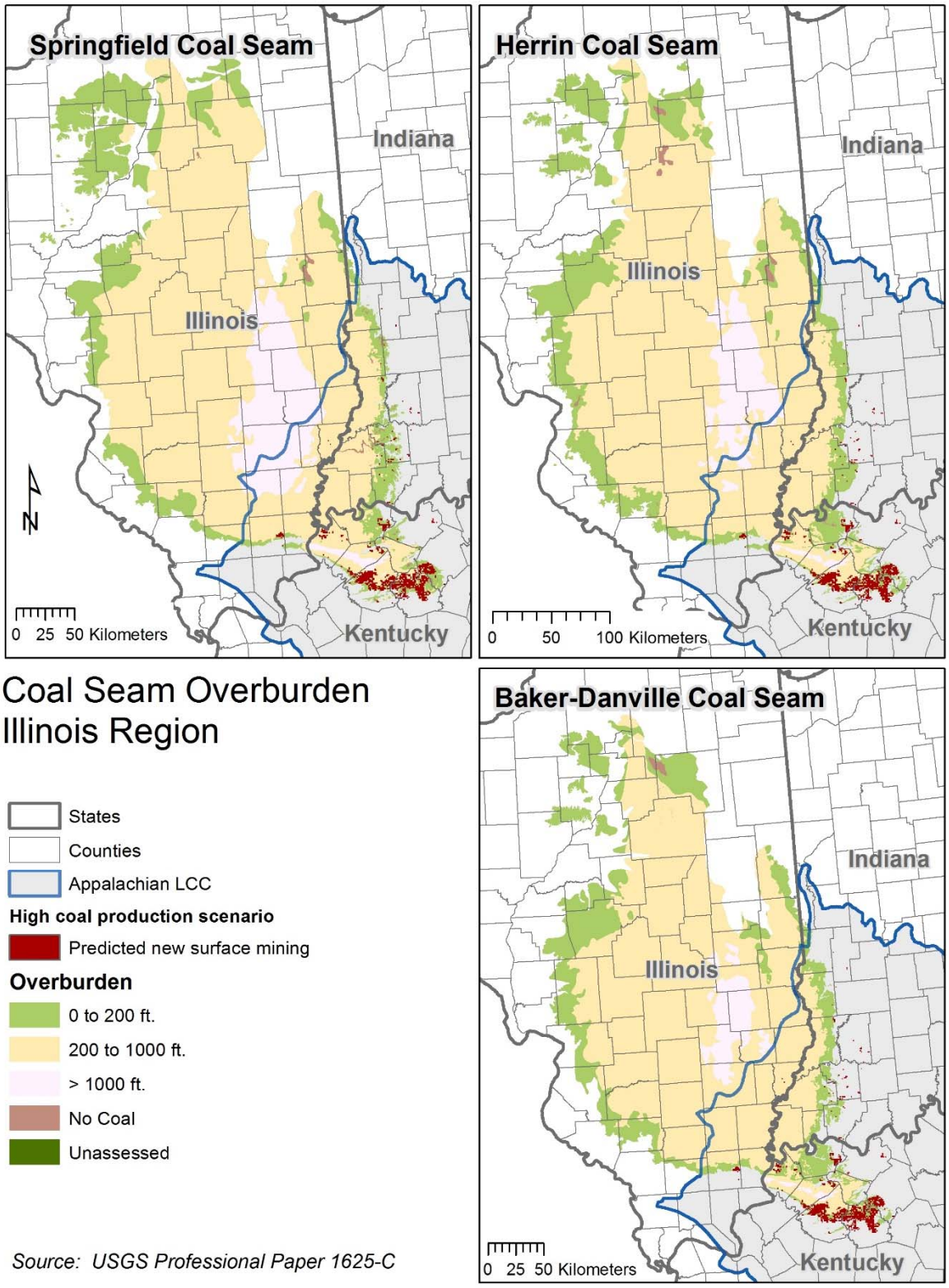


Figure 24. Comparison of model results (high coal production scenario) with existing data on coal seam overburden for three coal seams in the Appalachian region.



Source: USGS Professional Paper 1625-C

Figure 25. Comparison of model results (high coal production scenario) with existing data on coal seam overburden for three coal seams in the Illinois region.

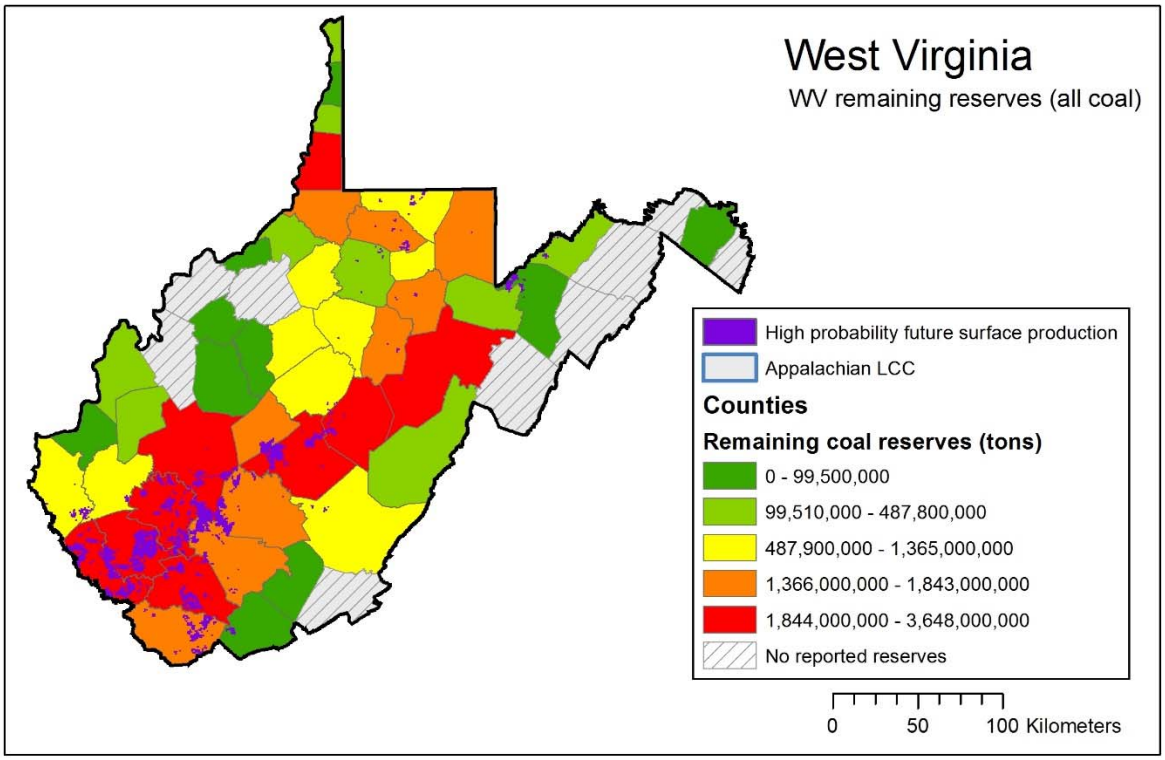


Remaining coal reserves data

Comparison of model results with published coal reserve figures indicates close correspondence between areas of future high surface coal production (from this model) and established coal reserves. County-level coal reserves (amount of remaining coal) have been published for many of the states within the Appalachian LCC study area. Comparison of model results for future surface mining probability may be corroborated with county reserve data for West Virginia, Pennsylvania, Kentucky, and Ohio.

