Image from NASA:
https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

GIScience

Spatial Predictive Modeling

WV VIEW

AmericaView℠
Empowering Earth Observation Education
AmericaView.org - Est. 2003

Modifiable Areal Unit Problem (MAUP) and the Ecological Fallacy

What is the modifiable areal unit problem (MAUP)?

- A challenge that occurs during the spatial analysis of aggregated data in which the results differ when the same analysis is applied to the same data, but different aggregation schemes are used.
- MAUP takes two forms: the scale effect and the zone effect

Modifiable Areal Unit Problem (MAUP)

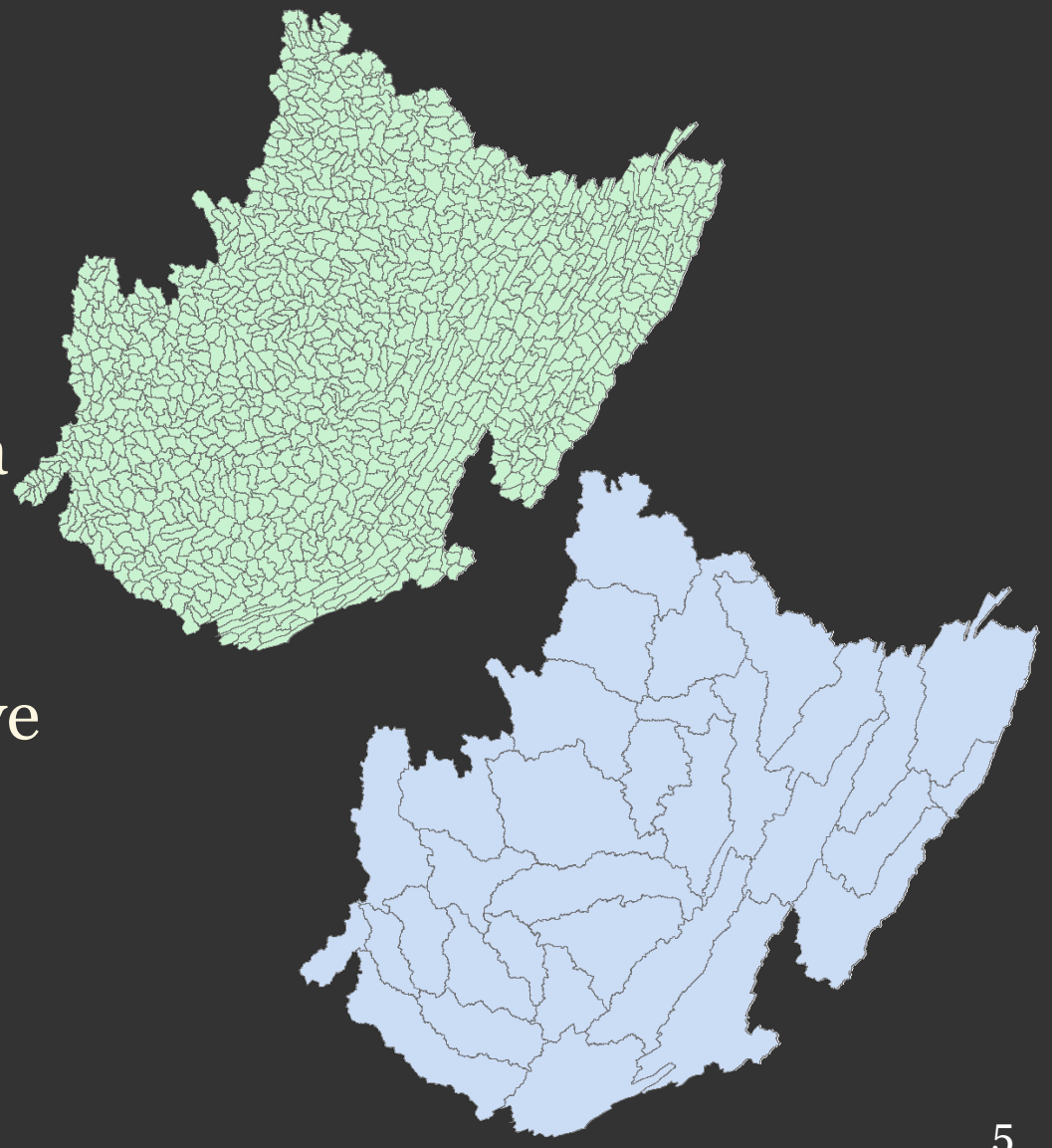
Openshaw (1984)

“The areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating.“

Example: Gerrymandering

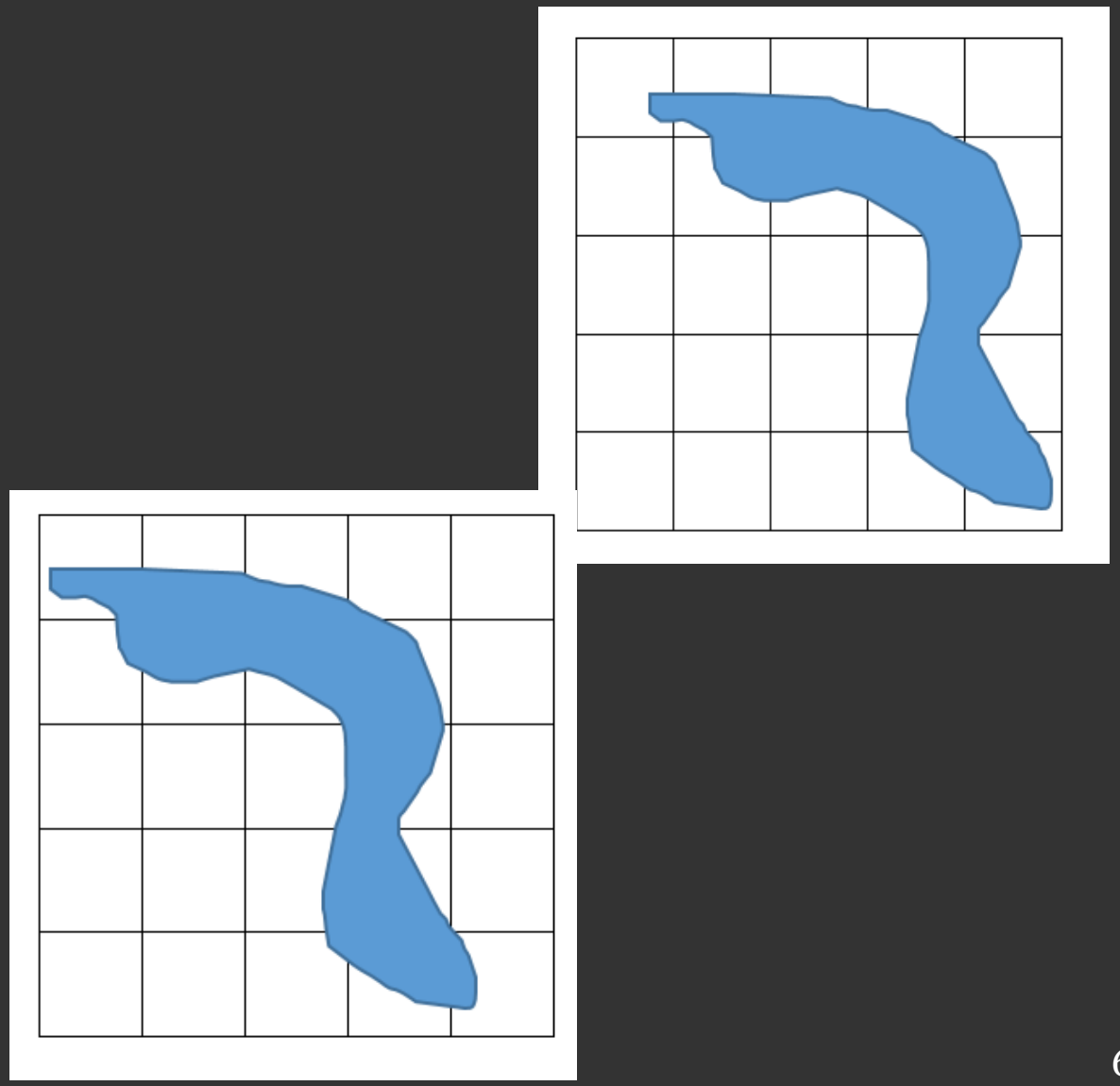
Scale Effect

- The scale effect suggest that analyses will exhibit different results when the same analysis is applied to the same data, but the scale of aggregation is changed.
- For example, analyses using data aggregated by county will differ from analyses using data aggregated by census tract. Often this differences in results are valid: each analysis asks a different question because each evaluates the data from a different perspective (different scale).
- Take home: It is important to consider the scale of your analysis and your aggregating units because you are addressing your question at that specific aggregation scale



Zone Effect

- The zone effect is observed when the scale of analysis is fixed, but the shape or position of the aggregation units change.
- For example, an analyses using data aggregated into one-mile grid cells will differ from an analysis using one-mile hexagon cells.
- The zone effect is a problem because it is an analysis, at least in part, of the aggregation scheme rather than the data itself.



Spatial Heterogeneity/Spatial Autocorrelation

- Patterns can vary across space
- Observations or correlations noted at one location may not be true at other locations
- Samples may be correlated based on their proximity to each other in space

Ecological Fallacy

- The assumption that an individual from a specific group or area will exhibit a trait that is predominant in the group as a whole

Temporal Heterogeneity/Temporal Autocorrelation

- Observed patterns and correlations may change over time
- Observations near in time tend to be more similar than observations that are further apart in time

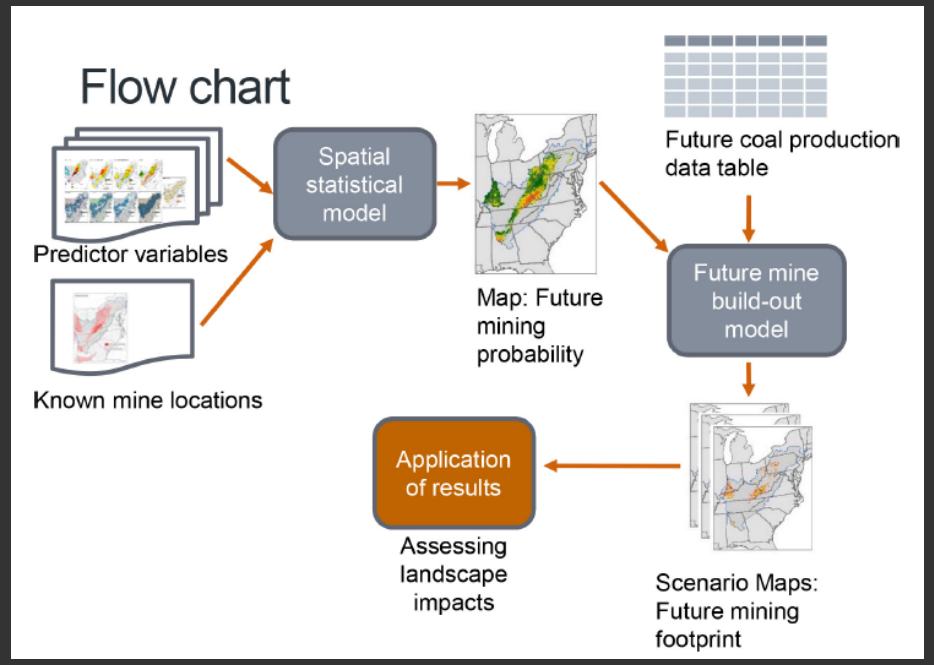
Practical Considerations

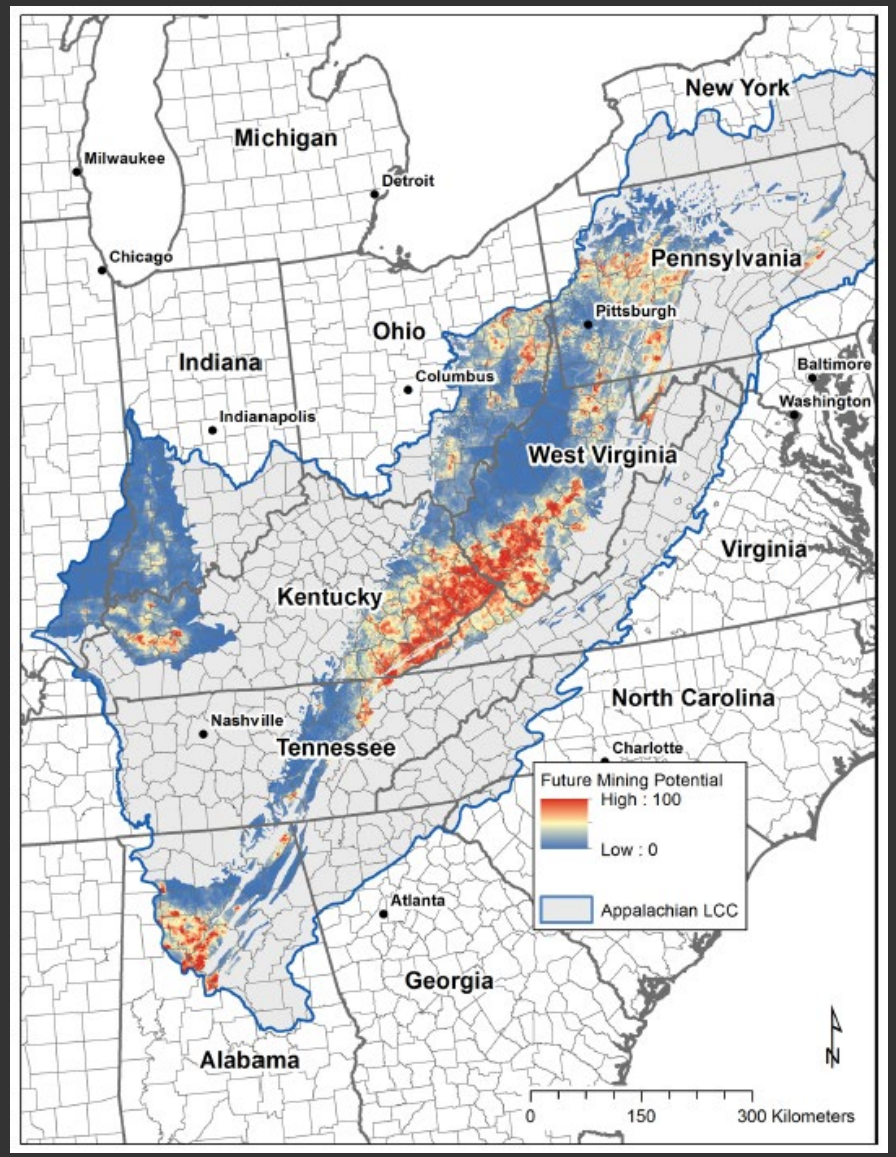
1. Scale matters
2. Can't assume patterns are consistent over space
3. Do not extrapolate from the population to the individual
4. Do not assume that measures, patterns, or results will be consistent over time
5. Do not assume that samples collected over space are independent

Overview of Modeling

What is a model?

A simplified representation of a phenomenon or a system





Strager, M.P., J.M. Strager, J.S. Evans, J.K. Dunscomb, B.J. Kreps, and A.E. Maxwell, 2015. Combining a spatial model and demand forecasts to map future surface coal mining in Appalachia, *PLoS ONE*.

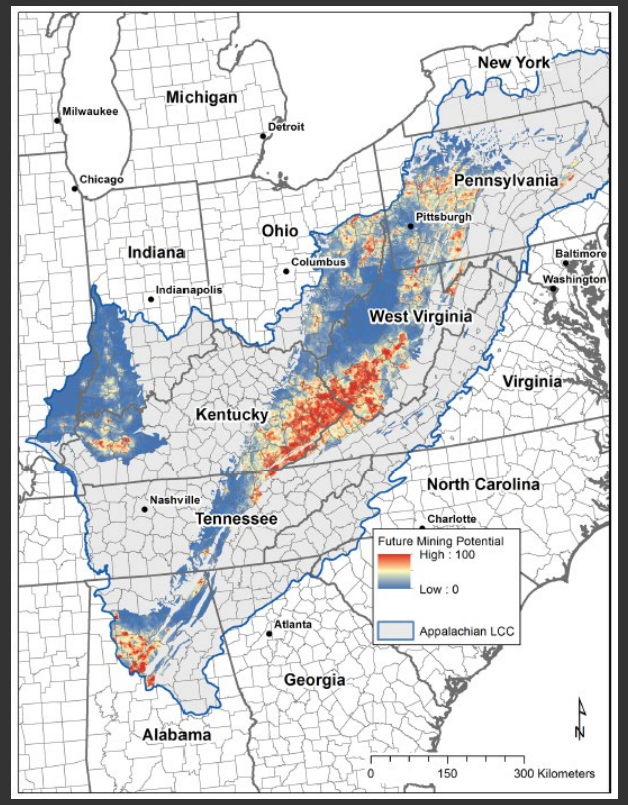
What is a model?

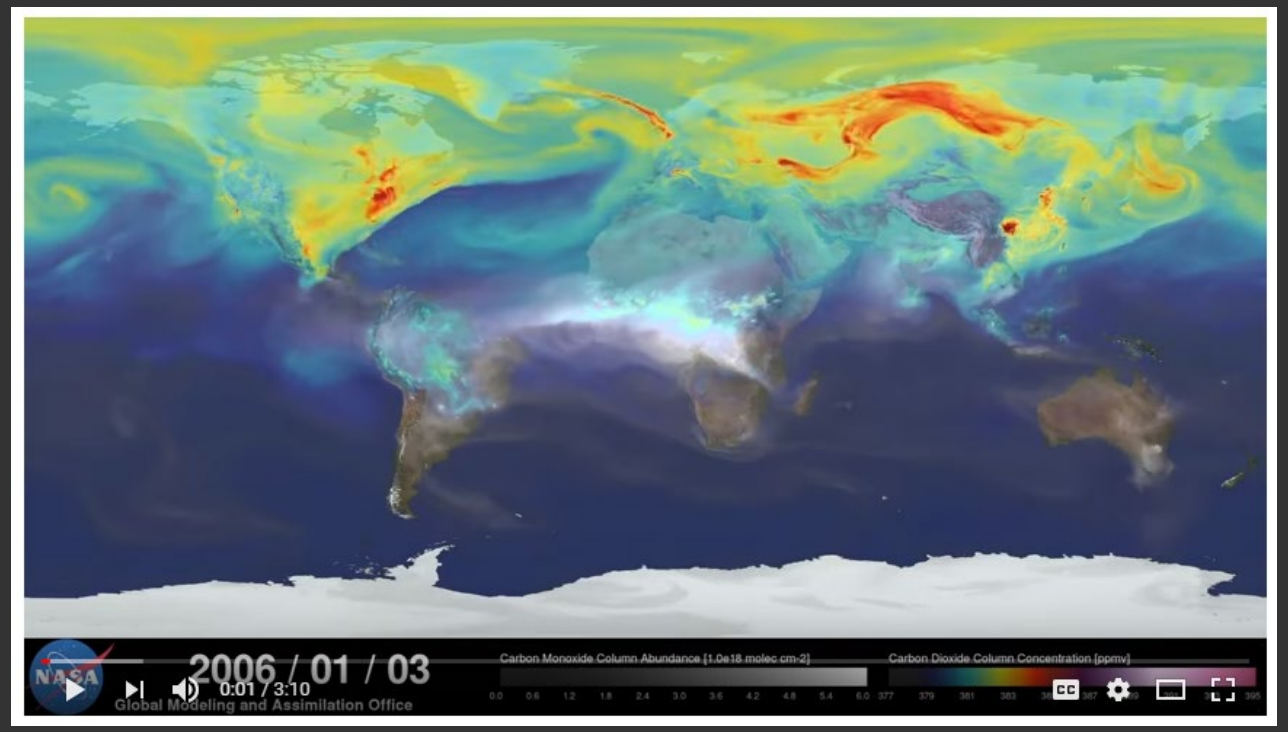
"Essentially, all models are wrong, but some are useful."