Within West Virginia, reserve data available from the WV Coal Association (for all types of coal – surface and underground) (WV Coal Association 2012) indicate that areas mapped as high probability for future surface mining bear a strong correspondence with counties with high remaining reserves in southwestern West Virginia (Figure 26). One county modeled to have future surface production lacking a high amount of reserves, however, was Grant County (in the eastern panhandle). Within Pennsylvania, high coal reserve counties in the southwestern portion of the state also have high probability areas for future mining (Figure 26). Pennsylvania coal reserve figures (PA Coal Alliance 2011) do not include anthracite coal. Kentucky coal reserves (including non-recoverable coal) (Kentucky Foundation 2013) are shown in Figure 27, and comparison with Ohio coal reserves (surface only) is also shown in Figure 27 (Ohio DNR 2013a).

Figure 26. Coal reserves as reported by county, West Virginia and Pennsylvania, compared with model results (high coal production scenario).



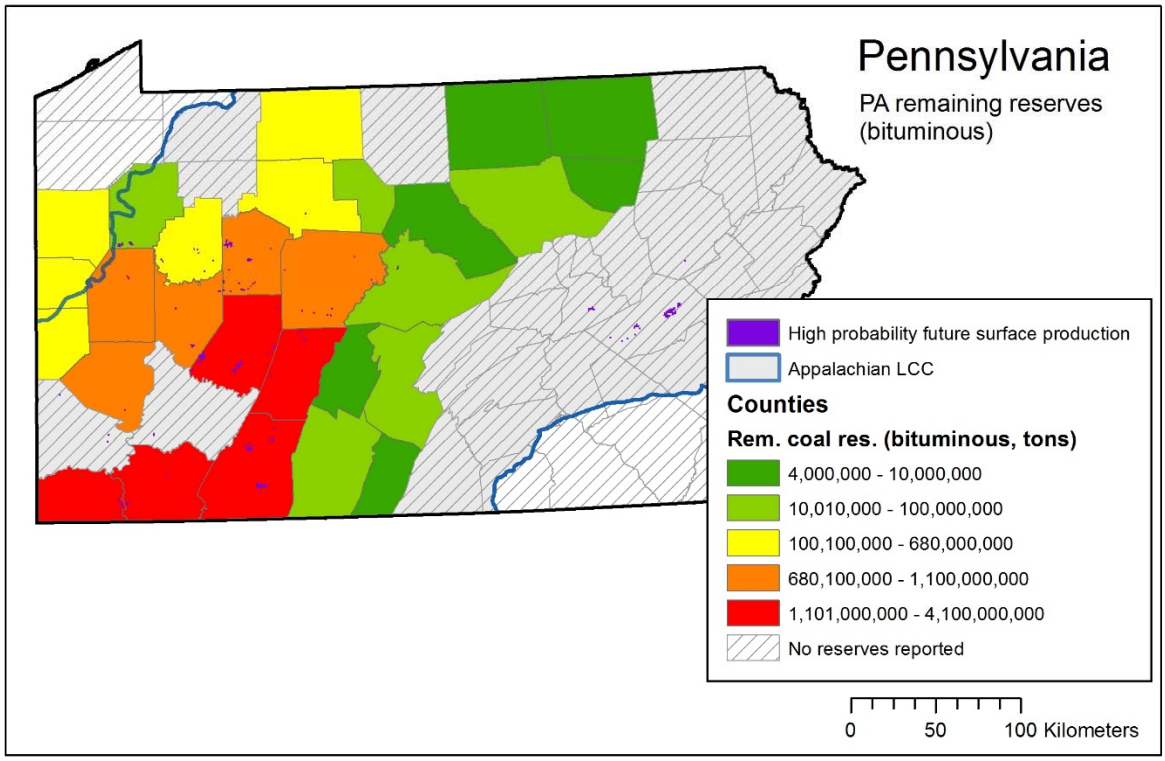
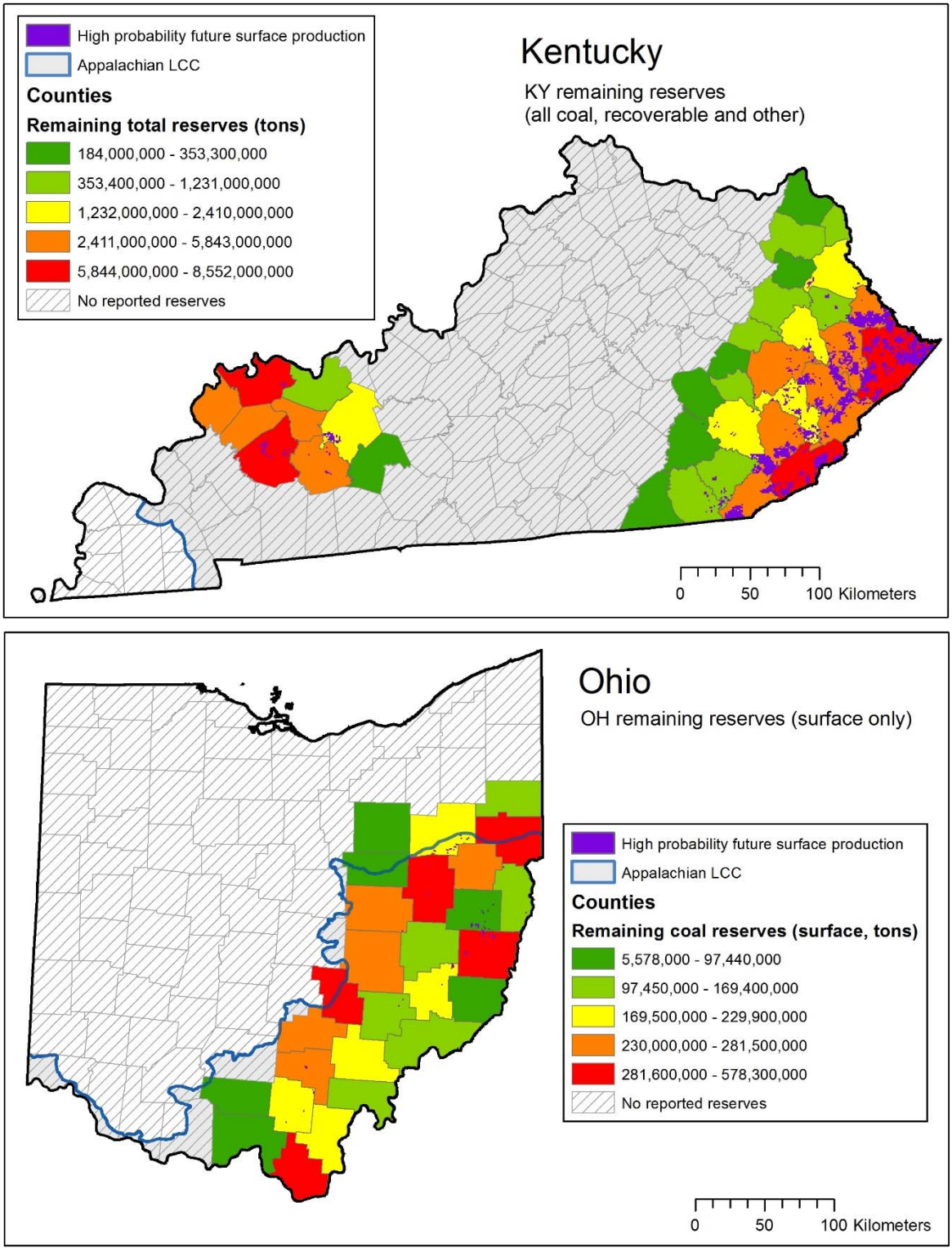


Figure 27. Coal reserves as reported by county, Kentucky and Ohio, compared with model results (high coal production scenario).