-Statistician George Box

Static vs. Dynamic Models

- Static = does not change with time
- Dynamic = changes with time

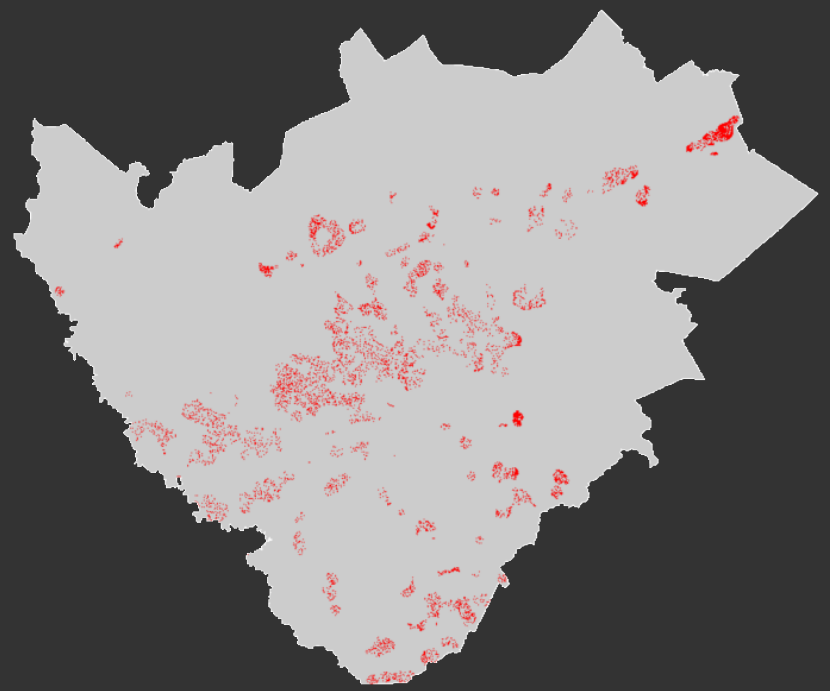


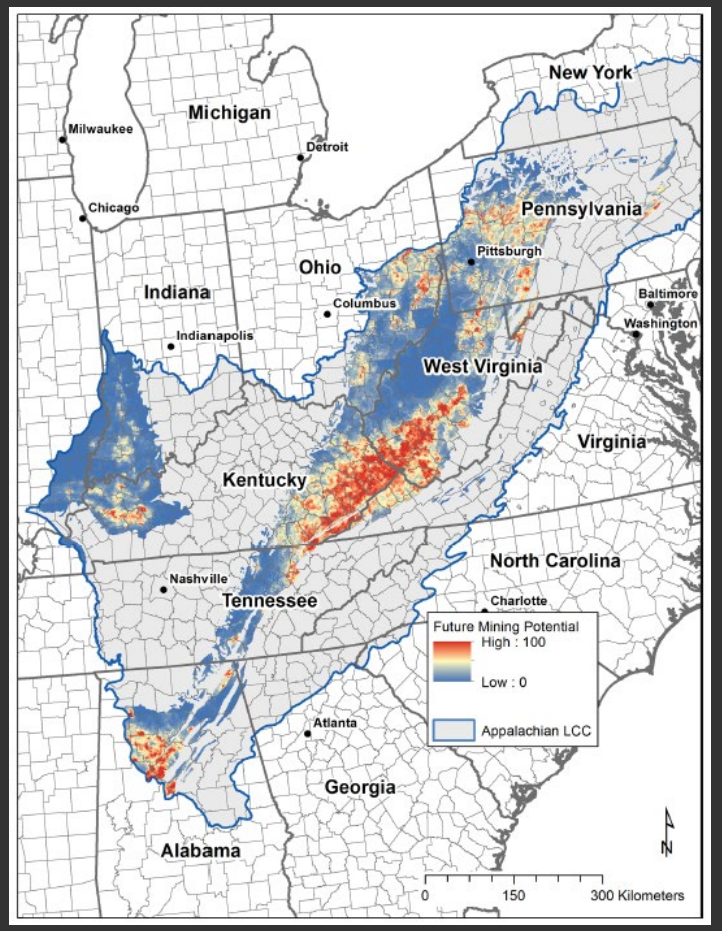


https://www.youtube.com/watch?v=x1SgmFa0r04

Conservative vs. Liberal Models

- Conservative = binary, 1/0, yes/no
- Liberal = represents gray levels





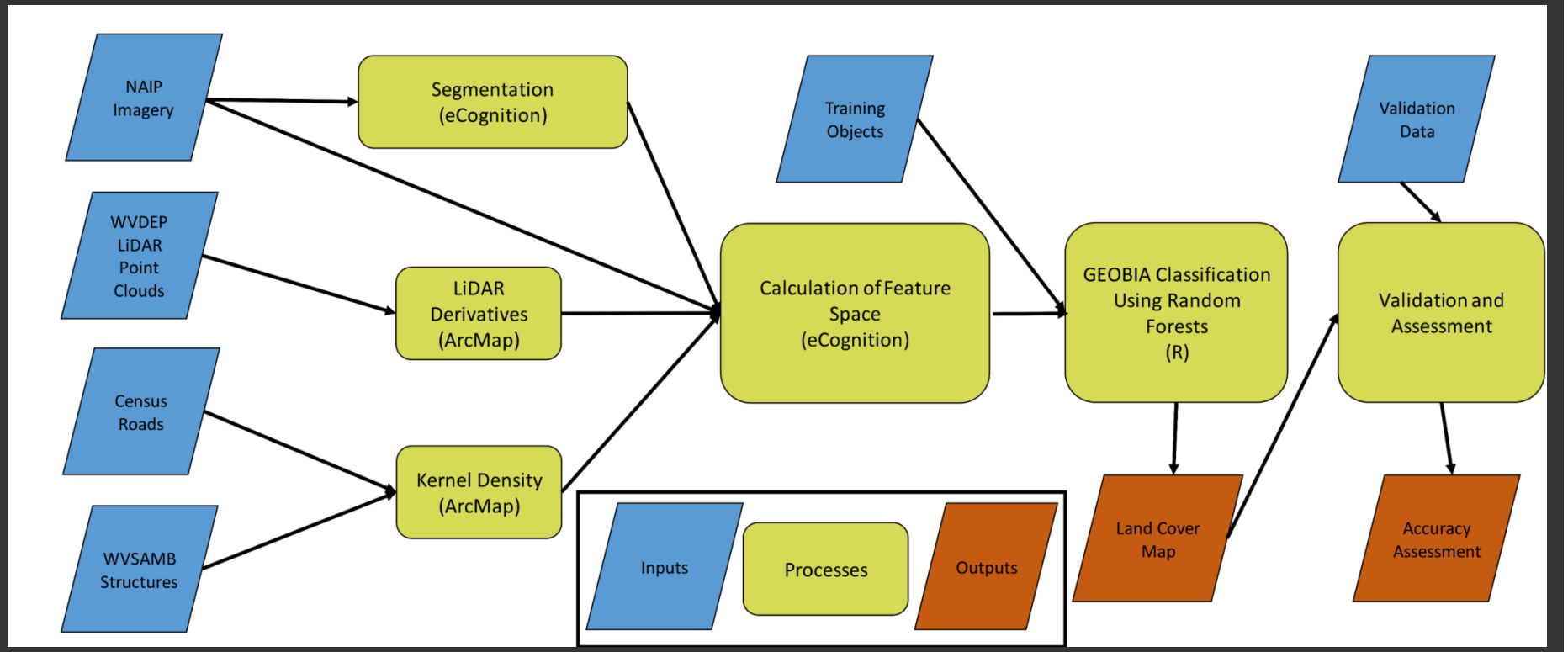
Deterministic vs. Stochastic

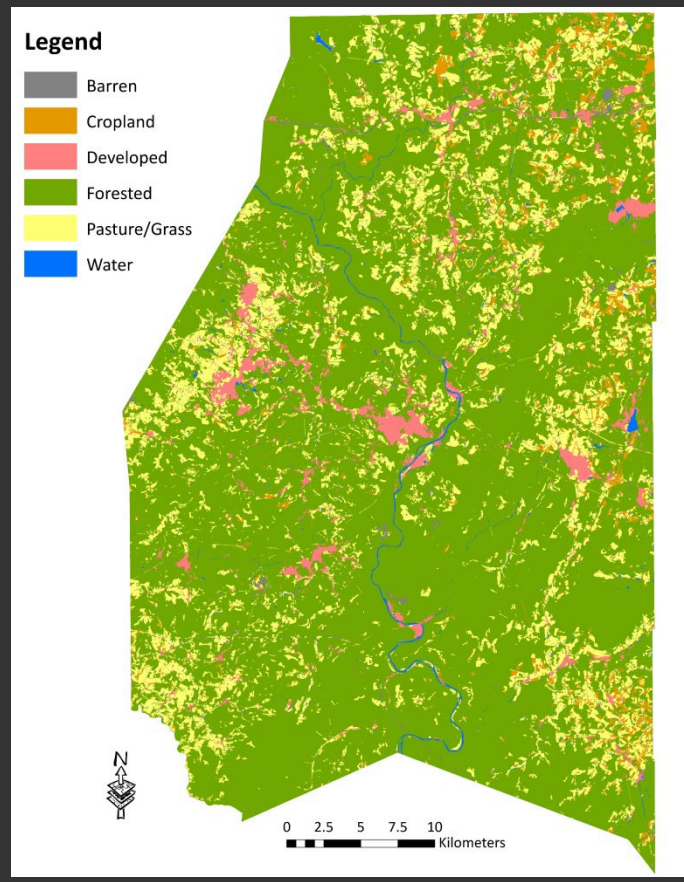
- Deterministic = does not incorporate randomness
- Stochastic = does incorporate randomness

Deductive vs. Inductive

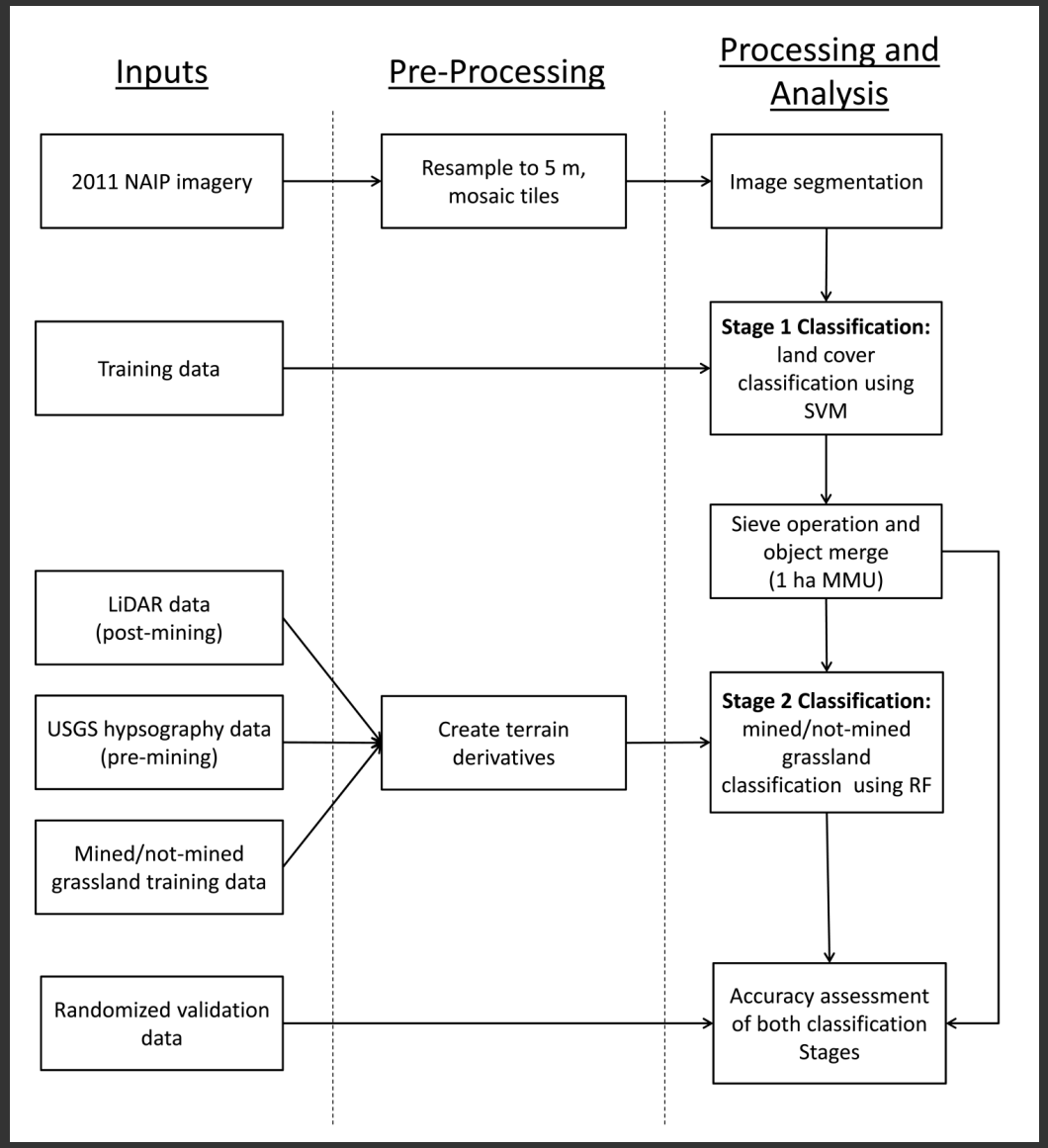
- Deductive = conclusion derived from a set of premises (general → specific)
- Inductive = conclusion derived from empirical data and observations (specific → general)

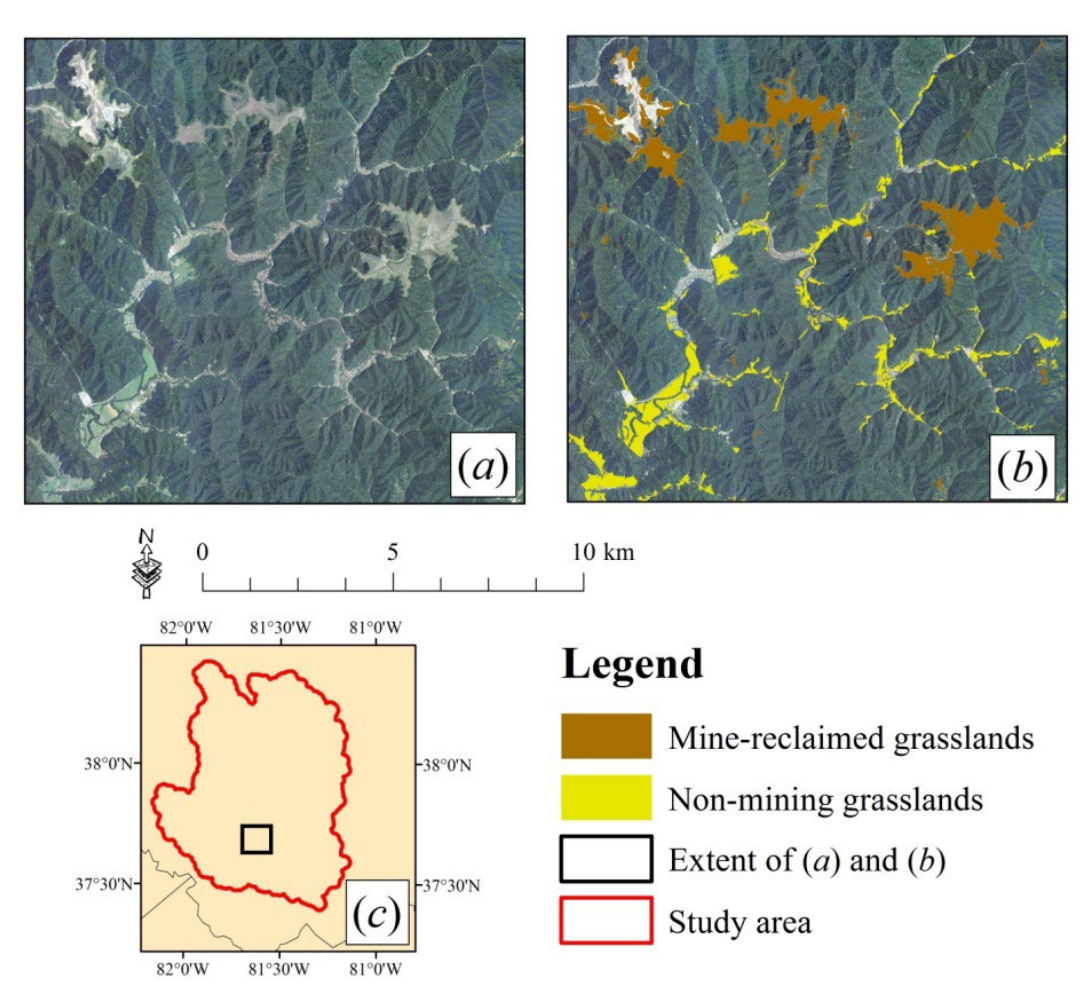
Modeling Example: Land Cover Classification





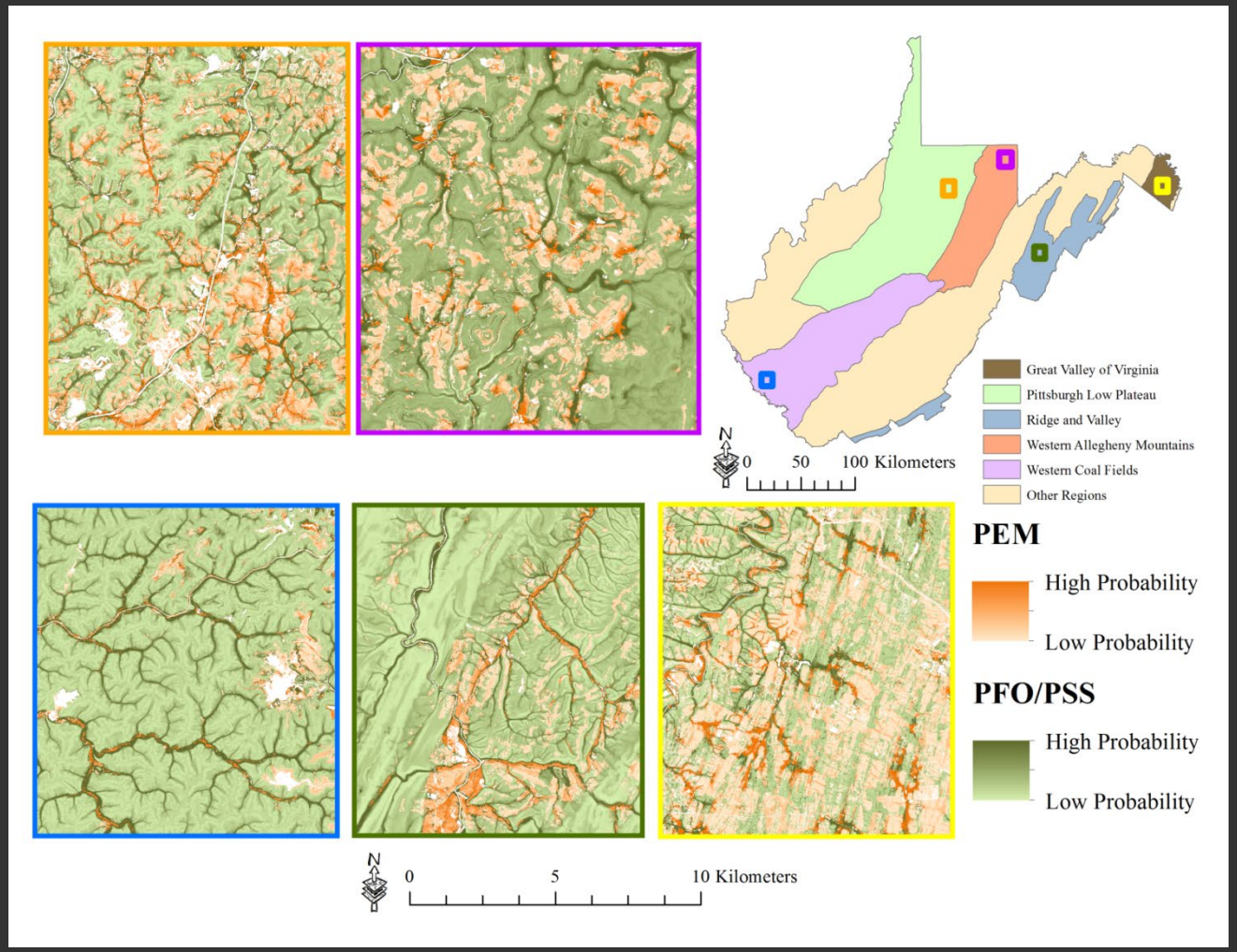
Modeling Example: Land Cover Classification





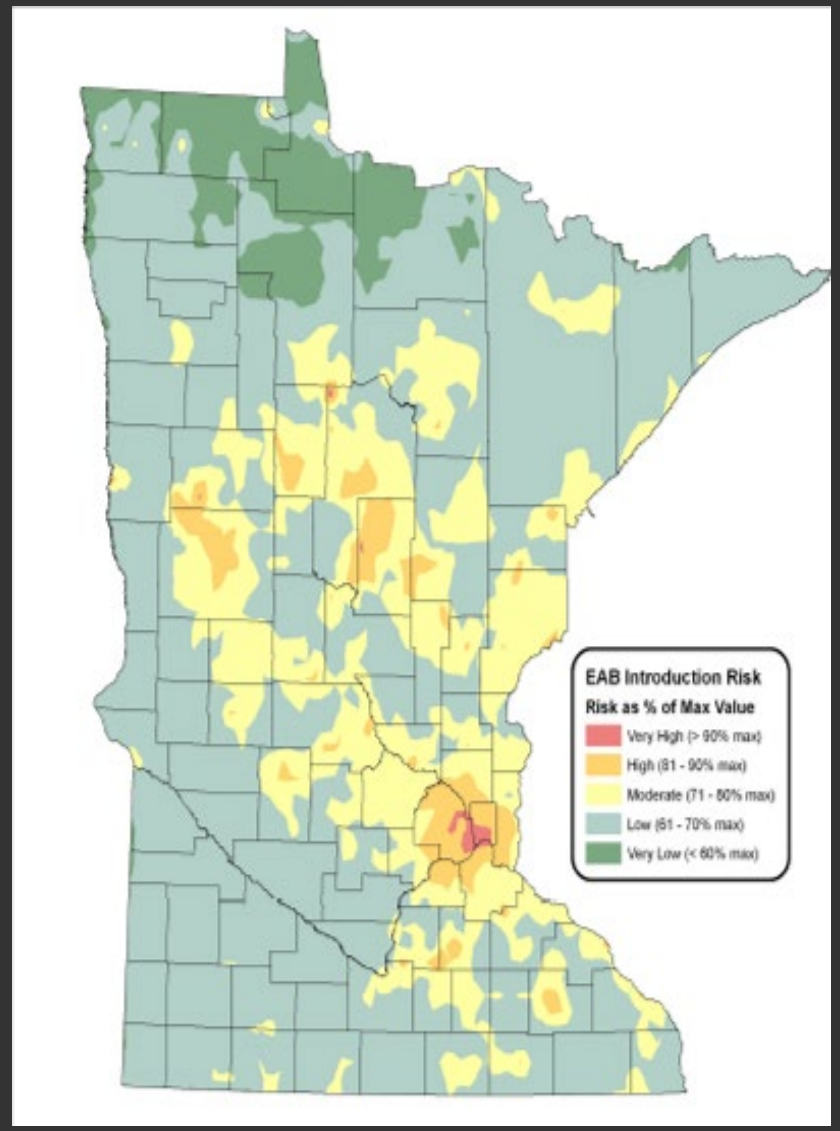
Maxwell, A.E., and T.A. Warner, 2015. Differentiating mine-reclaimed grasslands from spectrally similar land cover using terrain variables and object-based machine learning classification, *International Journal of Remote Sensing*.