Newly permitted areas

Areas of recent surface mine permit activity may also be used to qualitatively evaluate model results. Areas modeled to have high probability of future surface mining should theoretically be associated with areas of high current permit activity (newly approved permits, permits approved but not yet started etc.). Recent permit activity is available for Alabama and West Virginia. Within Alabama, the Alabama Surface Mining Commission lists recent permit decisions, including renewals, revisions and applications (ASMC 2013b). Based on information from the ASMC, there are 86 permit polygons within the Appalachian LCC study area in Alabama that have recent permit activity in 2013 (permit activity includes renewal, revision or approval). Some permits consist of more than one polygon. Of these 86 polygons, 77 (89.5%) intersect areas of high future mining likelihood (probability over 90 as modeled) (Figure 28. Table 4). For West Virginia, recent surface permits that are mapped but have not been started yet (personal communication, Nick Schaer, WVDEP 2013) may also be used in a similar fashion. Within West Virginia, there are a total of 43 surface mine permits that have been issued but have not yet been started, and of these, 26 (60.4%) intersect areas of high future mining probability as modeled (Figure 29, Table 4).

Table 4. For Alabama and West Virginia, detailed summary of mine permits with new/recent activity in relation to Random Forests (RF) model results.

State	Permits/polygons applicable	Average RF score for permits/polygons	Number of permits/polygons intersecting high likelihood areas (RF score > 90)
Alabama (permit polygons with 2013 activity)	86	96.5	77
West Virginia (permits approved but not started)	42	80.2	26

Figure 28. Surface mine permits in Alabama with recent permit activity, compared with model results (high coal production scenario).

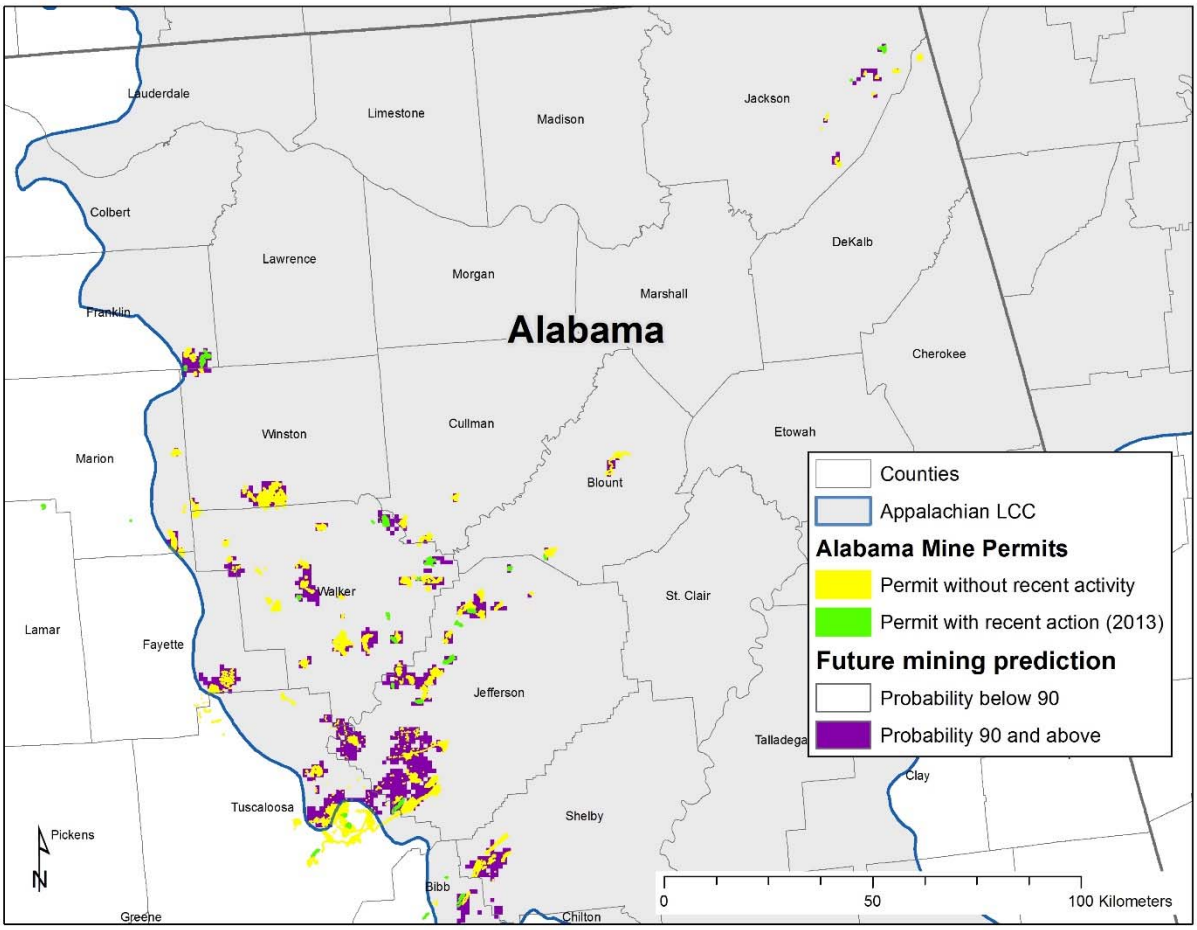
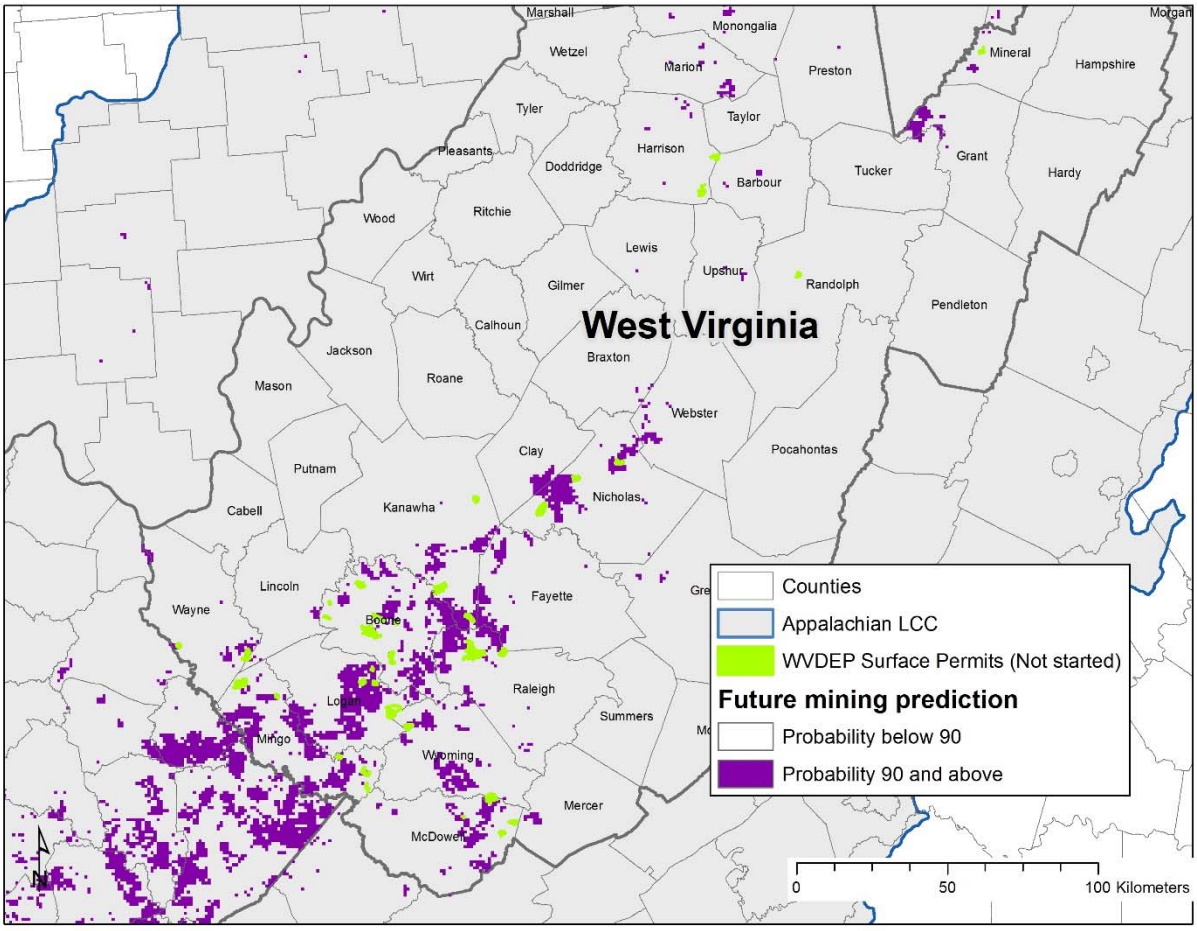


Figure 29. Recent surface mine permits in West Virginia (permits approved but not started), compared with model results (high coal production scenario).



4. CONCLUSIONS AND SUMMARY

This project maps future surface mining footprint across the Appalachian region, based on varying estimates of future coal production. The Random Forests modeling technique was used to predict areas with high likelihood of future mining. Through the modeling process, we determined that key determining factors of future mining locations at the regional scale include coal geology type, coal sulfur content, coal btu content, and distance to transportation related infrastructure. The extent of future surface mining will vary regionally, with highest probability areas concentrated in the mountaintop removal/valley fill mining region of central Appalachia.

This study provides a framework for continued surface mine predictive modeling at a more local scale of analysis. As mentioned earlier, much more detailed information could be used to predict surface mine activity, particularly in areas with more readily available detailed geological and mining spatial datasets such as West Virginia and Kentucky. In addition, future focused work in site specific areas may also consider land ownership and other variables related to topography, access roads, etc. with higher resolution than the 1 km^2 unit used in this study.

The requirement to include consistent spatial datasets available for the entire study area as well as the unit cell size limits taking our results and applying them to individual mine sites or even locations within an individual HUC 12 watershed. We feel the best use of our predicted cell locations would be summarization at the county or larger HUC 10 watershed extent. We encourage caution with any site-specific application of model results. However, even with these limitations we feel the correlation between our identified areas and actual planned surface mine permits indicates the value of our modeled results. We look forward to the analysis and summary of this information to aid in planning for future surface mine activity.

Acknowledgments

This work was supported by The Nature Conservancy (TNC) through an Appalachian Landscape Conservation Cooperative grant: Assessing Future Impacts of Energy Extraction in the Appalachian LCC. Judy Dunscomb and Brad Kreps of TNC worked closely with the WVU team in order to guide the project, answer questions, and provide technical input. Jeff Evans of TNC and Andy Tri from WVU contributed key assistance with R and Random Forests statistical modeling. The work was greatly enhanced by cooperation with an extensive team of peer reviewers assembled by TNC. We are particularly grateful for the input received from Joseph Kiesecker, Emily Medine of Energy Ventures Analysis, Gene Kitts with Alpha Natural Resources, Nick Schaer of the West Virginia Department of Environmental Protection, and Tom Galya of the U.S. Office of Surface Mining.

References

Alabama Surface Mining Commission (ASMC). 2013a. Permit decisions issued by ASMC. Alabama Surface Mining Commission, Jasper, AL. URL: http://surface-mining.alabama.gov/PermitDecisions.html. Retrieved 10-18-2013.

Alabama Surface Mining Commission (ASMC). 2013b. Alabama coal mine geospatial data. Alabama Surface Mining Commission, Jasper, AL. URL: http://www.surface-mining.state.al.us/. Retrieved 3/25/13.

Bragg, L.J., Oman, J.K., Tewalt, S.J., Oman, C.L., Rega, N.H., Washington, P.M., and Finkelman, R.B. 1997. U.S. Geological Survey Coal Quality (COALQUAL) Database: Version 1.3: U.S. Geological Survey Open-file Report 97-134. URL: http://energy.er.usgs.gov/products/databases/CoalQual/intro.htm. Retrieved 10-21-2013.

Breiman, L., 2001. Random Forests, Machine Learning, 45(1): 5-32.

Bureau of Transportation Statistics. 2011. National Transportation Atlas Database 2011. United States Department of Transportation, Research and Innovative Technology Administration, Washington, DC. URL: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_atlas_database/2011/index.html. Retrieved 6/18/13.

Conservation Biology Institute. 2012. Protected Areas Database – United States, version 2. Conservation Biology Institute, Corvallis, OR. URL: http://consbio.org/products/projects/pad-us-cbi-edition. Retrieved 6/18/13.

Cutler, D.R., T.C. Edwards, Jr., K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler, 2007. Random forests for classification in ecology, Ecology, 88(11): 2783-2792.

EPA (2012) "Cross-State Air Pollution Rule." U.S. Environmental Protection Agency. Accessed online: http://www.epa.gov/airtransport.

ESRI 2012. U.S. National Transportation Atlas Railroads. ESRI Data & Maps 2012. Environmental Systems Research Institute, Redlands CA.

ESRI 2012a. U.S. Highways. ESRI Data & Maps 2012. Environmental Systems Research Institute, Redlands CA.

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J., 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States, *PE&RS*, Vol. 77(9):858-864.

Government Accounting Office. 2009. Surface coal mining characteristics of mining in mountainous areas of Kentucky and West Virginia. United States Government Accountability Office. GAO 10-21, Washington, DC. URL: http://www.gao.gov/new.items/d1021.pdf. Retrieved 9-23-13.

Hatch, J.R., and Affolter, R.H., 2002, Resource Assessment of the Springfield, Herrin, Danville and Baker Coals in the Illinois Basin: U.S. Geological Survey Professional Paper 1625-D. [CD-ROM]. http://greenwood.cr.usgs.gov/energy/coal/PP1625D/

Illinois State Geological Survey. 2012. Illinois coal resource shapefiles. Illinois State Geological Survey, Prairie Research Institute, University of Illinois at Urbana-Champaign, Champaign, IL. URL: http://www.isgs.uiuc.edu/maps-data-pub/coal-maps/coalshapefiles.shtml. Retrieved 3/5/13.

Indiana Department of Natural Resources, Division of Reclamation. 2013. Indiana DNR mine permit data provided by Forrest Brown, GIS Specialist, IN DNR, Jasonville, IN.

Indiana Geological Survey. 2000. The availability of the Springfield Coal Member for mining in Indiana. Indiana Geological Survey open-file study 99-07, Indiana Geological Survey, Bloomington, IN. URL: http://igs.indiana.edu/arcims/statewide/download.html. Accessed 10-14-2013.

ISGS 2013. Illinois Coal Resource Shapefiles. Illinois State Geological Survey, Prairie Research Institute, University of Illinois at Urbana-Champaign. URL: http://www.isgs.illinois.edu/research/coal/shapefiles. Accessed 10-15-2013.