Modeling Example: Wetland Probability



Maxwell, A.E., T.A. Warner, and M.P. Strager, 2016. Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables, *Photogrammetric Engineering & Remote Sensing*.

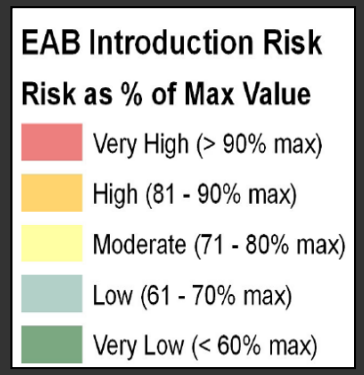
Model Example: Pest Risk Mapping (Emerald Ash Borer)



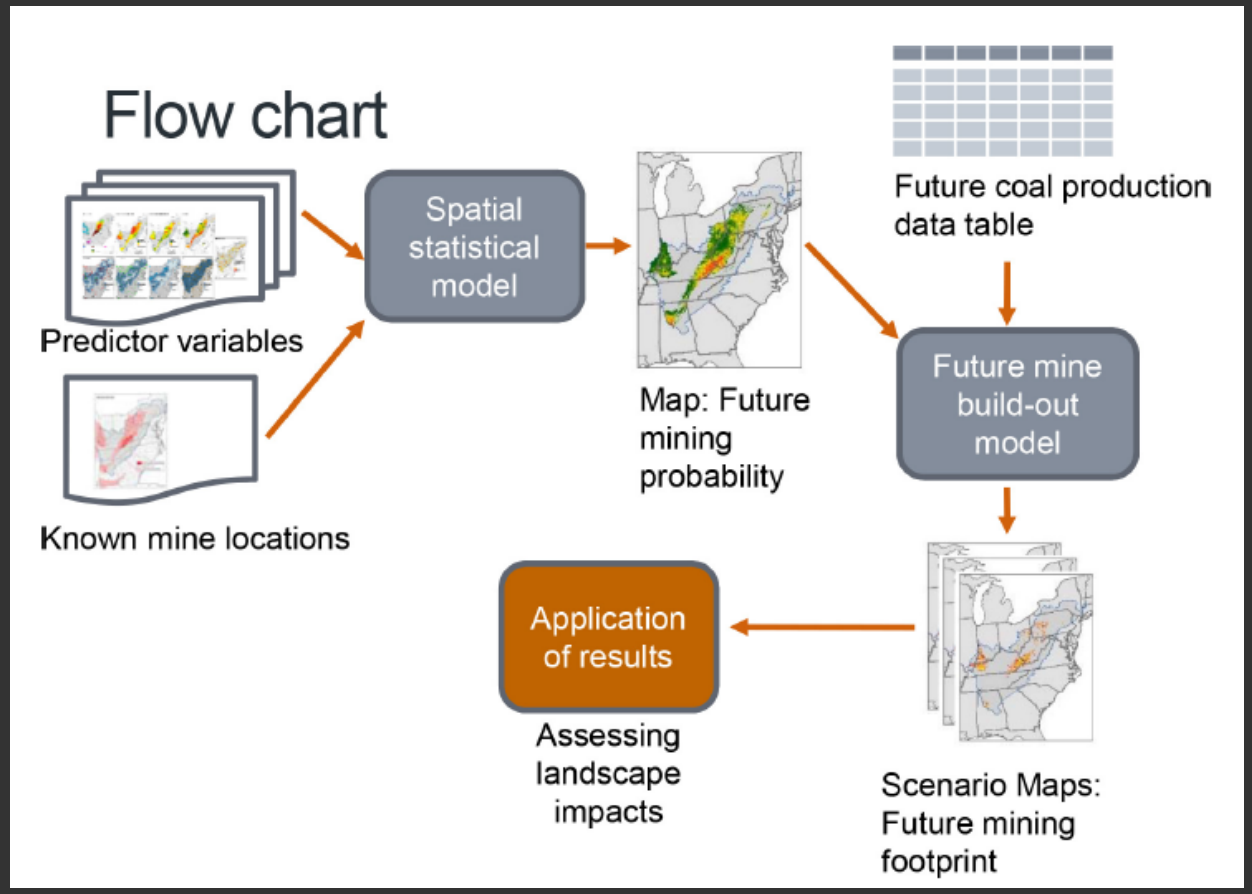
Equation 4 = Same as Equation 2 but Campsites multipled by 25, Seasonal homes multiplied by 25 and Accessibility multiplied by 1 (Figure 13):

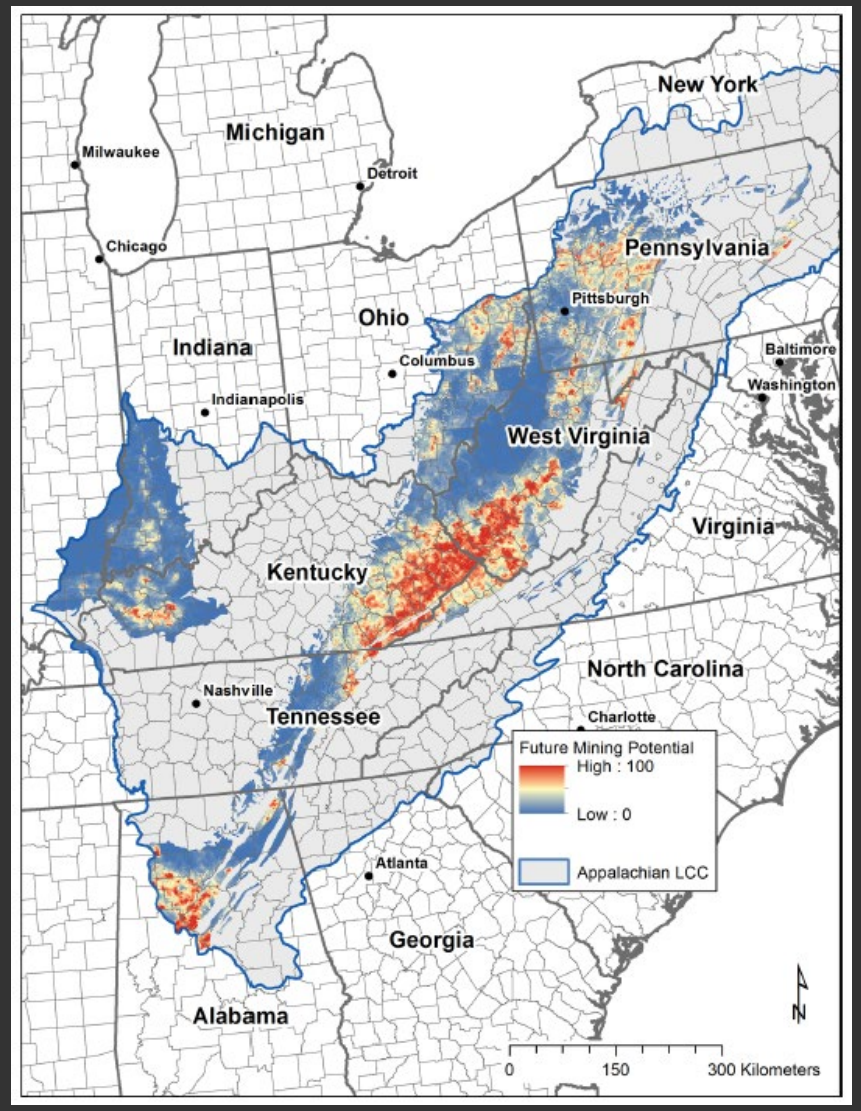
Log((25 * [Reclassified Campsite Density]) +
(25 * [Reclassified Cabin Density]) +
(10 * [Reclassified Urban Centroid Density]) +
(5 * [Reclassified Sawmill Risk Density]) +
(5 * [Reclassified Firewood Dealer Density]) +
(5 * [Reclassified High Risk Nursery Density]) +
(1 * [Reclassified Accessibility]))
= Risk Map Grid Surface

https://www.mda.state.mn.us/



Model Example: Mine Expansion

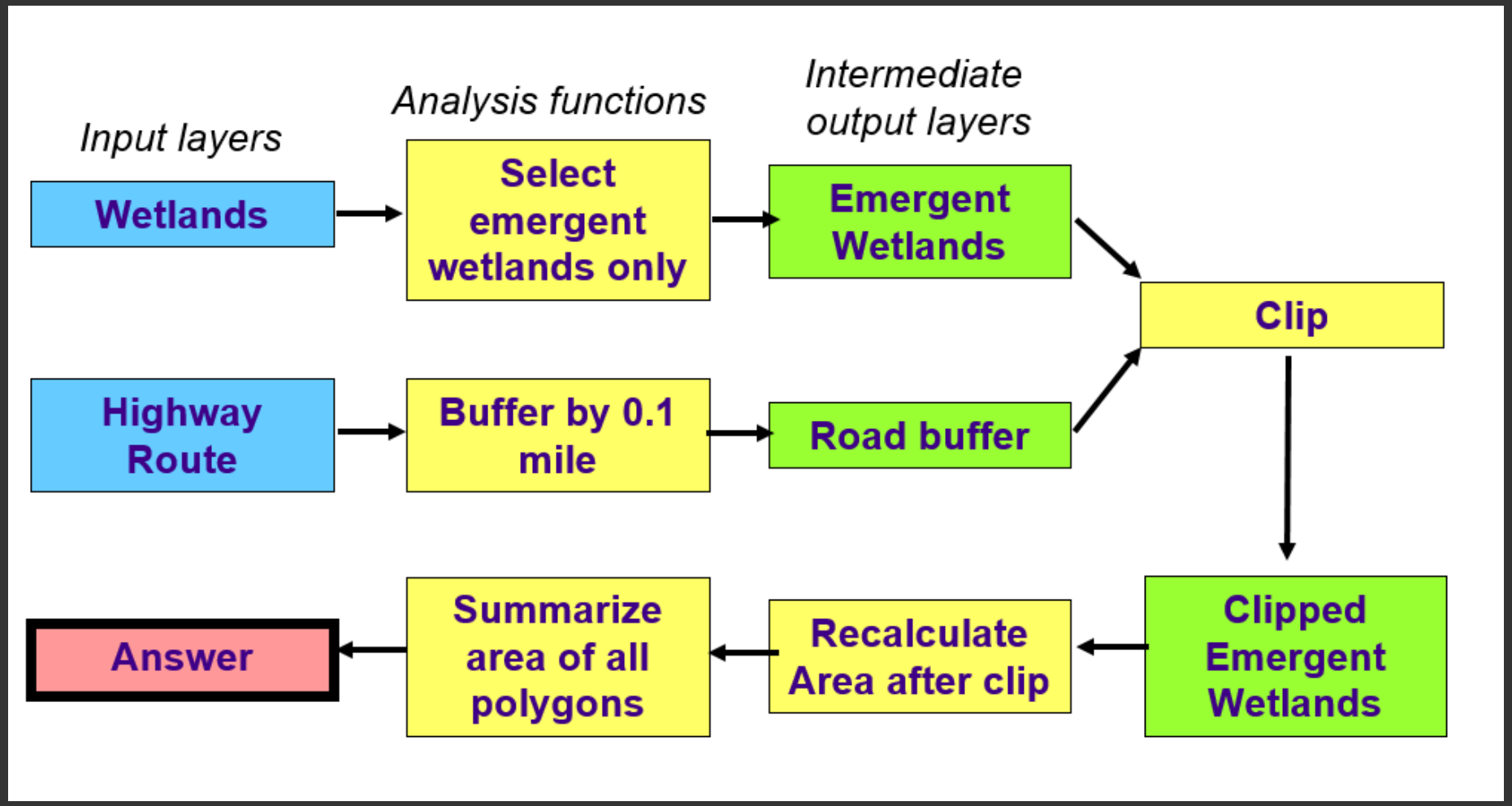




Strager, M.P., J.M. Strager, J.S. Evans, J.K. Dunscomb, B.J. Kreps, and A.E. Maxwell, 2015. Combining a spatial model and demand forecasts to map future surface coal mining in Appalachia, *PLoS ONE*.

Conservative Models

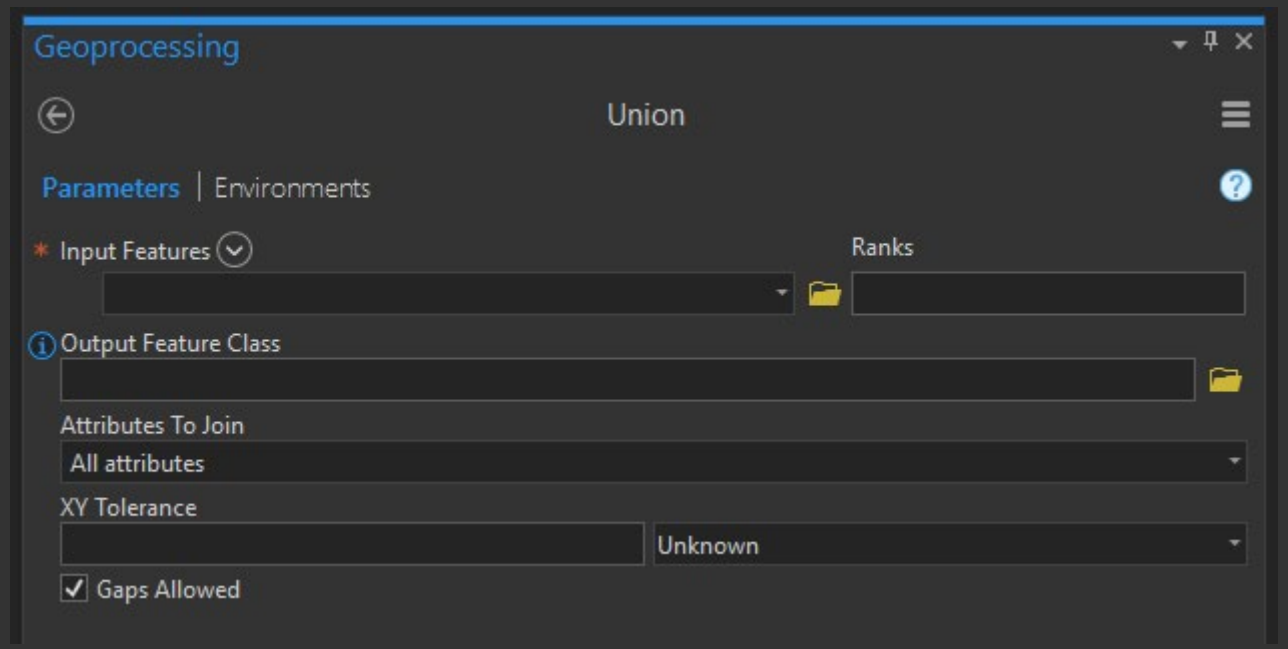
Vector Overlay



Example from Mike and Jackie Strager

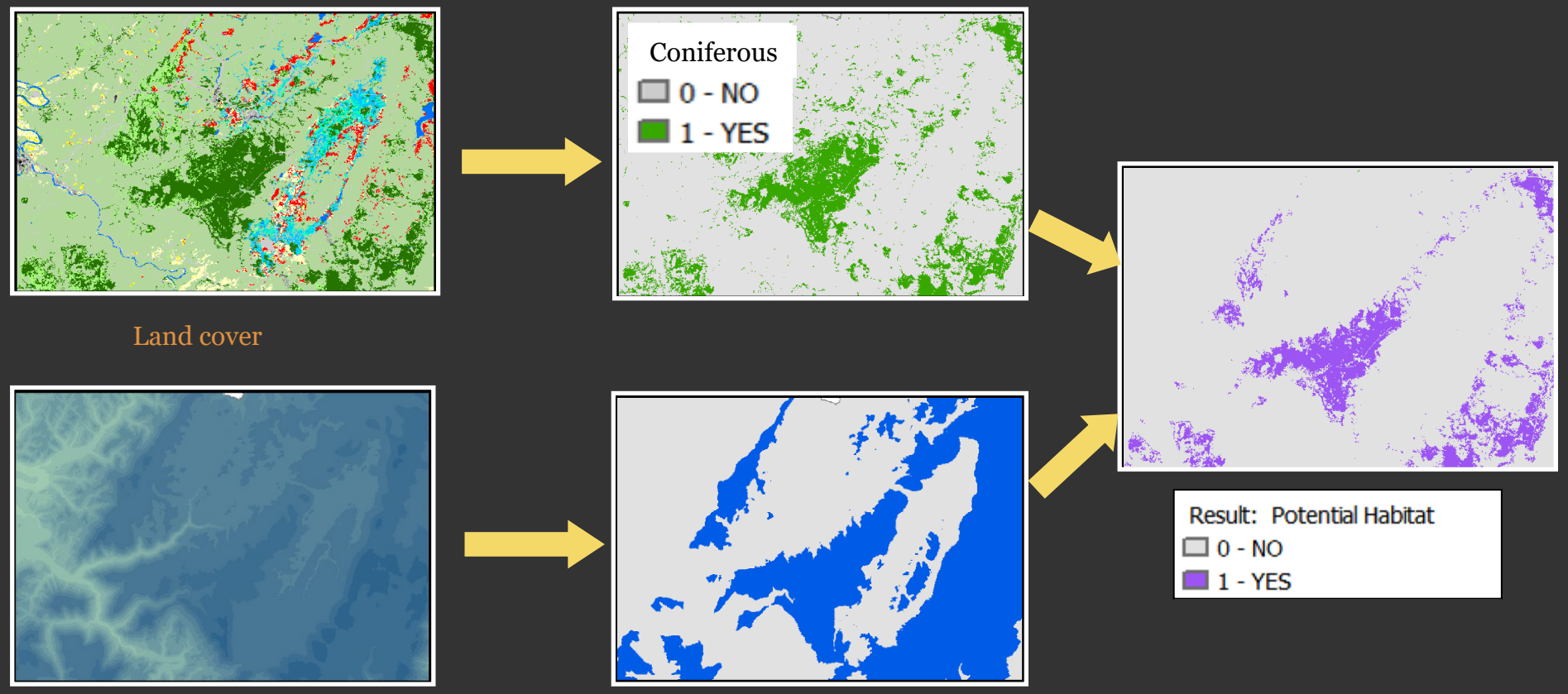
Vector Overlay Tools

- Dissolve
- Buffer
- Clip
- Intersect
- Union
- Spatial Query
- Table Query
- Etc.



Raster Overlay

- Habitat = coniferous forested areas over 1,000 meters in elevation

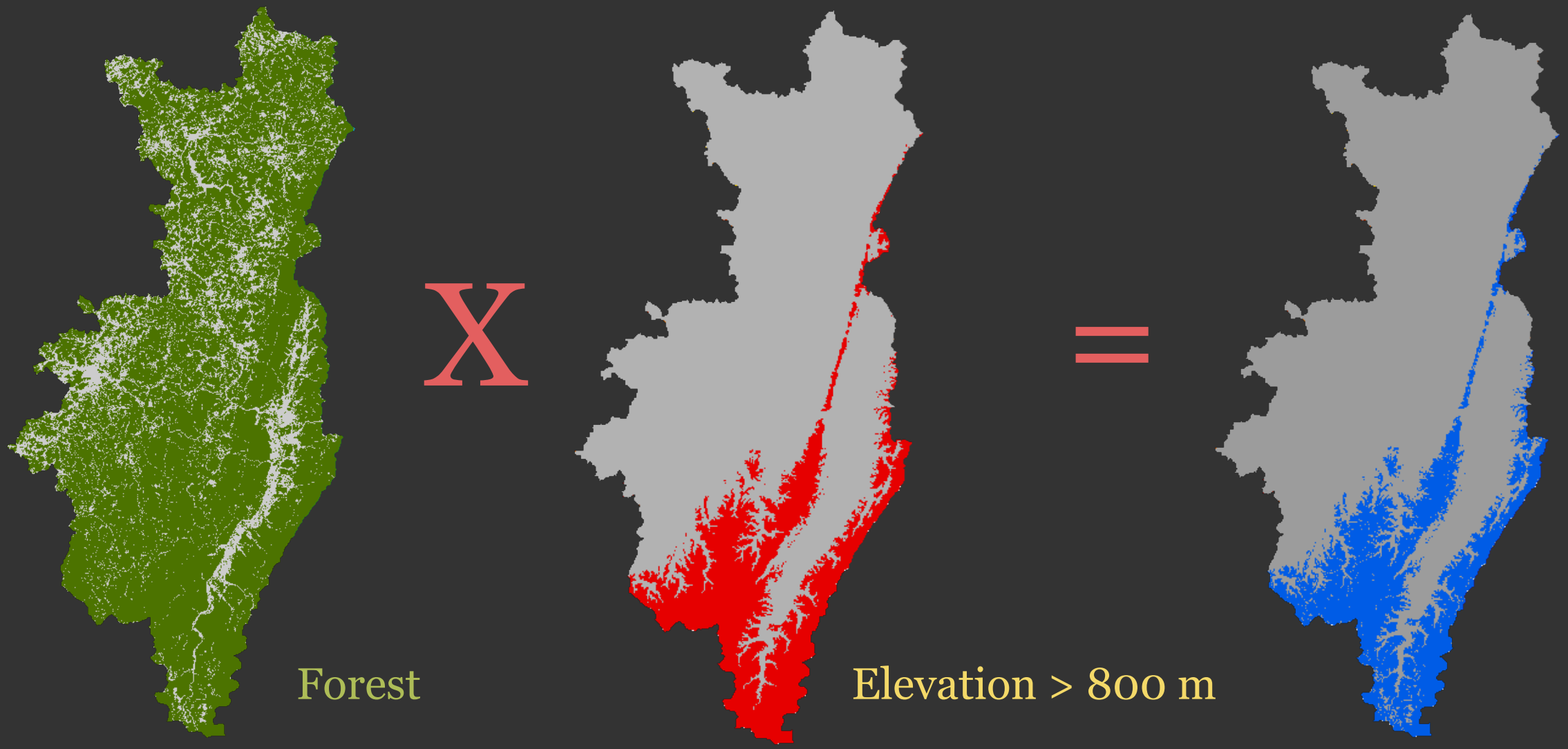


Land cover

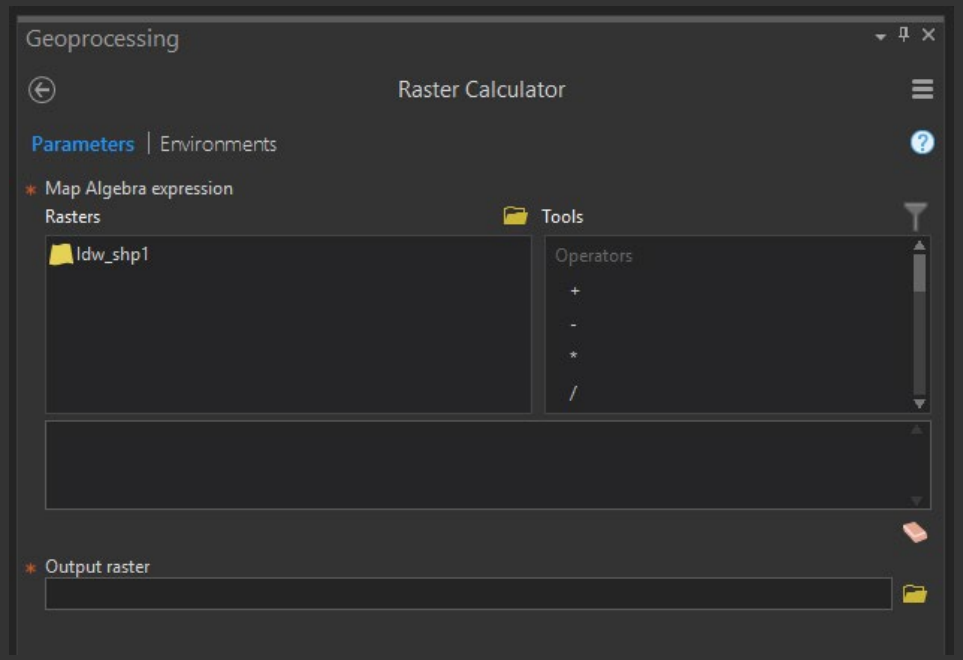
Elevation

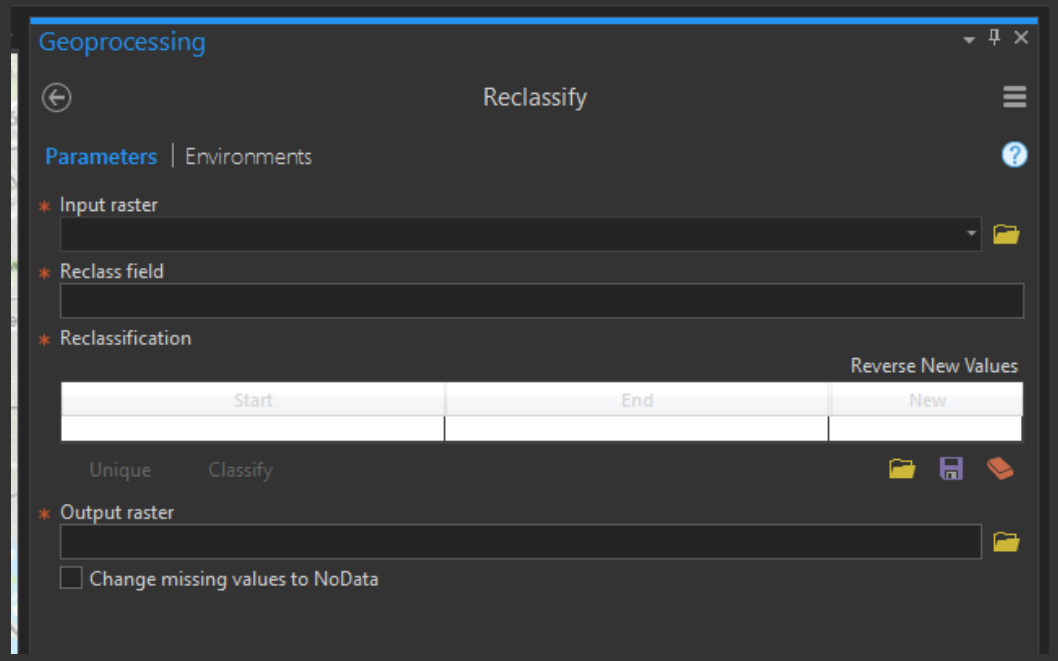
Example from Mike and Jackie Strager

Raster Overlay



Raster Overlay Tools





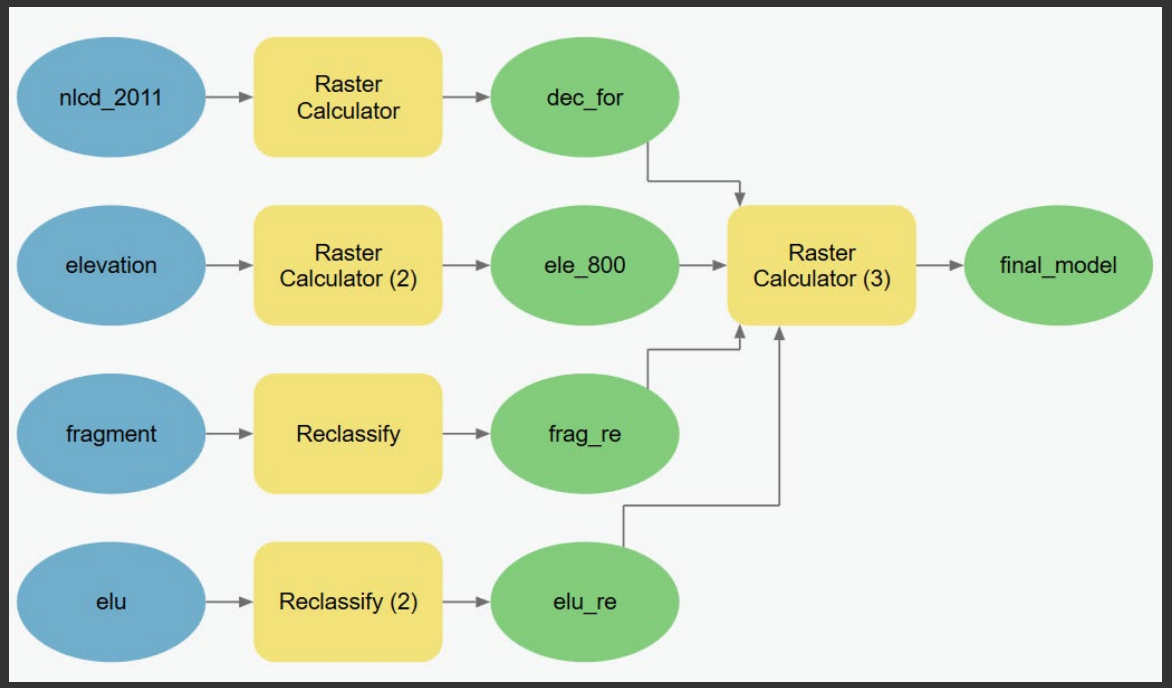
Limitations

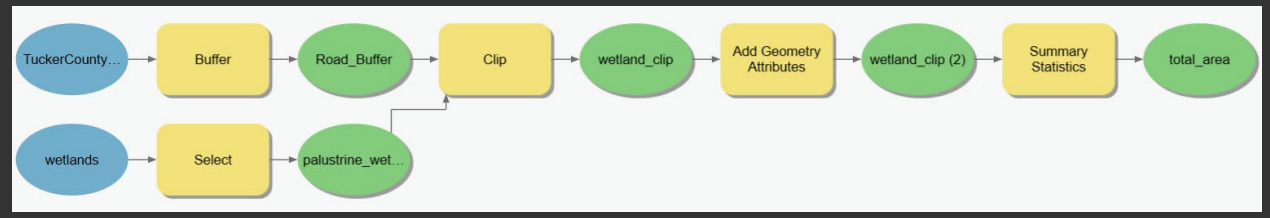
- Binary output may not be optimal or appropriate
- Difficult to assign thresholds/cut-offs
- Error can compound

ArcGIS ModelBuilder

Uses of ModelBuilder

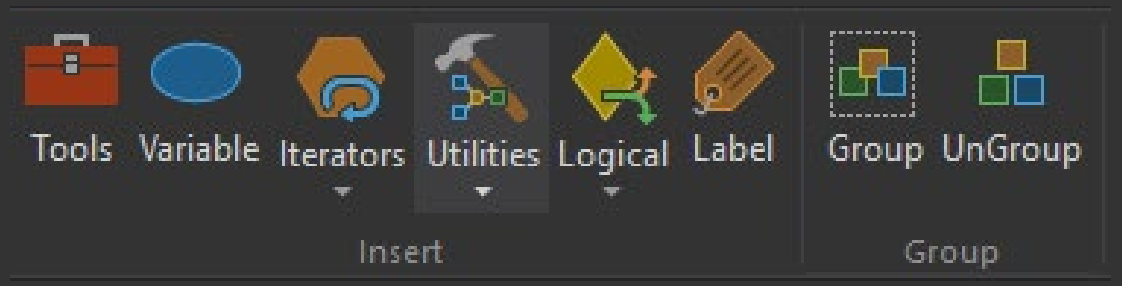
- Allows you to combine analysis tasks and tools into a single process chain
- Graphical interface
- No coding required





Components of ModelBuilder

- Input variables and data
- Input parameters
- Tools
- Intermediate outputs
- Final outputs
- Python scripts
- Other models
- Tables
- Iterators
- Etc.

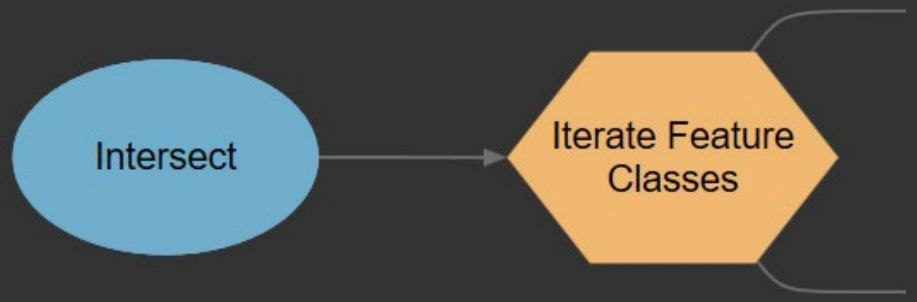


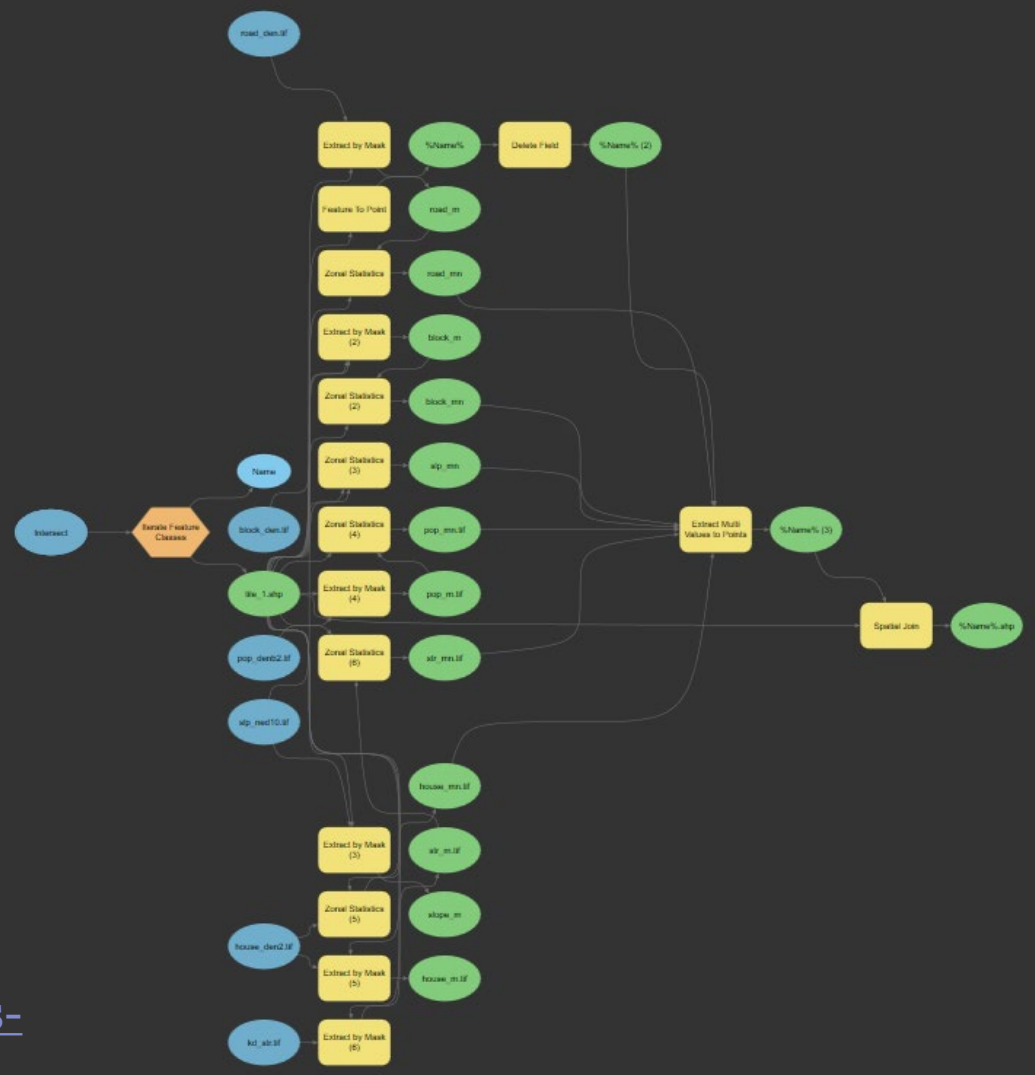
Components of ModelBuilder

Input Tool Output

Iterators

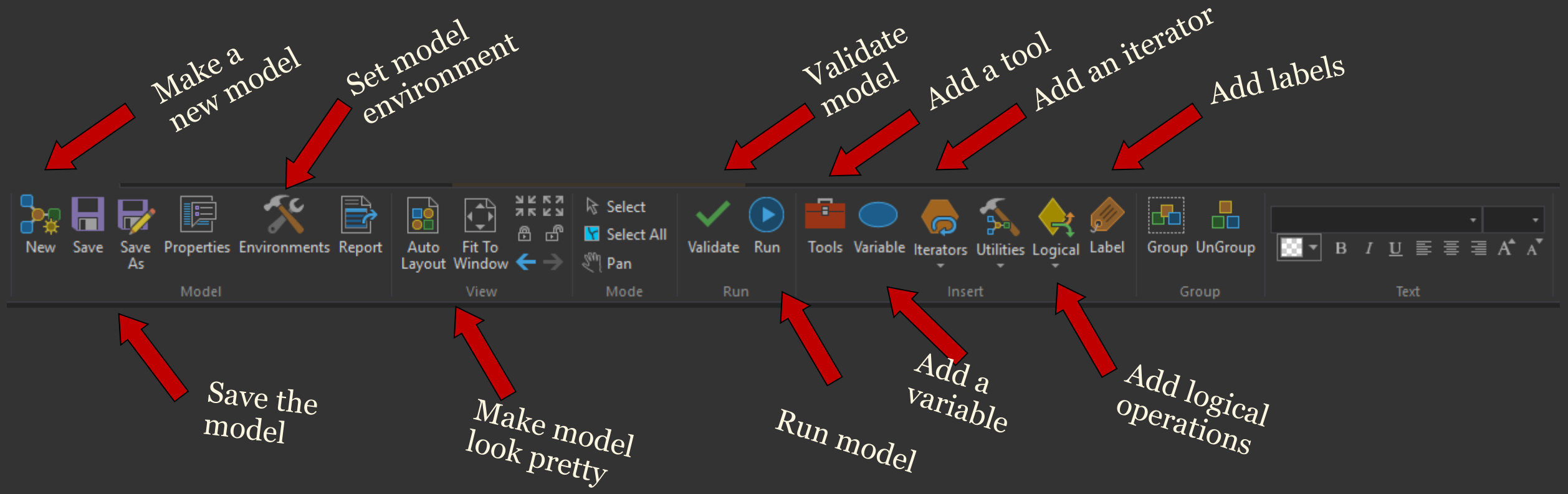
- For, While, Iterate Feature Selection, Iterate Row Selection, Iterate Field Values, Iterate Multivalues, Iterate Datasets, Iterate Feature Classes, Iterate Files, Iterate Rasters, Iterate Tables, Iterate Workspaces





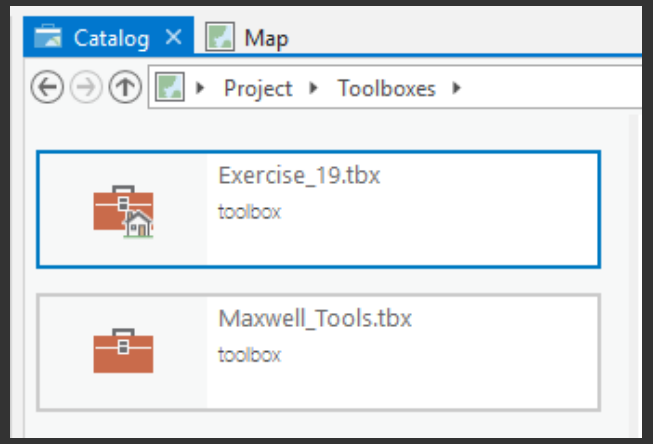
http://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/modelbuilder/iterators-for-looping.htm