Johnston, Kevin, Jay M. Ver Hoef, Konstantin Krivoruchko, and Neil Lucas. Using ArcGis Geostatistical Analyst, 2001. Environmental System Research Institute, Redlands, CA.

Kentucky Division of Mine Permits 2012. Kentucky Permitted Mine Boundaries (GIS dataset). Kentucky Division of Mine Permits, Frankfort, KY. URL: ftp://data.gis.eppc.ky.gov/shapefiles/Permitted_Mine_Boundaries.zip. Retrieved 3-22-2013.

Kentucky Foundation. 2013. Kentucky Coal Resources. URL: http://www.coaleducation.org/Ky_coal_facts/coal_resources/ky_resources.htm. Retrieved 10-17-2013.

Korose, C.P., C.G. Treworgy, R.J.Jacobson, and S.D. Elrick. 2002. Availability of the Danville, Jamestown, Dekoven, Davis, and Seelyville Coals for Mining in selected areas of Illinois. Illinois Minerals 124: Department of Natural Resources, Illinois State Geological Survey, Champaign, IL. URL: http://www.isgs.illinois.edu/sites/isgs/files/files/coal-maps/im124.pdf.

Lawrence, R.L., S.D. Wood, and R.L. Sheley, 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest), Remote Sensing of Environment, 100: 356-362.

Lutz, B.D., E.S. Bernhardt, W.H. Schlesinger. 2013. The environmental price tag on a ton of mountaintop removal coal. PLOS ONE 8(9):1-5.

Maryland Department of the Environment. 2012. Maryland surface mine permit shapefiles. Provided by Maryland Department of the Environment, Mining Program, Abandoned Minelands Division, Frostburg, MD.

National Energy Technology Laboratory. 2012. Tracking new coal-fired power plants (data update 1/13/2012). National Energy Technology Laboratory, Office of Strategic Energy Analysis & Planning, U.S. Department of Energy, Morgantown, WV. URL: http://www.netl.doe.gov/coal/refshelf/ncp.pdf. Retrieved 6/12/13.

Office of Surface Mining Reclamation and Enforcement. 2013. Tennessee active surface mine permit data. Digital dataset provided by OSMRE Field Office, Knoxville, TN.

Ohio Department of Natural Resources 2013a. Ohio coal reserves by county. Ohio Division of Natural Resources, Columbus, OH. URL:

http://www.ohiodnr.com/OhioGeologicalSurvey/EnergyAndMineralResources/tabid/24204/Default.aspx . Retrieved 10-17-2013.

Ohio Department of Natural Resources. 2013b. Issued coal permits. Ohio Department of Natural Resources, Division of Mineral Resources Management, Columbus, OH. URL: http://www.ohiodnr.com/Portals/11/mining/pdf/issued.pdf, Retrieved 3/21/13.

PA Coal Alliance Inc. 2011. Coal data book 2011. Pennsylvania Coal Alliance, Inc., Harrisburg, PA. URL: http://www.pacoalalliance.com/wp-content/uploads/downloads/2012/07/coaldatabook2011.pdf. Retrieved 10-17-2013.

Palmer, M.A., E.S. Bernhardt, W.H. Schelsinger, K.N. Eschleman, E. Foufoula-Georgiou, M.S. Hendryx, A.D. Lemly, G.E. Likens, O.L. Loucks, M.E. Power, P.S. White, and P.R. Wilcock. 2010. Mountaintop mining consequences. Science 327:148-149.

Pennsylvania Department of Environmental Protection. 2012. Anthracite and bituminous mining permits. Pennsylvania Department of Environmental Protection, Bureau of Mining and Reclamation, Harrisburg, PA. URL: http://www.pasda.psu.edu/data/dep/ Retrieved 12/20/12.

Peters, J., B. De Baets, N.E.C. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts, 2007. Random forests as a tool for ecological distribution modelling, Ecological Modelling, 207: 304-318.

Pino-Mejías, R., M. Dolores Cubiles-de-la-Vega, M. Anaya-Romero, A. Pascual-Acosta, A. Jordán-López, and N. Bellinfante-Crocci, 2010. Predicting the potential habitat of oaks with data mining models and the R system, Environmental Modelling & Software, 25: 826-836.

Prasad, A.M., L.R. Iverson, and A. Liaw, 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, Ecosystems, 9: 181-199.

Ruppert, L.F., Trippi, M.H., and Slucher, E.R., 2010, Correlation chart of Pennsylvanian rocks in Alabama, Tennessee, Kentucky, Virginia, West Virginia, Ohio, Maryland, and Pennsylvania showing approximate position of coal beds, coal zones, and key stratigraphic units: U.S. Geological Survey Scientific Investigations Report 2010–5152, 9 p., 3 plates (online only). http://pubs.usgs.gov/sir/2010/5152/

Saylor, K.L. 2008. Land Cover Trends: Central Appalachians. U.S. Department of the Interior, U.S. Geological Survey (USGS), Washington, DC. URL: http://landcovertrends.usgs.gov/east/eco69Report.html. Accessed 12-13-2013.

Sierra Club. 2013. Stopping the coal rush. Sierra Club, San Francisco, CA. URL: http://www.sierraclub.org/environmentallaw/coal/plantlist.aspx. Retrieved 5/7/13.

SourceWatch. 2013. Existing U.S. coal plants. Center for Media and Democracy, Madison, WI. URL: http://www.sourcewatch.org/index.php?title=Existing_U.S._Coal_Plants. Retrieved 3/12/13.

Thompson, E.C., M.C. Berger, S.N. Allen, and J.M. Roenker. 2001. A study on the current economic impacts of the Appalachian coal industry and its future in the region. Center for Business and Economic Research, Gatton College of Business and Economics, University of Kentucky, Lexington, KY.

Treworgy, C.G., C.P. Korose, C.A. Chenoweth, and D.L. North. 1999. Availability of the Springfield Coal for Mining in Illinois. Illinois Minerals 118: Department of Natural Resources, Illinois State Geological Survey, Champaign, IL. URL: http://www.isgs.illinois.edu/sites/isgs/files/files/coal-maps/im118.pdf.

Treworgy, C.G., C.P. Korose, and C.L. Wiscombe. 2000. Availability of the Herrin Coal for mining in Illinois. Illinois Minerals 120: Department of Natural Resources, Illinois State Geological Survey, Champaign, IL. URL: http://www.isgs.illinois.edu/sites/isgs/files/files/coal-maps/im120.pdf.

U.S. Energy Information Administration. 2011. Existing units by energy source (existing_gen_units_2011.xls). United States Department of Energy, Energy Information Administration, Washington, DC. URL: http://www.eia.gov/electricity/capacity/. Retrieved 1/24/13.

U.S. Energy Information Administration. 2012a. Power plants (GIS shapefile). United States Department of Energy, Energy Information Administration, Washington, DC. URL: http://www.eia.gov/maps/map_data/ EIA_States_MapLayer_PowerPlants.zip. Retrieved 6/12/13.

U.S. Energy Information Administration. 2012b. Annual Coal Distribution Report 2011. United States Department of Energy, Energy Information Administration, Washington, DC. URL: http://www.eia.gov/coal/distribution/annual/pdf/acdr_fullreport2011.pdf. Retrieved 6/17/13.