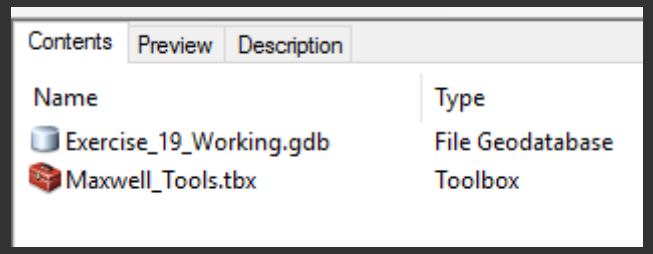
ModelBuilder Tools

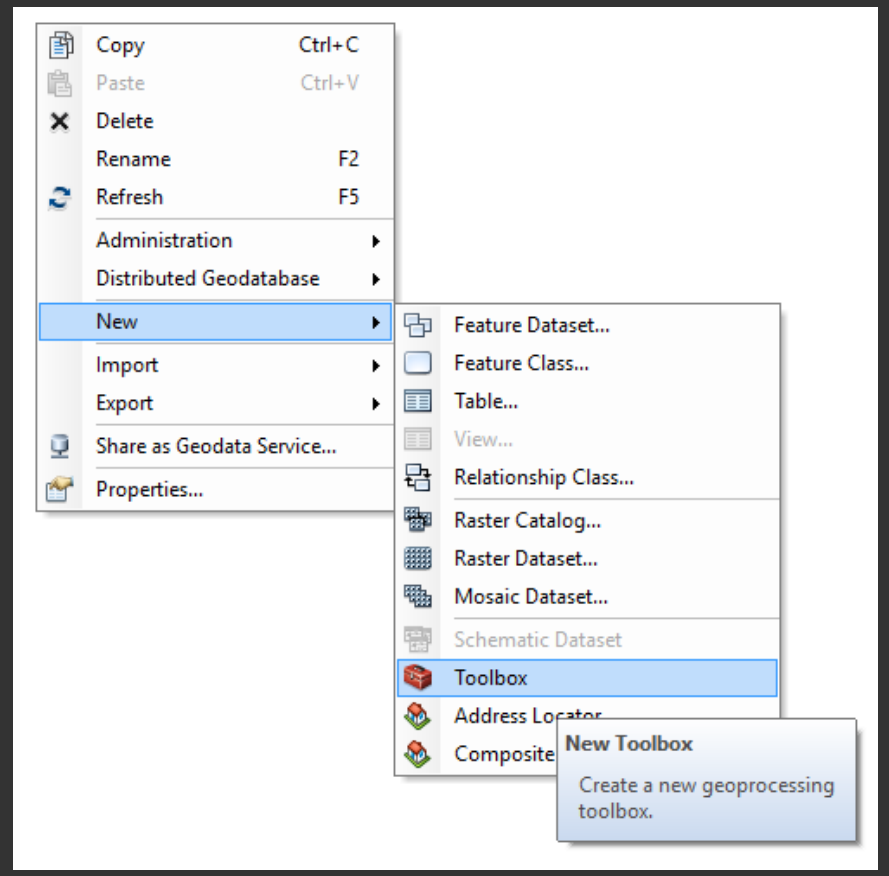


Saving a Model

- Models are saved inside of a Toolbox

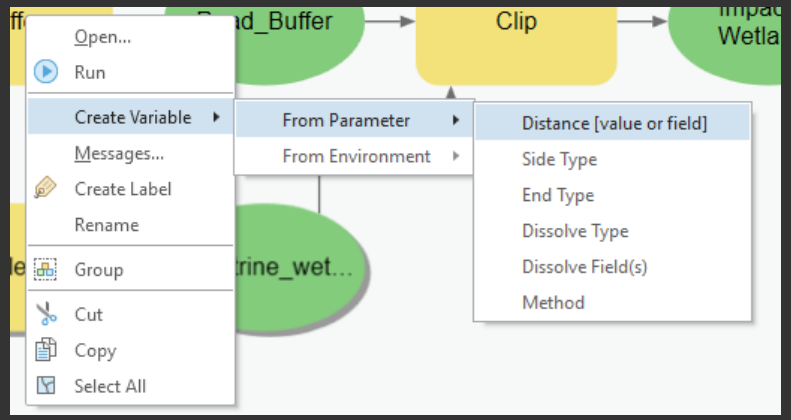


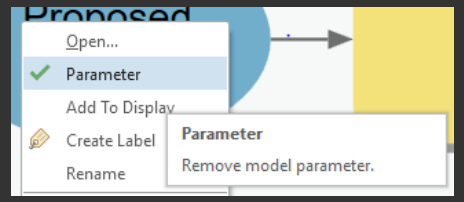


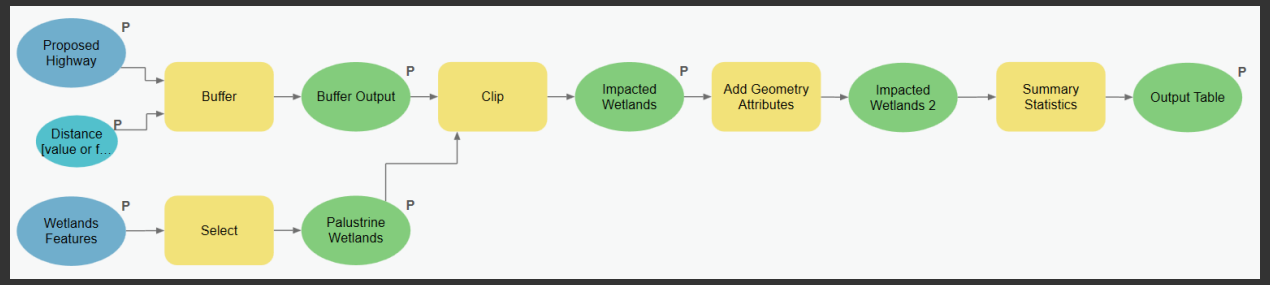


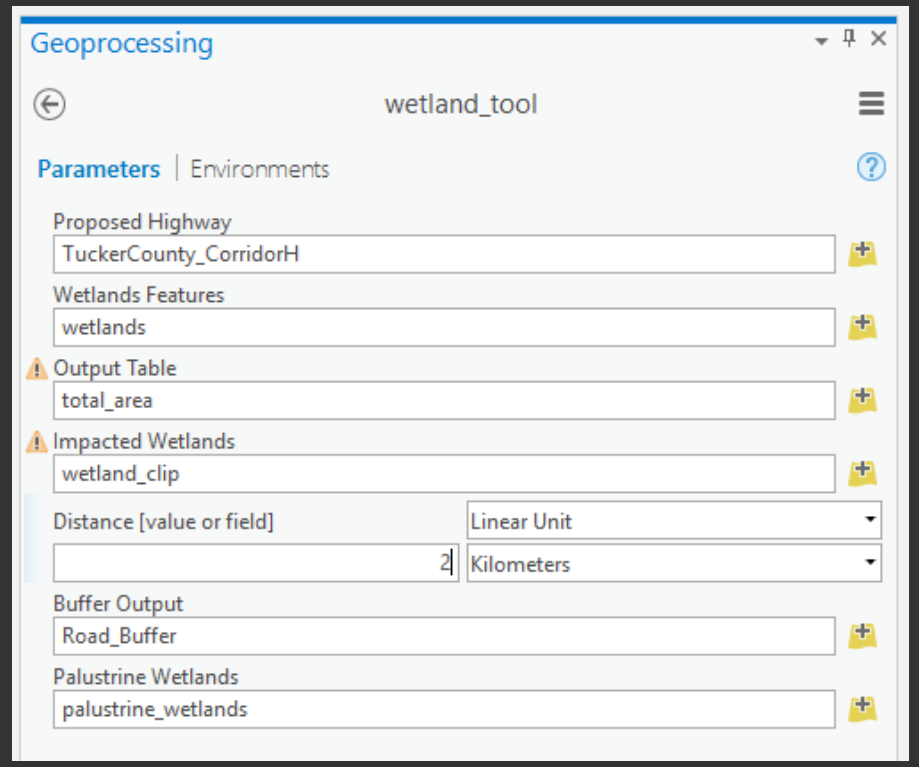
Creating a Tool from a Model

- Must specify user-defined parameters
- May want to rename inputs and outputs

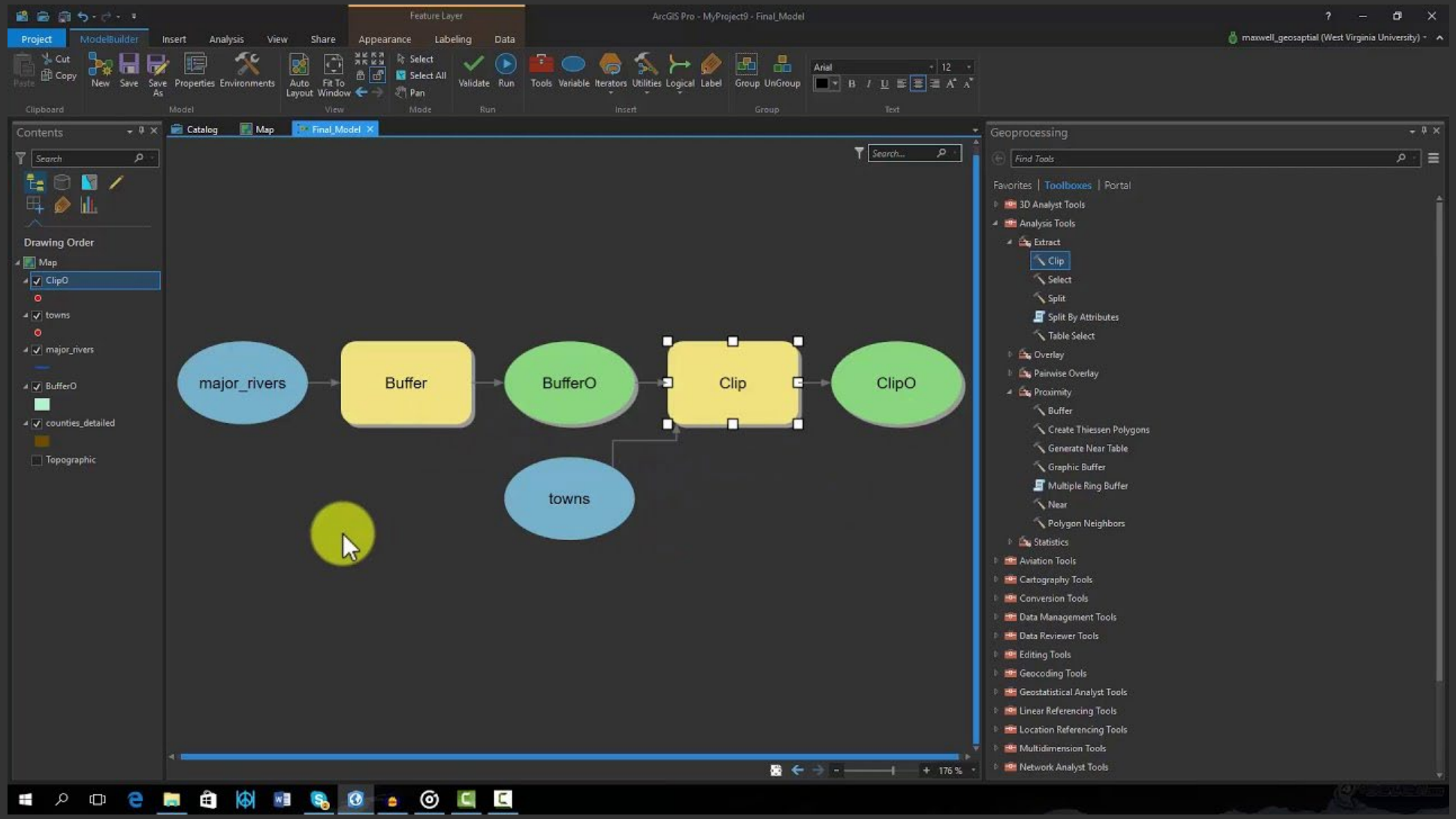








Video: ArcGIS ModelBuilder



Weighted Overlay

Overview of Weighted Overlay

- Used to create liberal models

Overall Score = [Criteria Score 1 X Weight 1] + [Criteria Score 2 X Weight 2] + [Criteria Score 3 X Weight 3] + [Criteria Score 4 X Weight 4] + [Criteria Score 5 X Weight 5] + ...

Steps in Weighted Overlay

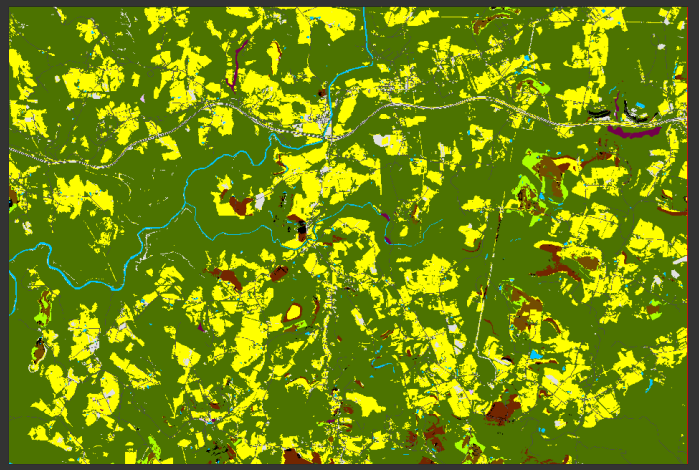
1. Define the overall problem or goal
   - What are you ranking?
   - What is the required result?
2. Determine evaluation criteria
3. Find spatial data that represent your criteria
4. Determine scoring within criteria
5. Determine weights for individual criteria
6. Calculate result
7. Evaluate and present results

It is important to consider:

1. What is the goal?
2. How will the model be used?
3. Who will use the model?
4. Does your model meet the need?

Defining Criteria

- Land cover/land use
- Landscape patterns
- Land ownership
- Human Infrastructure
  - Structures
  - Roads
  - Utilities
  - Airports
  - Etc.
- Distance from X
- Density of X
- Soil characteristics
- Terrain characteristics
  - Elevation
  - Slope
  - Aspect
  - Slope position
  - Etc.
  - Climate considerations
  - Temperature
  - Rainfall
  - Etc.
  - Surface hydrology
  - Etc.



Based on:

- Expert knowledge
- Stakeholder preferences
- Consult the literature

Data Considerations

1. Is the dataset available?
2. Is the dataset of adequate quality?
3. Is the dataset spatial or can it be made spatial? Can criteria be interpreted into spatial and quantifiable variables?
4. What is the spatial resolution?
5. Is it up to date?

Defining Scores

- What values are most suitable? Least suitable?
- Methods
- Discrete = Yes/No, 1/0, binary
- Ranges = Break into ranges
- Continuous = Linear increasing, linear decreasing, something more complex

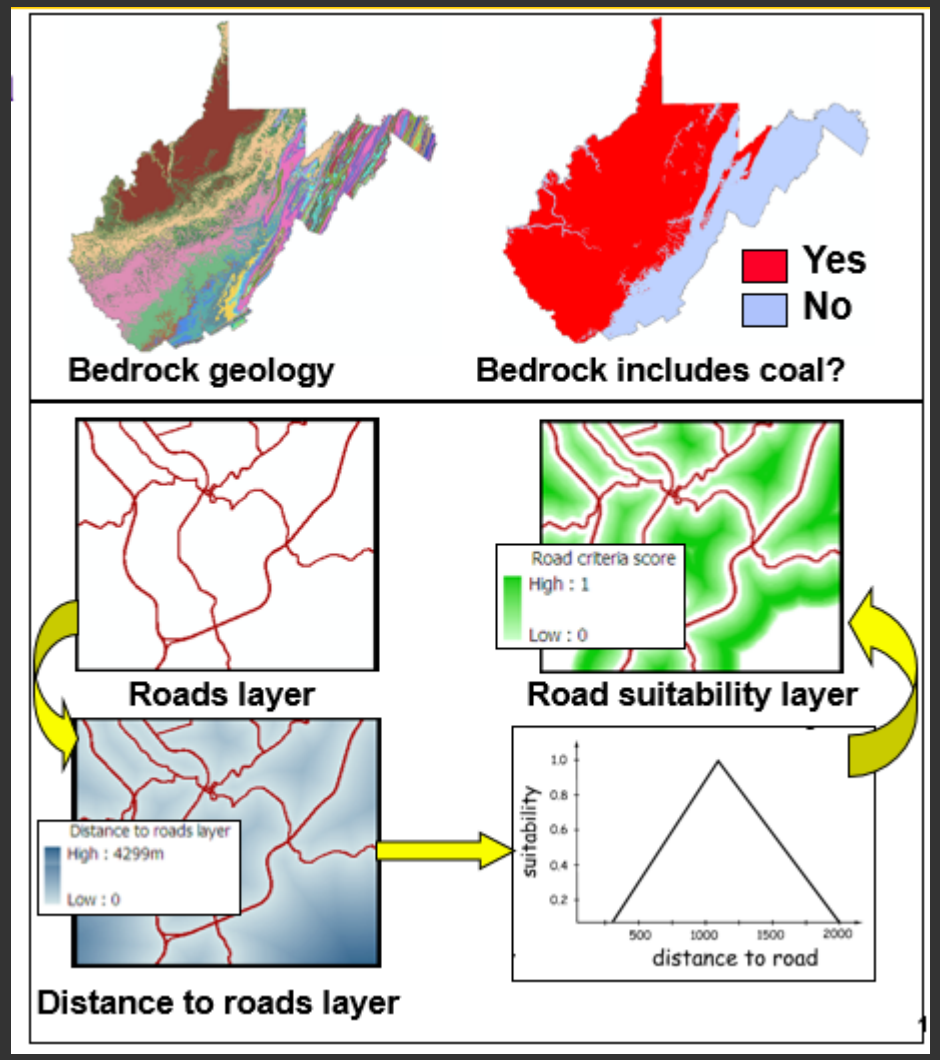
Distance to Towns	
Distance	Score
≤ 0.5 km	10
0.5 to 2 km	7
2 to 4 km	5
> 4 km	3

$$\frac{X_i - X_{min}}{X_{max} - X_{min}}$$

- It is important that all criteria are scored on the same scale

Defining Scores

- What values are most suitable? Least suitable?
- Methods
- Discrete = Yes/No, 1/0, binary
- Ranges = Break into ranges
- Continuous = Linear increasing, linear decreasing, something more complex
- It is important that all criteria are scored on the same scale



Example from Mike and Jackie Strager

Defining Weights

What values are most important?

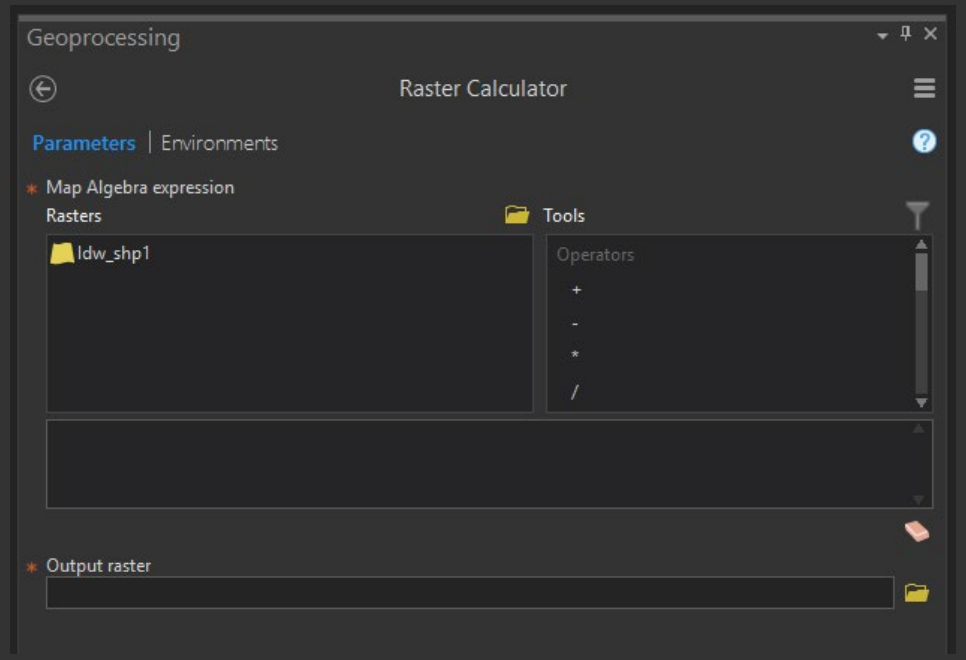
Based on:

1. Best judgement
2. Expert knowledge
3. Literature
4. Stakeholder preference (surveys)

Note: There are statistical methods to determine if multiple stakeholders value the same criteria similarly

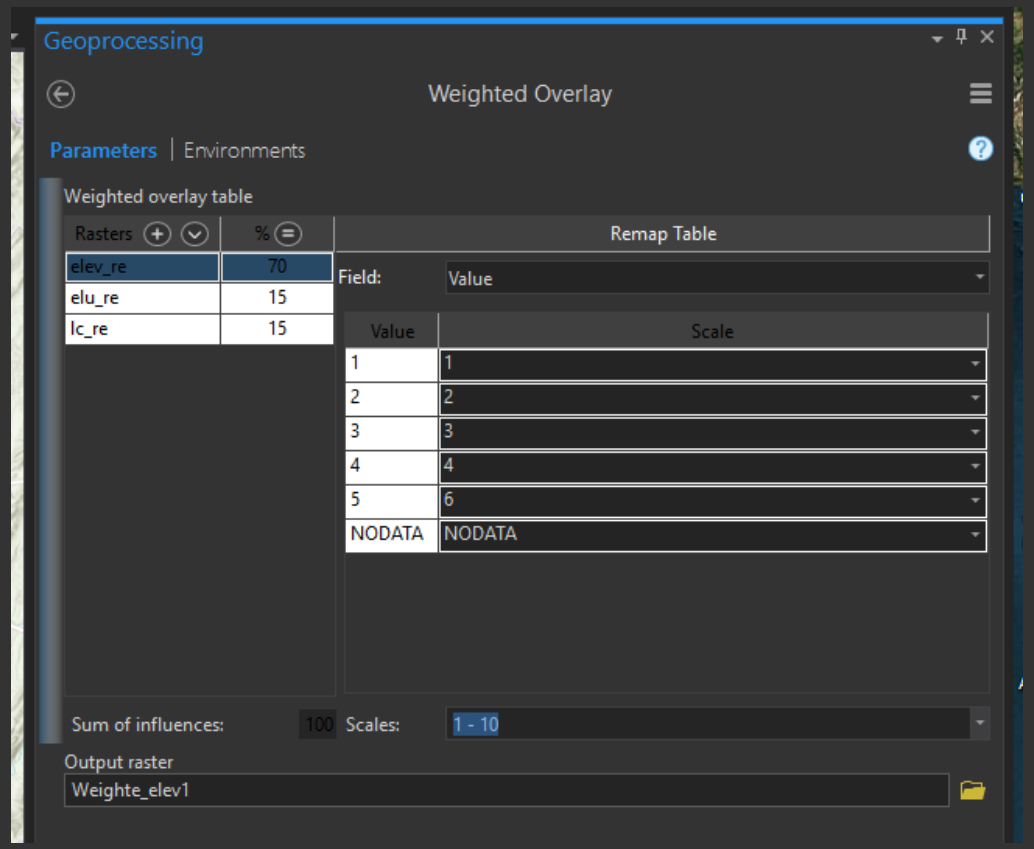
Calculating the Model: Raster Calculator

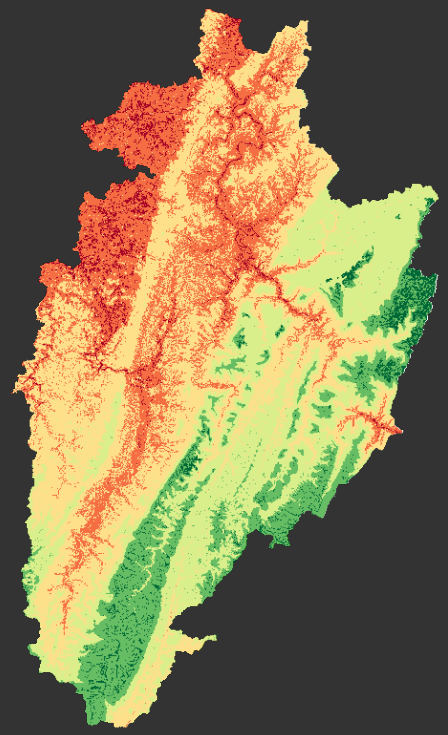
Overall Score = [Criteria Score 1 X Weight 1] + [Criteria Score 2 X Weight 2] + [Criteria Score 3 X Weight 3] + [Criteria Score 4 X Weight 4] + [Criteria Score 5 X Weight 5] + ...



Calculating the Model: Weighted Overlay Tool

- Provide scored grids
- Define % of influence for each criterion
- Sum of influence must by 100%
- Can define the scale of the output





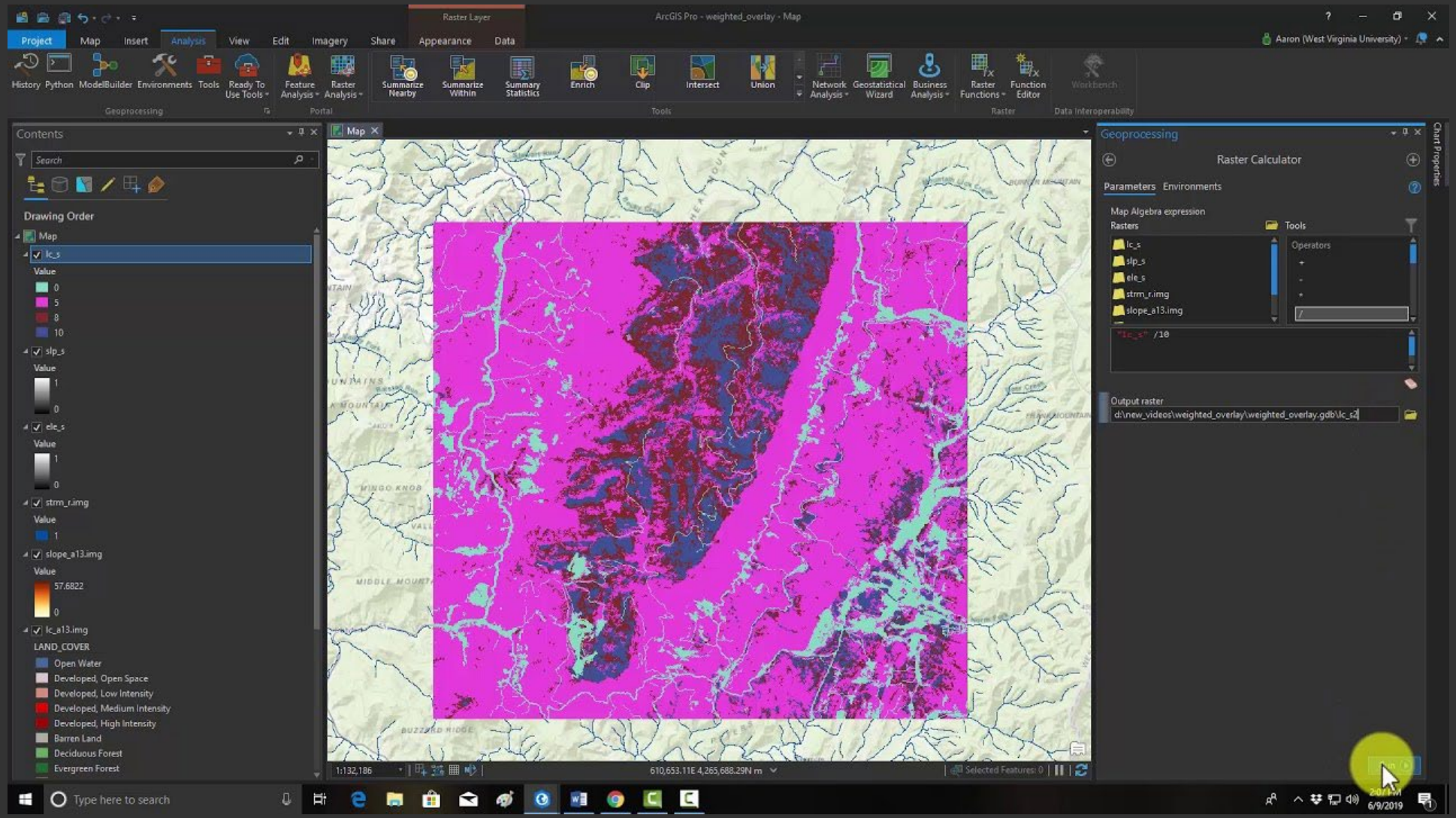
Calculating the Model

Criteria (C)	Scoring for criterion (C)		Weight of criterion (W)	Score for Site "X"
1. Distance to nearest interstate	10 5 1	within 1 mile within 5 miles within 10 miles	10	5 X 10 = 50
2. Distance to nearest airport	10 5 1	within 10 mile within 50 miles within 100 miles	15	10 X 15 = 150
3. Population in community	10 5 1	> 10,000 5,000 to 9,999 < 5,000	5	1 X 5 = 5

Overall score for site = sum((C1)(W1) + (C2)(W2) + (C3)(W3))= (50 + 15+5) = 205

Example from Mike and Jackie Strager

Video: Weighted Overlay



Model Validation

How do you know if your model makes sense?

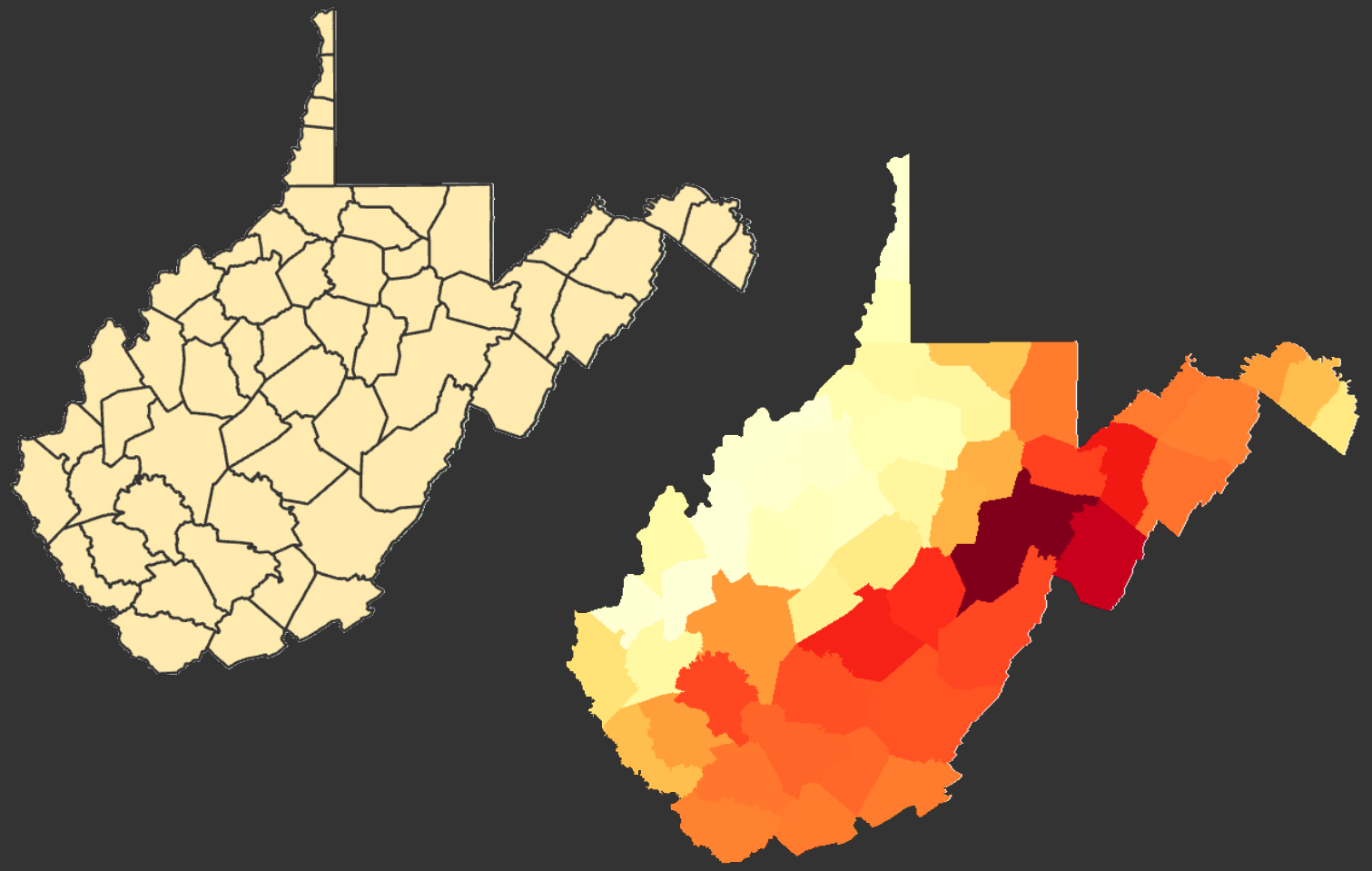
- Not always easy to do in a quantitative/statistical manner
- Visual assessment
- Present to:
  - Experts
  - Stakeholders
  - Public
- Solicit feedback
- Sources of uncertainty and error should be made known

Sensitivity Analysis

- How does your model change when varying scores/weights?
- How does your model change with changes in criteria?
- How does your model change with data source choice?

Summarize Model: Zonal Statistics

- Used to obtain statistics for a continuous raster grid within an area or zone as defined by another raster or a polygon vector layer
- Can obtain results as a raster surface for a single statistic



Elevation range by county

Summarize Model: Zonal Statistics as Table

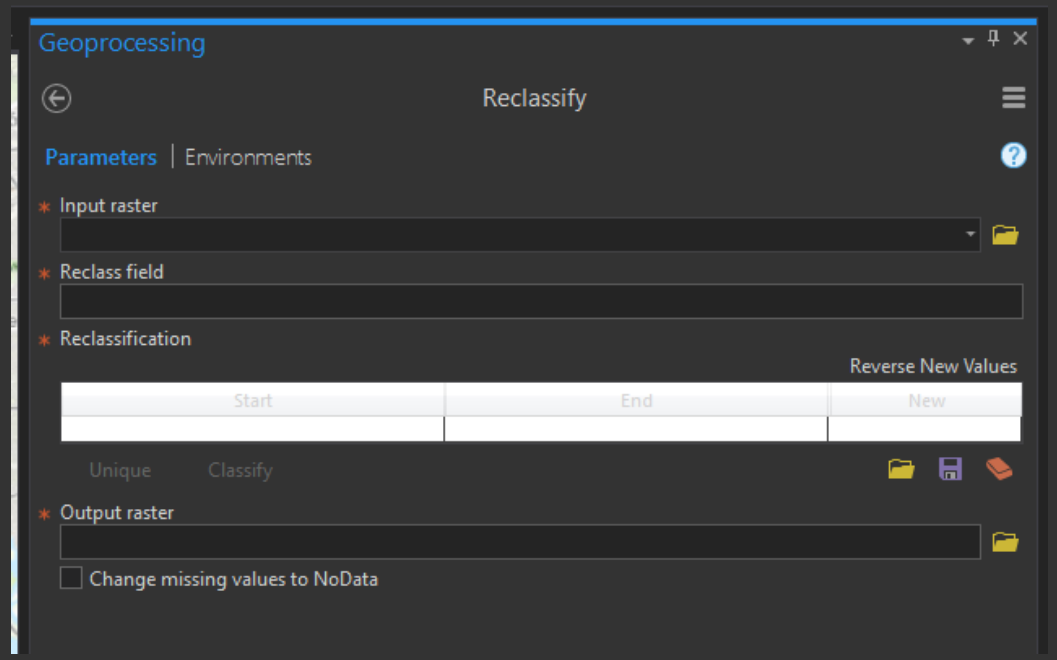
- Used to obtain statistics for a continuous raster within an area or zone as defined by another raster or a polygon vector layer
- Can obtain multiple statistics as a table (mean, majority, maximum, median, minimum, minority, range, standard deviation, sum, and variety)
- Can use table join to join the results back to the zone data

NAME	ZONE_CODE	COUNT	AREA	MIN	MAX	RANGE	MEAN	STD	SUM	VARIETY	MAJORITY	MINORITY	MEDIAN
Ohio	1	313037	281733300	190	434	244	335.185646	53.554422	104925509	245	190	433	349
Marshall	2	896296	806666400	181	488	307	346.481389	58.372259	310549883	308	190	181	361
Preston	3	1872347	1685112300.000001	257	1047	790	626.009406	122.181059	1172106833	791	542	257	606
Morgan	4	661649	595484100	0	711	711	278.286232	97.964891	184127807	606	261	0	260
Monongalia	5	1052070	946863000	214	788	574	404.379005	94.363254	425435020	561	265	214	381
Wetzel	6	1038769	934892100	170	518	348	356.66806	61.03432	370495724	346	380	170	365
Mineral	7	947525	852772500	167	960	793	399.815696	179.897201	378835367	794	262	167	332
Berkeley	8	925921	833328900	69	664	595	201.454417	81.850591	186530875	596	156	71	174
Marion	9	895803	806222700	249	612	363	374.991877	51.121075	335918848	361	363	249	372
Tyler	10	749712	674740800	158	471	313	288.145992	44.391612	216026508	303	290	159	289
Hampshire	11	1854635	1669171500.000001	143	924	781	378.330457	114.098121	701664907	779	354	143	362
Jefferson	12	608694	547824600	71	520	449	164.008096	43.217139	99830744	447	162	72	159
Pleasants	13	386824	348141600	165	381	216	269.91643	42.829357	104410153	217	179	165	276
Harrison	14	1198449	1078604100	258	559	301	363.909574	39.328835	436127065	301	354	258	361
Taylor	15	505225	454702500	256	636	380	414.409028	58.841298	209369801	381	356	257	409

Elevation statistics by county

Model Generalization

- You can bin your output
  - Not suitable
  - Moderately suitable
  - Highly suitable



Summarize Model: Tabulate Area

- Used to obtain areas for categories in a categorical raster within an area or zone as defined by another raster or polygon vector layer
- Areas will be in square map units
- Can use table join to join the results back to the zone data

NAME	VALUE_11	VALUE_21	VALUE_22	VALUE_23	VALUE_24	VALUE_31	VALUE_41	VALUE_42	VALUE_43	VALUE_52	VALUE_71	VALUE_81	VALUE_82	VALUE_90	VALUE_95
Ohio	7220700	32448600	12725100	9791100	4060800	1828800	163184400	243900	27000	5400	5736600	33041700	11615400	10800	5400
Marshall	17972100	52911900	8149500	7511400	3639600	2680200	594241200	1294200	90000	61200	17165700	90684900	10204200	368100	233100
Preston	12051000	109386000	8920800	5368500	1096200	11108700	1264432500	23017500	12316500	691200	2385900	173321100	58838400	1421100	1750500
Morgan	5271300	38037600	5807700	919800	235800	339300	448716600	15448500	14912100	1154700	1809000	57774600	4842000	153900	11700
Monongalia	16850700	75492900	25077600	17461800	5606100	6173100	688201200	1125900	1607400	403200	10475100	74536200	23842800	216000	14400
Wetzel	7335900	48271500	5332500	2608200	595800	570600	832127400	1025100	86400	146700	6072300	23953500	6749100	71100	18900
Mineral	7317900	51754500	9358200	4086900	863100	3462300	625183200	14517900	18436500	1299600	2419200	110389500	3740400	216000	113400
Berkeley	5900400	57793500	40197600	11976300	4634100	1073700	387274500	19776600	16367400	1209600	3294900	257097600	23208300	2637900	891000
Marion	7502400	64860300	17891100	8419500	2449800	2124900	618721200	819000	654300	75600	7243200	58851000	16449300	9900	29700
Tyler	11124000	31230900	2791800	820800	271800	66600	556369200	4665600	219600	117000	3649500	53750700	9626400	40500	22500
Hampshire	13815900	46972800	7014600	1631700	397800	535500	1241870400	43697700	69669000	4070700	4116600	234529200	1488600	62100	59400
Jefferson	9701100	59721300	18992700	5191200	1332000	877500	112617000	24759900	8029800	481500	1954800	268405200	31912200	2241000	2356200
Pleasants	13662900	21668400	3581100	1895400	518400	375300	276406200	5797800	348300	194400	2730600	18234000	2688300	0	133200
Harrison	6086700	81247500	29097000	16684200	4040100	6741900	757297800	470700	2953800	93600	6548400	156815100	10492200	20700	39600
Taylor	8799300	33958800	5025600	1863900	459900	1993500	329850000	334800	882000	0	1338300	65516400	4537800	4500	152100

Area of land cover by county

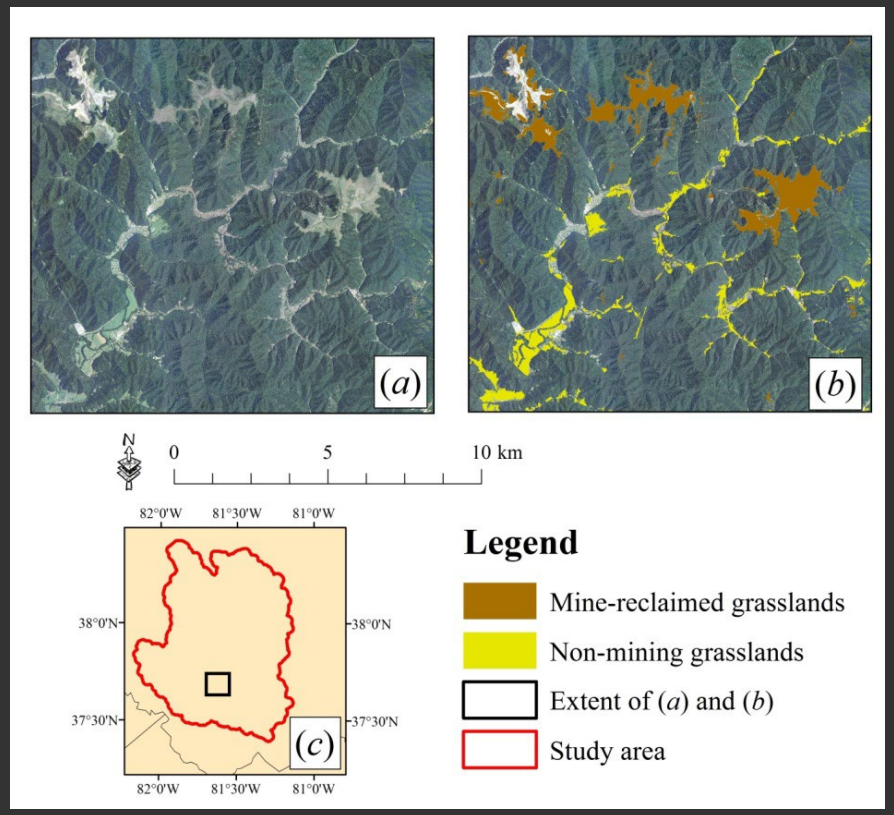
Limitations

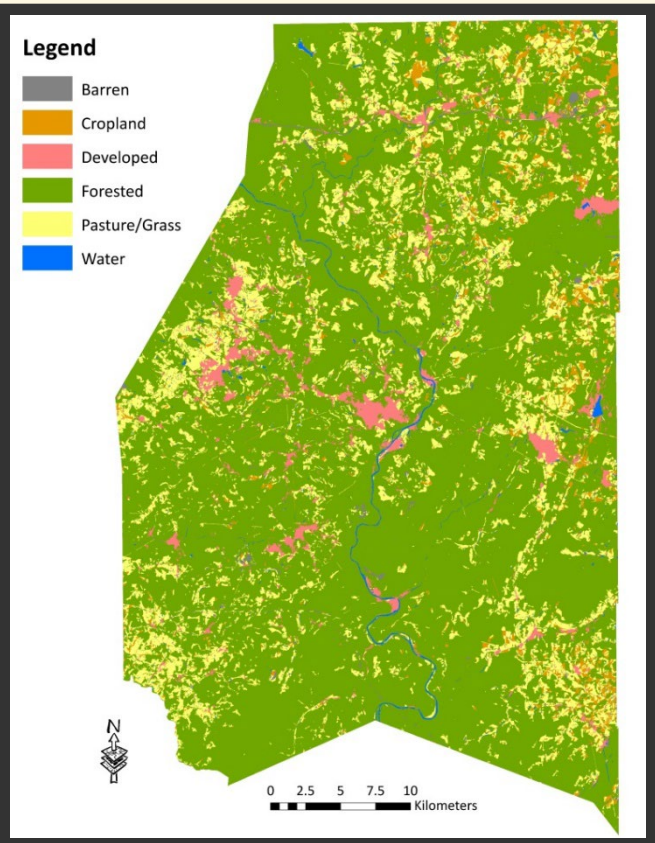
- Error can compound
- Difficulty determining scores
- Difficulty determining weights
- Difficulty with assessment

Advanced Modeling

Predicting a Categorical or Binary Outcome

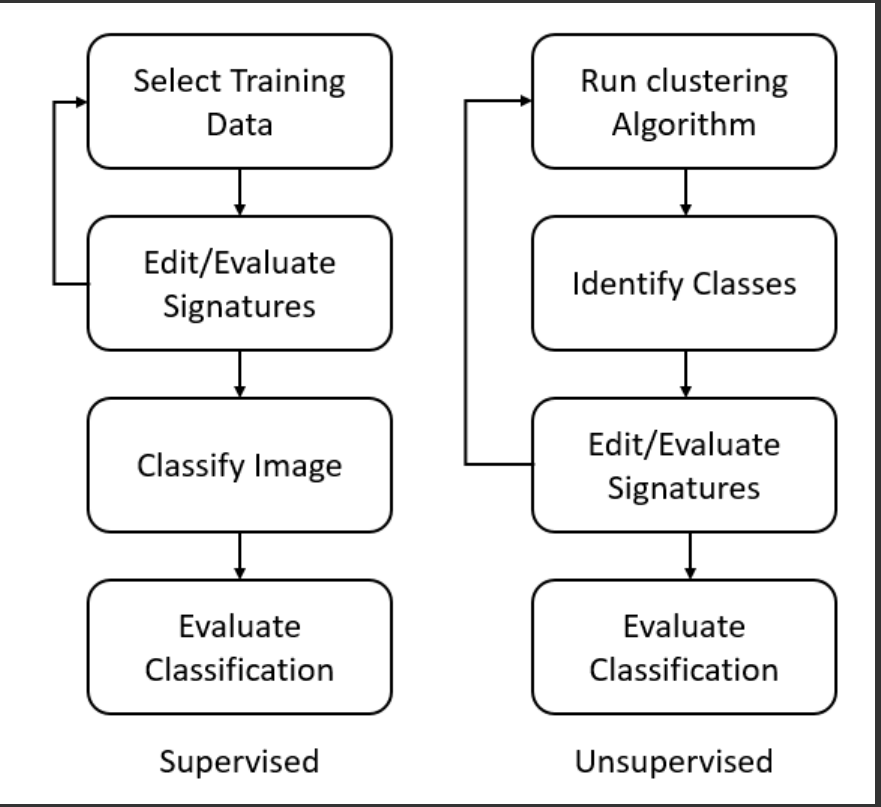
- Logistic regression
- Maximum likelihood
- Parallelepiped
- Minimum-Distance-to-Means
- Machine learning





Example: Land Cover Classification

- Supervised and Unsupervised classification



Example: Land Cover Classification

INTERNATIONAL JOURNAL OF REMOTE SENSING, 2018
VOL. 39, NO. 9, 2784–2817
https://doi.org/10.1080/01431161.2018.1433343

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

Check for updates

Implementation of machine-learning classification in remote sensing: an applied review

Aaron E. Maxwell, Timothy A. Warner and Fang Fang

Department of Geology and Geography, West Virginia University, Morgantown, WV, USA

ABSTRACT
Machine learning offers the potential for effective and efficient classification of remotely sensed imagery. The strengths of machine learning include the capacity to handle data of high dimensionality and to map classes with very complex characteristics. Nevertheless, implementing a machine-learning classification is not straightforward, and the literature provides conflicting advice regarding many key issues. This article therefore provides an overview of machine learning from an applied perspective. We focus on the relatively mature methods of support vector machines, single decision trees (DTs), Random Forests, boosted DTs, artificial neural networks, and *k*-nearest neighbours (*k*-NN). Issues considered include the choice of algorithm, training data requirements, user-defined parameter selection and optimization, feature space impacts and reduction, and computational costs. We illustrate these issues through applying machine-learning classification to two publically available remotely sensed data sets.