U.S. Energy Information Administration. 2013a. Annual Energy Outlook 2013 with projections to 2040. United States Department of Energy, Energy Information Administration, Washington, DC. URL: http://www.eia.gov/forecasts/aeo/pdf/0383(2013).pdf

U.S. Energy Information Administration. 2013b. Annual coal report 2012. U.S. Energy Information Administration, Washington, DC. URL: http://www.eia.gov/coal/annual/pdf/acr.pdf. Accessed 12-13-2013.

U.S. Environmental Protection Agency. 2005. Mountaintop mining/valley fills in Appalachia: Final Programmatic Environmental Impact Statement. EPA Region 3, Philadelphia, PA. EPA 9-03-R-05002.

U.S. Geological Survey. 2000. Resource assessment of selected coal beds and zones in the Northern and Central Appalachian Basin coal regions. Northern and Central Appalachian basin coal regions assessment team, U.S. Geological Survey, U.S. Department of the Interior, Washington, DC. URL: http://pubs.usgs.gov/pp/p1625c/. Retrieved 10-29-2013.

U.S. Geological Survey. 2008. Coal fields of the United States. National Atlas of the United States. U.S. Geological Survey, Eastern Energy Team, John Tully (Comp.), Reston, VA. URL: http://nationalatlas.gov/atlasftp.html#coalfdp. Retrieved 6/18/13.

U.S. Geological Survey. 2011. A summary of the relationship between GAP status codes and IUCN definitions. U.S. Geological Survey, National Gap Analysis Program, Boise, ID. URL: U.S. Geological Survey. URL: http://gapanalysis.usgs.gov/blog/iucn-definitions/. Retrieved 6/18/13.

U.S. Geological Survey. 2013. Geologic maps of U.S. states. URL: http://mrdata.usgs.gov/geology/state/. Retrieved 10-26-2013.

Virginia Department of Mines Minerals and Energy. 2013. DMLR Mapping and resource page. Virginia Department of Mines Minerals and Energy, Division of Mined Land Reclamation, Richmond, VA .URL: http://www.dmme.virginia.gov/DMLR/DmlrMappingPage.shtml, Retrieved 3/21/13.

Virginia Tech. 1999. Virginia Coal Reserves. Virginia Tech Department of Mining and Minerals Engineering. URL: http://www.energy.vt.edu/reserves/. Retrieved 10/29/2013.

West Virginia Coal Association. 2012. West Virginia Coal Facts 2012. West Virginia Coal Association, Charleston, WV. URL: http://www.wvcoal.com/coal-facts-2012.html. Retrieved 6/11/13.

WV Department of Environmental Protection 2013. Mine permit boundaries (GIS dataset). West Virginia Department of Environmental Protection, Technical Applications and GIS Unit, Charleston, WV. URL: http://tagis.dep.wv.gov/data2.html. Retrieved 3/25/13.

West Virginia Geological and Economic Survey. 2013. Coal Bed Mapping Project. WV Geological and Economic Survey, Morgantown, WV. URL: http://www.wvgs.wvnet.edu/www/coal/cbmp/coalims.html. Retrieved 10/29/2013.

Appendices

Random Forests model notes

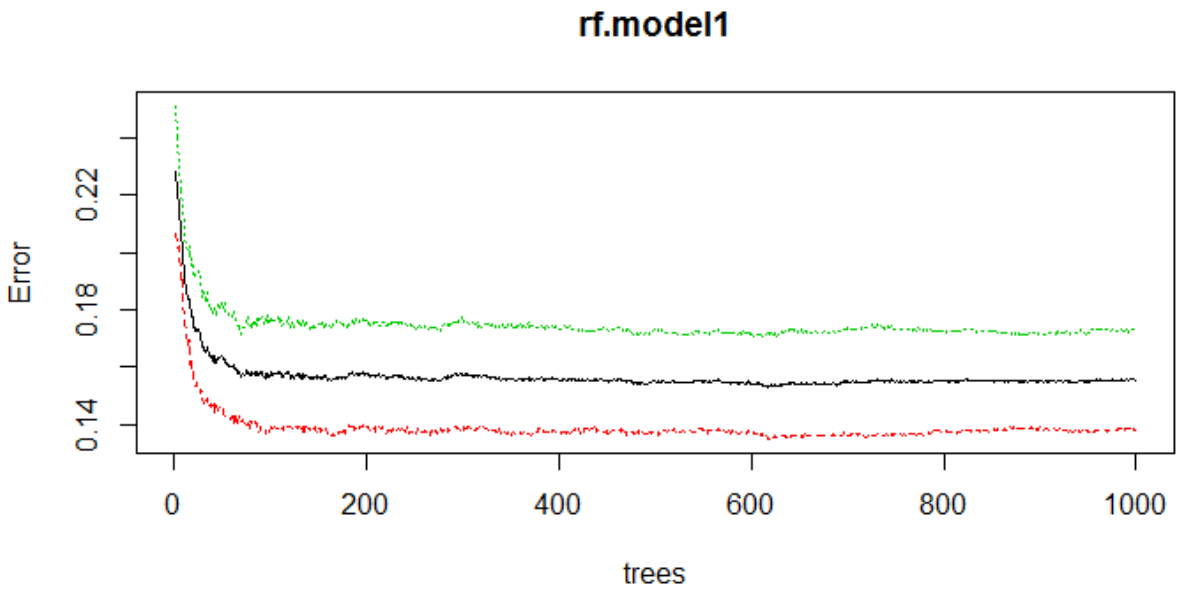
Revised random Forests model, October 2013