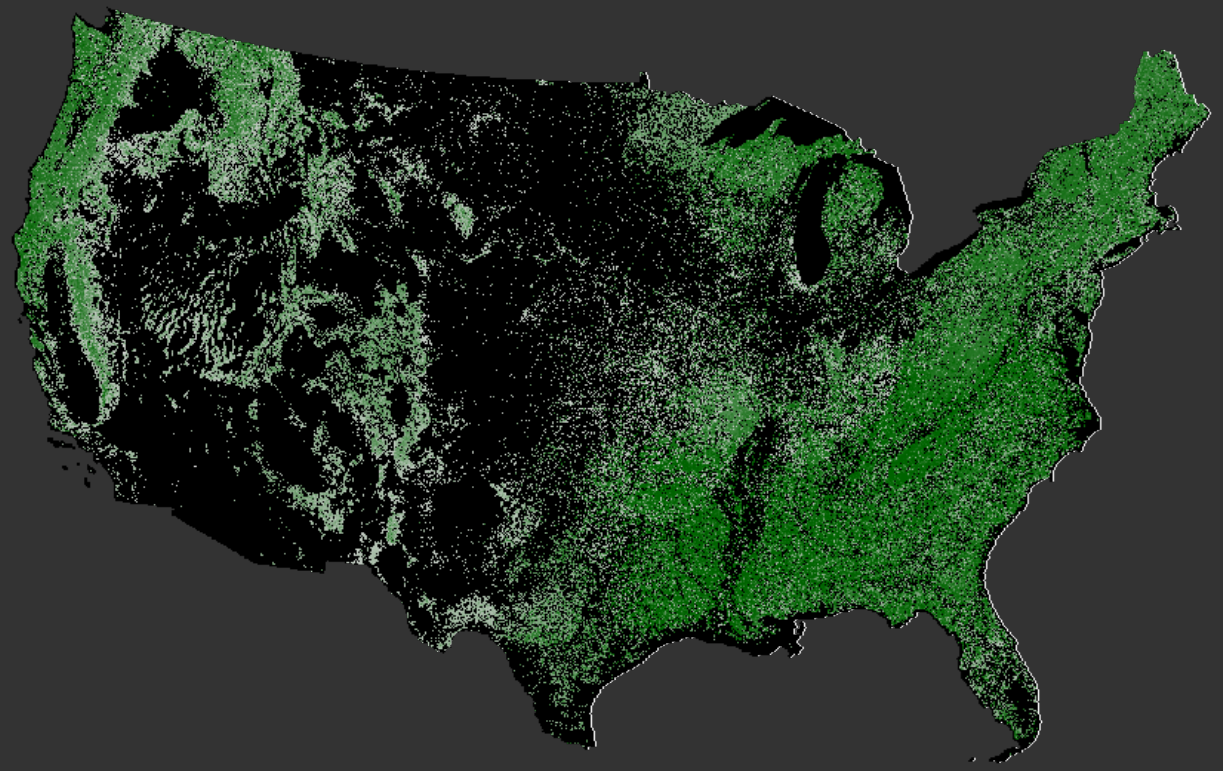
ARTICLE HISTORY
Received 28 October 2017
Accepted 16 December 2017

KEYWORDS
Land-cover classification; image classification; land-cover mapping; machine learning

https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1433343

Predicting a Numeric or Continuous Outcome

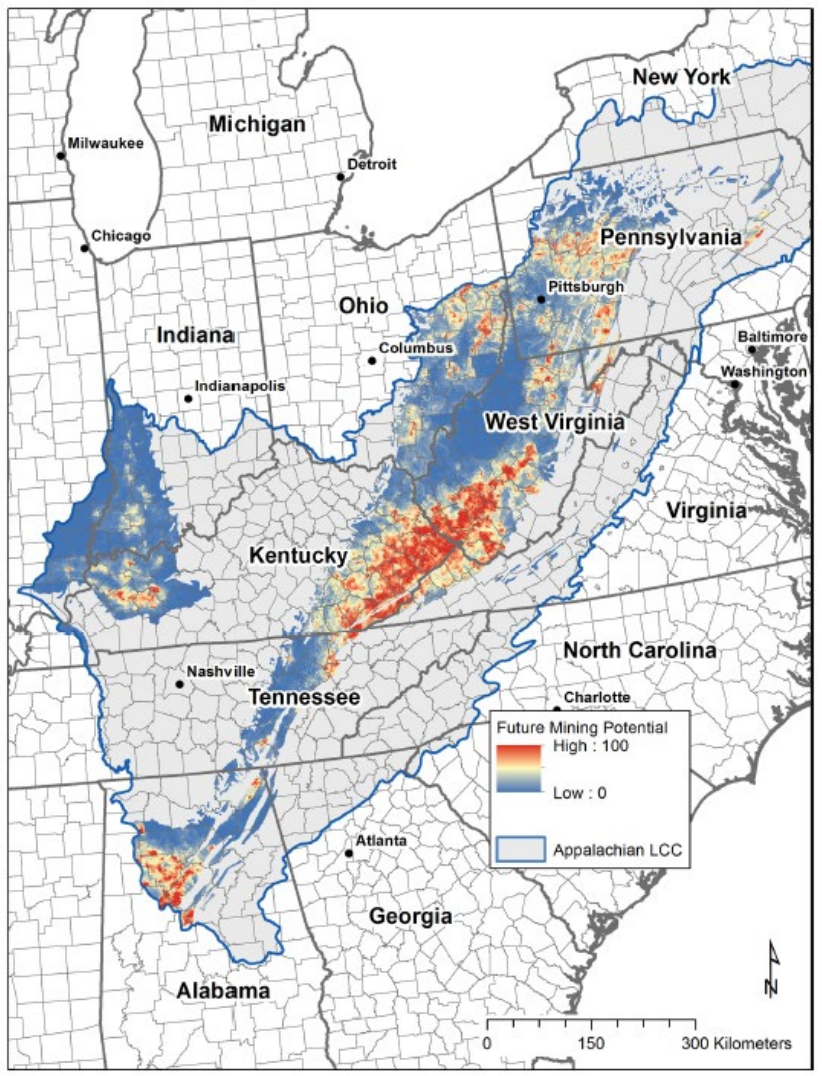
- Linear regression
- Multiple linear regression
- Polynomial regression
- Machine learning



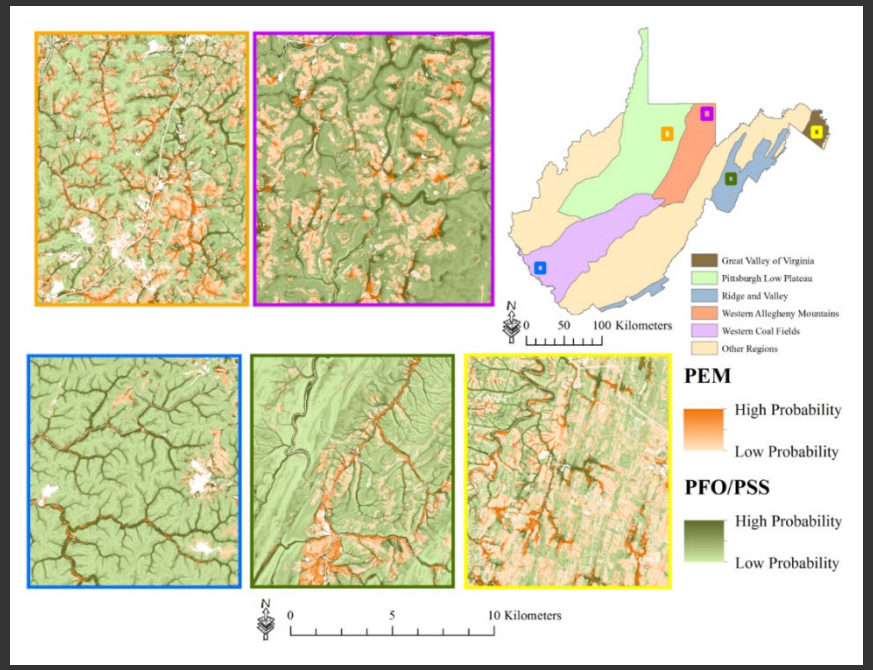
Percent Canopy Cover NLCD 2011

https://www.mrlc.gov/index.php

Predicting a Probabilistic Outcome



- Logistic regression
- Maxent
- Machine learning



Predicting Count Outcomes

- Poisson Regression
- Machine Learning

Modeling Physical Processes

- A model of a physical process
- Example: Revised Universal Soil Loss Equation (RUSLE)

A = RKLSCP

A = Average soil loss

R = Rainfall-runoff erosivity factor

K = Soil erodibility factor

L = Slope length factor

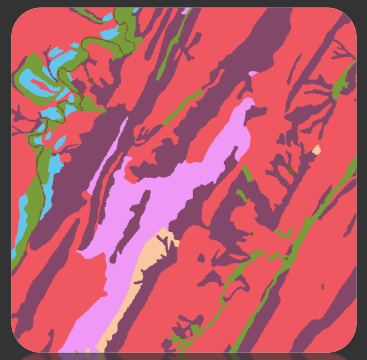
S = Slope steepness factor

C = Crop management factor

P = Support practice factor

https://www.ars.usda.gov/southeast-area/oxford-ms/national-sedimentation-laboratory/watershed-physical-processes-research/docs/revised-universal-soil-loss-equation-rusle-welcome-to-rusle-1-and-rusle-2/

Empirical Models



Thing you want to predict

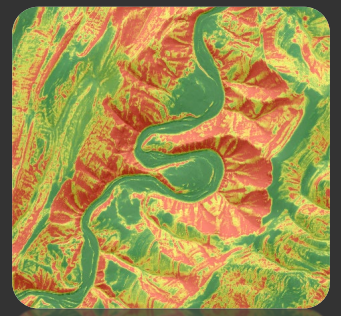
Things that might help predict the new thing

Learning algorithm

Trained Model

New things to predict

Prediction s

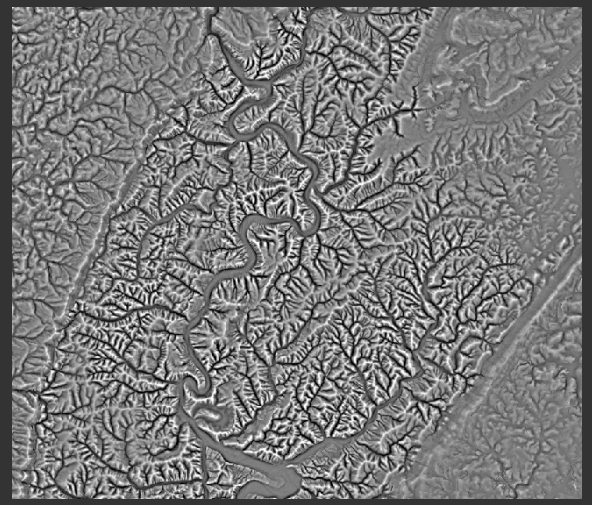


Machine Learning = Learning from Examples (Empirical)

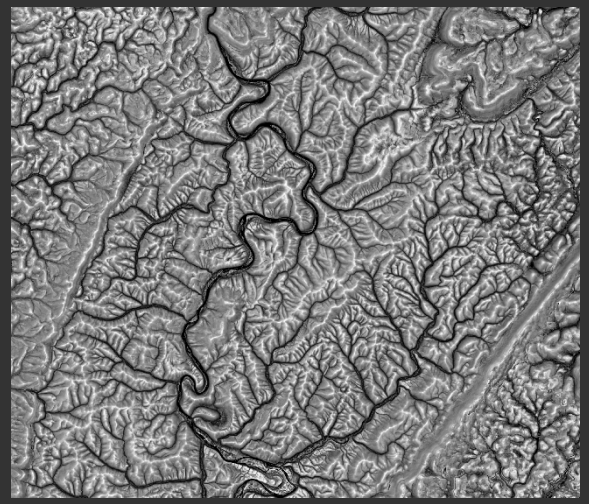
Developing Training Data

- Spatially explicit
- Representative of the population
- Adequate number of samples
- Adequate number of samples per category
- Accurate

Developing Predictor Variables



Slope Position

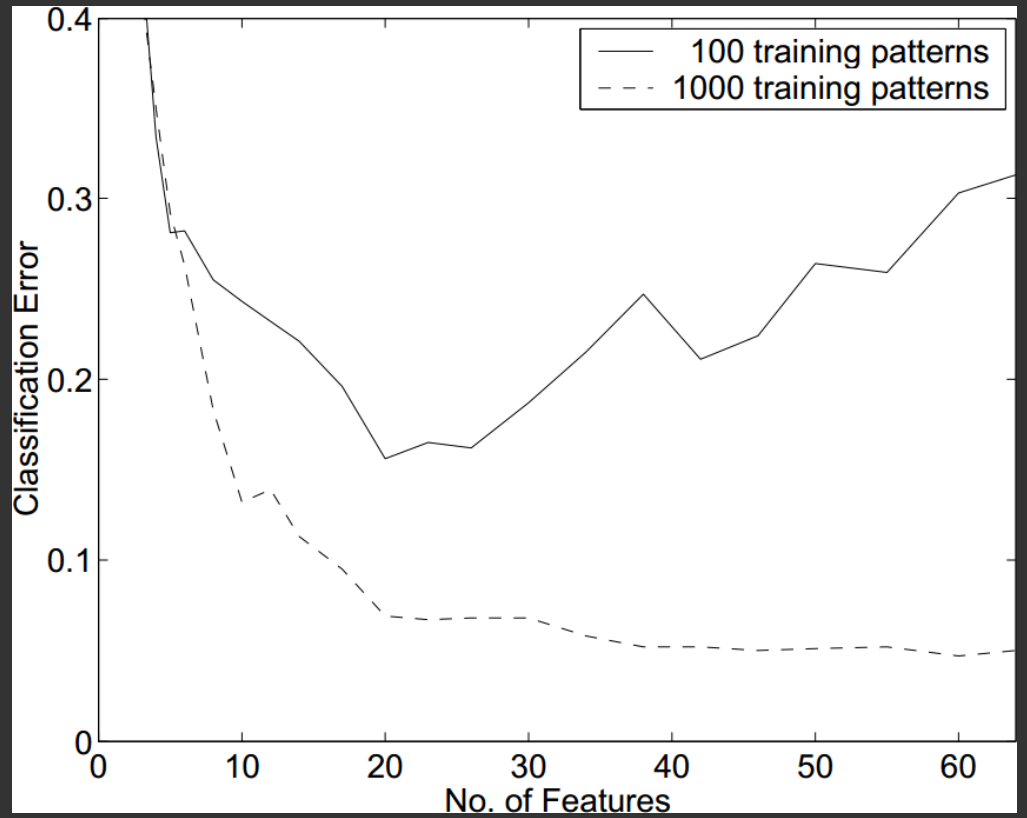


Dissection

Roughness

- Spatially explicit
- Accurately georeferenced
- Accurate/precise
- Timely
- Adequate spatial resolution
- Available

Developing Predictor Variables



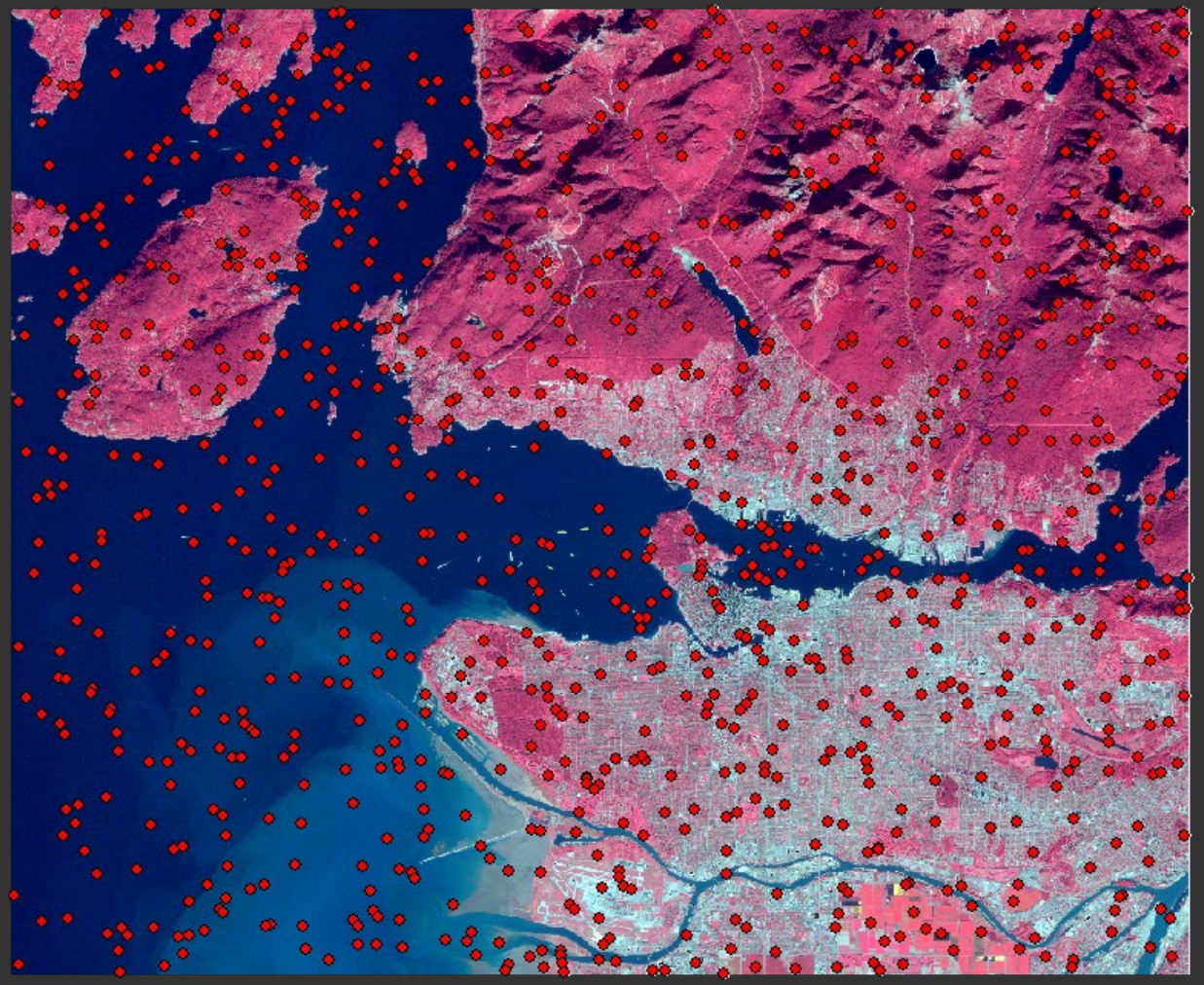
Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.1 (2000): 4-37.

- Hughes Phenomenon
- "Curse of dimensionality"

Hughes, G.F. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.