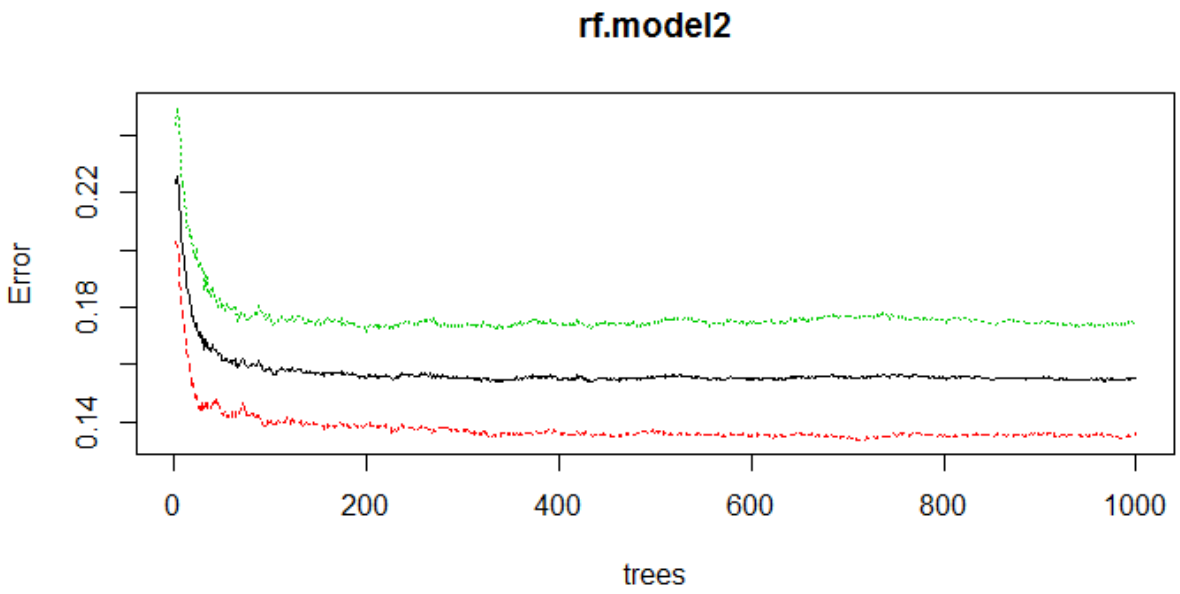
```
> set1 <- read.csv("set_1.csv")
> set2 <- read.csv("set_2.csv")
> set3 <- read.csv("set_3.csv")
> set4 <- read.csv("set_4.csv")
> set5 <- read.csv("set_5.csv")
>
set1$FID_ <- NULL
set2$FID_ <- NULL
set3$FID_ <- NULL
set4$FID_ <- NULL
set5$FID_ <- NULL
>
> raster <- stack(“ash”, “btu”, “distr”, “geotype”, “mod”, “popd”, “port”,
“pow”, “sulf”, "mtr", "eiareg")
Error: unexpected input in "raster <- stack(“"
> raster <- stack("ash", "btu", "distr", "geotype", "mod", "popd", "port",
"pow", "sulf", "mtr", "eiareg")
raster$geotype <- as.factor(raster$geotype)
set1$geotype <- as.factor(set1$geotype)
set2$geotype <- as.factor(set2$geotype)
set3$geotype <- as.factor(set3$geotype)
set4$geotype <- as.factor(set4$geotype)
set5$geotype <- as.factor(set5$geotype)

raster$mtr <- as.factor(raster$mtr)
set1$mtr <- as.factor(set1$mtr)
set2$mtr <- as.factor(set2$mtr)
set3$mtr <- as.factor(set3$mtr)
set4$mtr <- as.factor(set4$mtr)
set5$mtr <- as.factor(set5$mtr)

raster$eiareg <- as.factor(raster$eiareg)
set1$eiareg <- as.factor(set1$eiareg)
set2$eiareg <- as.factor(set2$eiareg)
set3$eiareg <- as.factor(set3$eiareg)
set4$eiareg <- as.factor(set4$eiareg)
set5$eiareg <- as.factor(set5$eiareg)

rf.model1 <- randomForest(formula = class ~ sulf + ash + btu + port + mod +
distr + popd + geotype + pow + mtr + eiareg, data= set1, importance = T,
confusion=T, err.rate=T, ntree=1000)

rf.model2 <- randomForest(formula = class ~ sulf + ash + btu + port + mod +
distr + popd + geotype + pow + mtr + eiareg, data= set2, importance = T,
confusion=T, err.rate=T, ntree=1000)

rf.model3 <- randomForest(formula = class ~ sulf + ash + btu + port + mod +
distr + popd + geotype + pow + mtr + eiareg, data= set3, importance = T,
confusion=T, err.rate=T, ntree=1000)

rf.model4 <- randomForest(formula = class ~ sulf + ash + btu + port + mod +
distr + popd + geotype + pow + mtr + eiareg, data= set4, importance = T,
confusion=T, err.rate=T, ntree=1000)

rf.model5 <- randomForest(formula = class ~ sulf + ash + btu + port + mod +
distr + popd + geotype + pow + mtr + eiareg, data= set5, importance = T,
confusion=T, err.rate=T, ntree=1000)
```

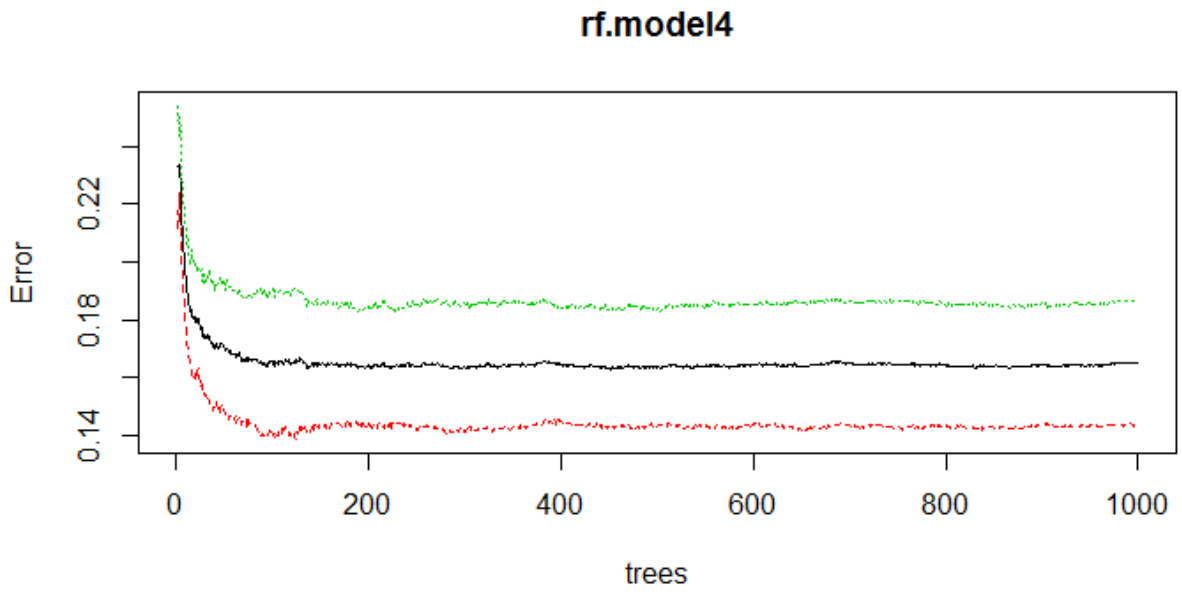
```
total.model <- combine(rf.model1, rf.model2, rf.model3, rf.model4, rf.model5)


predict(raster, total.model, type="prob", index=1, na.rm=TRUE,
progress="window", overwrite=TRUE, filename="model2new.img")

Loading required package: tcltk
class       : RasterLayer
dimensions  : 1303, 1303, 1697809  (nrow, ncol, ncell)
resolution  : 1000, 1000  (x, y)
extent      : 589837.8, 1892838, 1094030, 2397030  (xmin, xmax, ymin, ymax)
coord. ref. : +proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=23 +lon_0=-96 +x_0=0
+y_0=0 +ellps=GRS80 +units=m +no_defs
data source : C:\R\ModelData\DataToRun\model2new.img
names       : model2new
values      : 0, 1  (min, max)

> rf.model1$importance
              mine          NOT MeanDecreaseAccuracy MeanDecreaseGini
sulf    0.18024028  0.013024244           0.09648113         594.4364
ash     0.06591107  0.008392967           0.03708831         225.7069
btu     0.12771574  0.021233325           0.07435588         489.8110
port    0.14093593  0.012632069           0.07664152         527.9491
mod     0.13035583  0.016402501           0.07325908         500.0359
distr   0.10268957  0.011684316           0.05708882         462.0057
popd    0.07792855  0.014670125           0.04623679         397.6203
geotype 0.23150452  0.006925302           0.11899639         733.4620
pow     0.11236325  0.016442982           0.06430293         521.0641
mtr     0.15376096 -0.014349519           0.06954599         192.4539
eiareg  0.19811271 -0.012425979           0.09259842         476.8881
> rf.model2$importance
              mine          NOT MeanDecreaseAccuracy MeanDecreaseGini
sulf    0.17858038  0.016928788           0.09759745         608.9732
ash     0.05694270  0.007952359           0.03239669         220.9961
btu     0.12259524  0.019760048           0.07107897         488.6975
port    0.13666698  0.009497751           0.07293873         513.3651
mod     0.13120586  0.016999131           0.07398846         503.7707
distr   0.09983573  0.008764772           0.05420070         437.3988
popd    0.07016104  0.012722700           0.04137982         386.7707
geotype 0.22254097  0.016203508           0.11914574         769.9558
pow     0.10916662  0.014356112           0.06165392         509.3383
mtr     0.12826610 -0.008086872           0.05992834         189.1854
eiareg  0.19567671 -0.010277671           0.09248223         494.6217
> rf.model3$importance
              mine          NOT MeanDecreaseAccuracy MeanDecreaseGini
sulf    0.18140136  0.016951630           0.09900894         623.7124
ash     0.05830702  0.008105648           0.03315278         224.5400
btu     0.13019158  0.019050958           0.07451005         501.9171
port    0.13926593  0.012503211           0.07574482         528.3314
mod     0.13436748  0.020026453           0.07706217         528.5690
distr   0.10946632  0.013407648           0.06132853         460.0092
popd    0.07688747  0.012092635           0.04443099         387.3635
geotype 0.22692511  0.013194300           0.11982490         744.5327
pow     0.10918272  0.016847003           0.06291390         521.1553
mtr     0.13012945 -0.009113428           0.06034056         152.6249
eiareg  0.17833712 -0.010314637           0.08375349         448.7758
> rf.model4$importance
              mine          NOT MeanDecreaseAccuracy MeanDecreaseGini
sulf    0.17187952  0.015930731           0.09380560         606.8031
ash     0.05930049  0.006997447           0.03310909         222.9184
btu     0.13486632  0.023122427           0.07891794         521.3684
port    0.13987022  0.010337099           0.07501123         534.1551
mod     0.13760857  0.017402620           0.07742640         527.6842
```

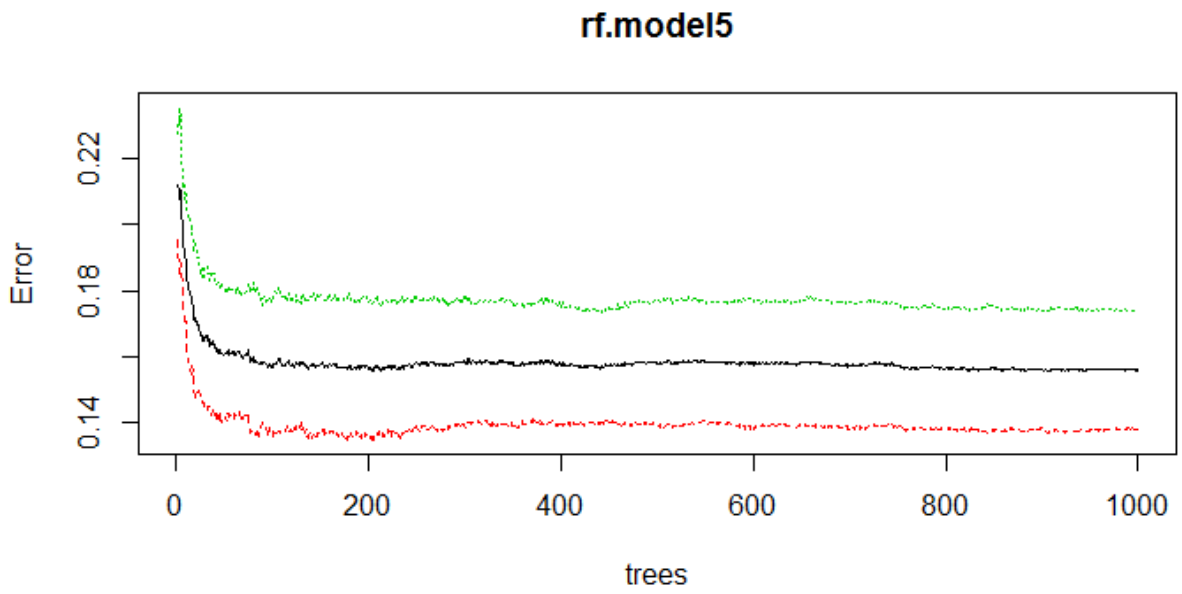
```
distr   0.10220166  0.013367159           0.05771412         438.8221
popd    0.07871415  0.012965211           0.04579505         401.7309
geotype 0.21984563  0.015052973           0.11726537         744.2954
pow     0.10859752  0.015880968           0.06218144         526.6931
mtr     0.13626252 -0.012226856           0.06186401         166.4477
eiareg  0.17222926 -0.010683506           0.08064315         428.5456
> rf.model5$importance
              mine          NOT MeanDecreaseAccuracy MeanDecreaseGini
sulf    0.16978617  0.014433324           0.09193848         576.2544
ash     0.06054685  0.008712426           0.03457209         219.2102
btu     0.12564723  0.023074382           0.07426369         484.1721
port    0.13859898  0.013059513           0.07569448         520.5819
mod     0.13985339  0.016662958           0.07813663         513.5633
distr   0.11332083  0.013372590           0.06324147         474.1342
popd    0.07591737  0.013929695           0.04485276         401.8855
geotype 0.22538338  0.018720062           0.12182665         759.6918
pow     0.10959507  0.016888852           0.06313905         517.0596
mtr     0.13199837 -0.009120261           0.06125895         165.0891
eiareg  0.18733137 -0.007122312           0.08984372         488.0499
>
> rf.model1

Call:
 randomForest(formula = class ~ sulf + ash + btu + port + mod +      distr +
popd + geotype + pow + mtr + eiareg, data = set1,      importance = T,
confusion = T, err.rate = T, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.54%
Confusion matrix:
     mine  NOT class.error
mine 4436  711   0.1381387
NOT   891 4274   0.1725073
> rf.model2

Call:
 randomForest(formula = class ~ sulf + ash + btu + port + mod +      distr +
popd + geotype + pow + mtr + eiareg, data = set2,      importance = T,
confusion = T, err.rate = T, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.54%
Confusion matrix:
     mine  NOT class.error
mine 4449  698   0.1356130
NOT   904 4261   0.1750242
> rf.model3

Call:
 randomForest(formula = class ~ sulf + ash + btu + port + mod +      distr +
popd + geotype + pow + mtr + eiareg, data = set3,      importance = T,
confusion = T, err.rate = T, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.61%
Confusion matrix:
     mine  NOT class.error
mine 4436  711   0.1381387
```

```
NOT   899 4266   0.1740561
> rf.model4

Call:
 randomForest(formula = class ~ sulf + ash + btu + port + mod +      distr +
popd + geotype + pow + mtr + eiareg, data = set4,      importance = T,
confusion = T, err.rate = T, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 16.51%
Confusion matrix:
     mine  NOT class.error
mine 4409  738   0.1433845
NOT   964 4201   0.1866409
>
> rf.model5

Call:
 randomForest(formula = class ~ sulf + ash + btu + port + mod +      distr +
popd + geotype + pow + mtr + eiareg, data = set5,      importance = T,
confusion = T, err.rate = T, ntree = 1000)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.58%
Confusion matrix:
     mine  NOT class.error
mine 4437  710   0.1379444
NOT   897 4268   0.1736689
> plot(rf.model1)
> plot(rf.model2)
> plot(rf.model3)
> plot(rf.model4)
> plot(rf.model5)
>
```


rf.model1
Error
0.14
0.18
0.22
0
200
400
600
800
1000
trees


rf.model2
Error
0.14
0.18
0.22
0
200
400
600
800
1000
trees


rf.model4
Error
0.14
0.18
0.22
0
200
400
600
800
1000
trees


rf.model5
Error
0.14
0.18
0.22
0
200
400
600
800
1000
trees