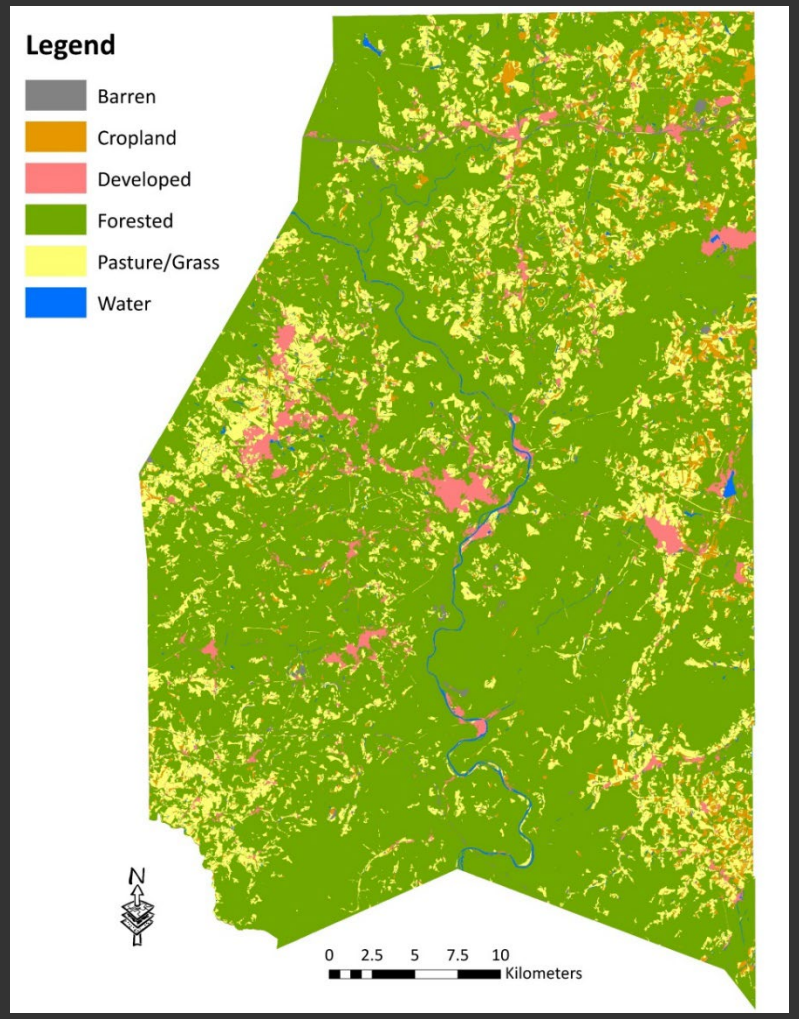
Modified by a slide created by Tim Warner

Developing Validation Data



- Unbiased
- Randomized
- Non-overlapping with your training data
- Correctly proportioned
- Accurate

Assessing a Categorical Prediction: Error or Confusion Matrix



		Reference Data							
		Forested	*Pasture/ Grass*	*Barren*	*Cropland*	*Developed*	*Water*	Total	1 – Commission Error
Classified Data	*Forested*	91	8	1	11	0	2	113	81%
	Pasture/ Grass	4	81	0	15	0	0	100	81%
	Barren	0	0	87	2	10	2	101	86%
	Cropland	3	11	0	69	0	0	83	83%
	Developed	0	0	10	0	73	0	83	88%
	Water	2	0	2	3	17	96	120	80%
	Total	100	100	100	100	100	100		
	1 – Omission Error	91%	81%	87%	69%	73%	96%		

Assessing a Numeric/Continuous Prediction: RMSE

RMSE = $\sqrt{\Sigma(y - \hat{y})^2/n}$

- Root Mean Square Error (RMSE) will be in the units of y

Mean Square Error (MSE) = $\Sigma(y - \hat{y})^2/n$

Assessing a Numeric/Continuous Prediction: R^2

$$R^2 = \frac{Total\ Sum\ of\ Squares - Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$$

Total Sum of Squares = $\Sigma(y - \bar{y})^2$

Residual Sum of Squares = $\Sigma(y - \hat{y})^2$

- Proportion of variance in y explained by the model

Assessing a Numeric/Continuous Prediction: Adjusted R^2

- When you have more than one x (multiple regression) you have to use adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{(N-p-1)}$$

N = sample size, ***p*** = number of variables

Other Metrics

Akaike Information Criterion (AIC)

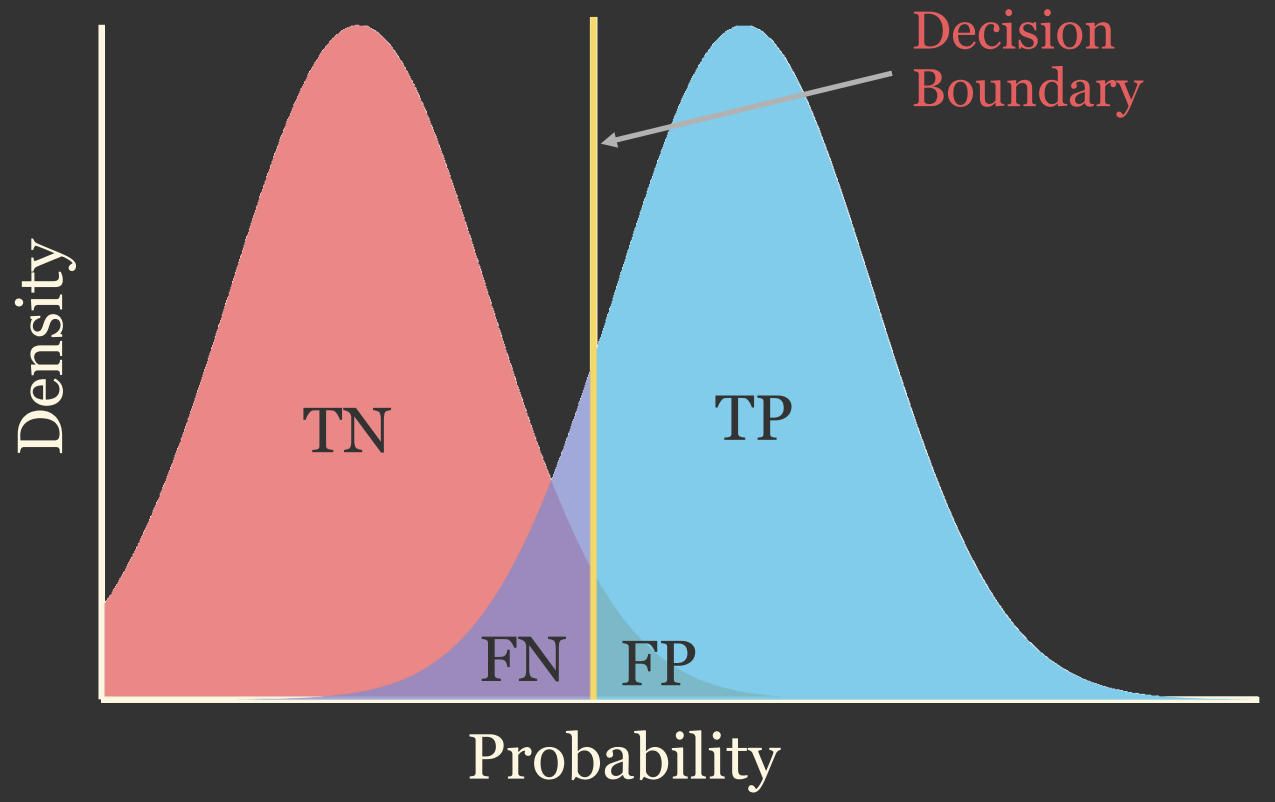
- Used for model comparisons
- Based in information theory
- Considers goodness of fit
- Penalizes for increased number of predictor variables/model complexity
- Larger value is better

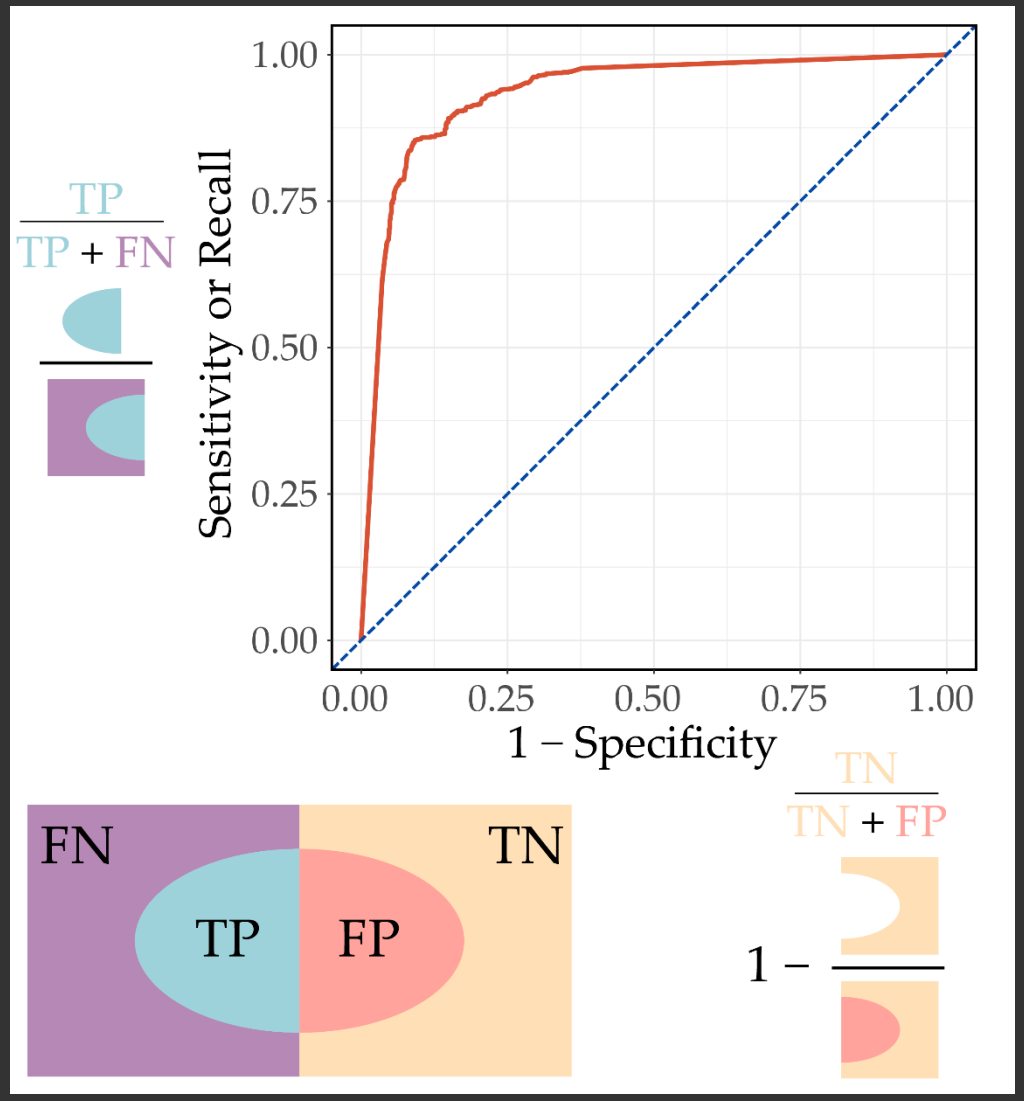
Bayesian Information Criterion (BIC)

- Based on Bayes' Theorem
- Considers goodness of fit
- Penalizes for increased number of predictor variables/model complexity
- Lower value is better

ROC Curves

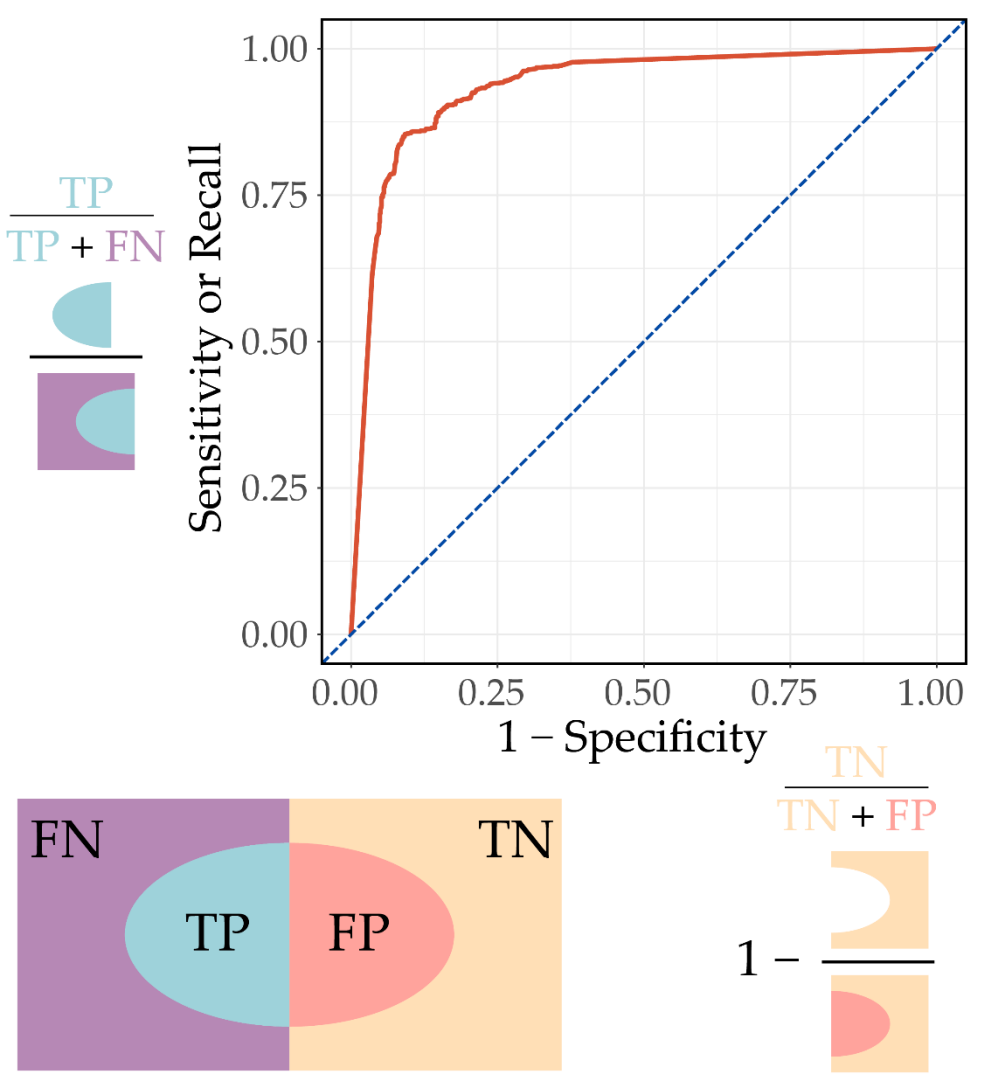
- True Positive Rate = fraction of cases predicted as true that were true
- False Positive Rate = fraction of cases predicted as true that were false





ROC Curves

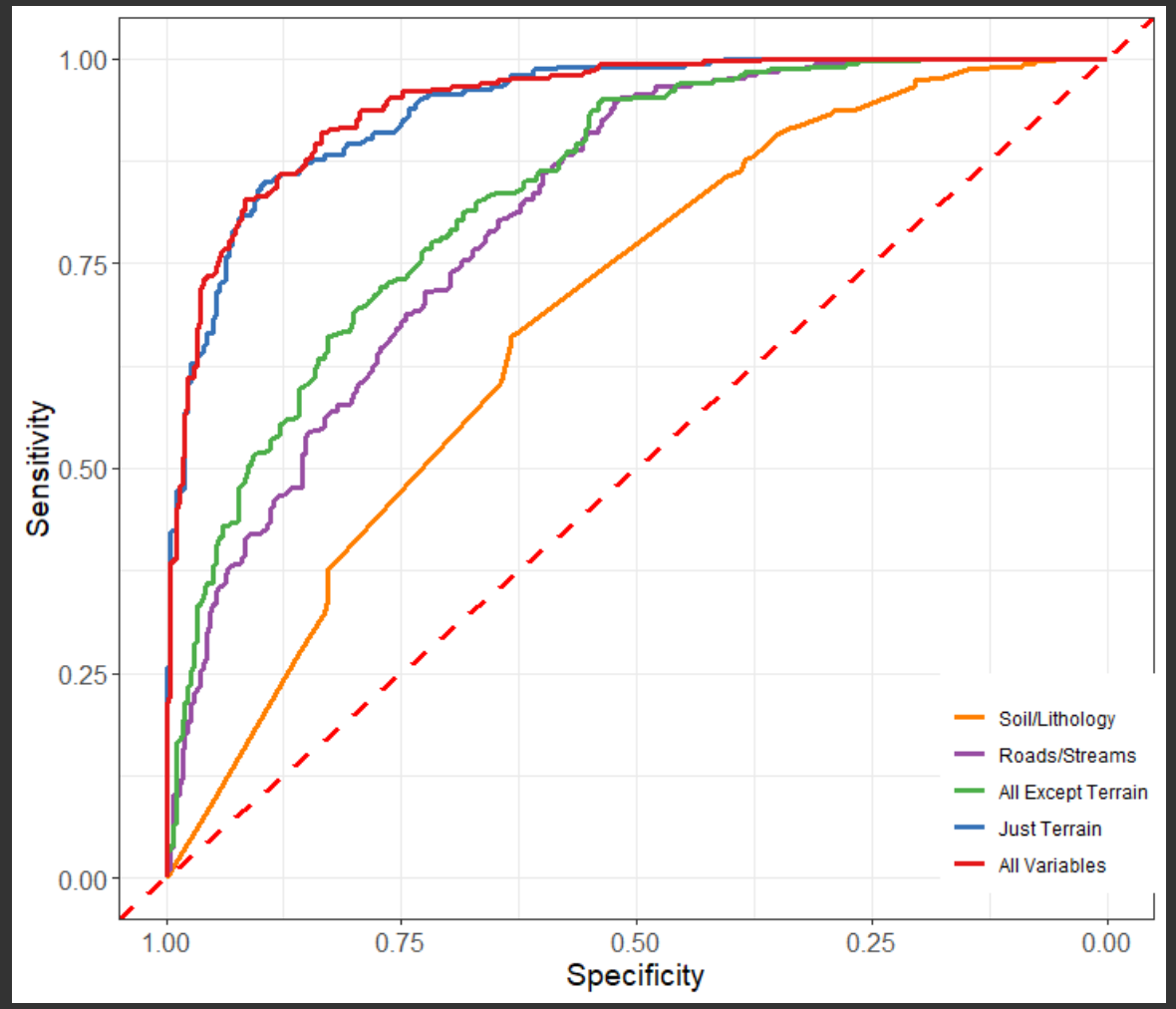
- Receiver Operating Characteristic (ROC) Curve
- Plots False Positive Rate and True Positive Rate at all decision/probability thresholds
- Sensitivity = True Positive Rate (Recall)
- Specificity = 1 – False Positive Rate



ROC Curves

- Area Under Curve (AUC) = area under ROC Curve (0-1)
- Larger suggest better models
- Can be used to compare models

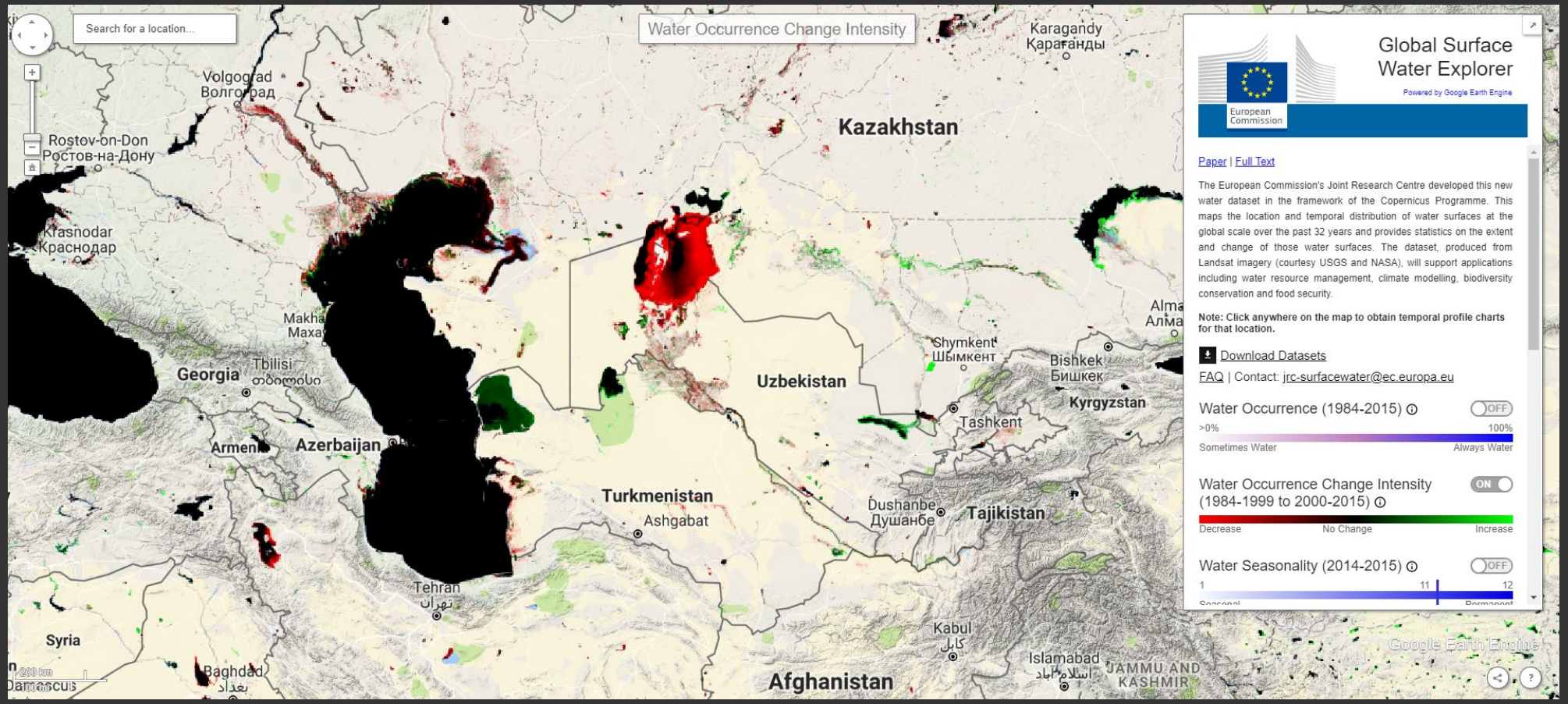
0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)



Modeling Limitations

1. World is inherently complex
2. Data are not perfect
3. There are always errors in predictions
4. Spatial heterogeneity
5. Spatial aggregation
6. MAUP

Future of Modeling



https://global-surface-water.appspot.com/

https://www.youtube.com/watch?v=KqyKXUMPHg8

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.