University of North Carolina in Chapel Hill

Gas Station Density Across Counties in North Carolina (2020)

S. Mouradkhanian, M. Dethlefs, E. Curley, J. Alvarez

Prof. Andrii Babii | Econ 573 | Spring 2025

Research Topic

1. Research Question

Using 2020 county-level data, what socioeconomic, demographic, and infrastructural factors most strongly predict gas station density across counties in North Carolina?

2. Research Relevance

- Improve transportation accessibility
- Inform policy decisions regarding infrastructure and urban development

Literary Review Overview

Estelaji et al. (2023)

- Model: GIS and Weighted Linear Combination(WLC) models
- Outcome: Gas Station Location in Tehran, Iran
- Used: Socioeconomic and geospatial variables

Chen(2020)

- Model: Random forest and regression models
- Outcome: EV Charger Locations
- Used: County-level socioeconomic variables

Ende (2021)

- Model: T-tests and CART Models
- Outcome: Density of branded gas stations
- Used: county-level wealth indicators

Data Collection and Overview

Main Data Sources

2020

US Census Bureau

Predictors:

1. Population Statistics
2. Educational Characteristics
3. Average Commuting Time
4. Economic Indicators

2020

NC Department of Agriculture & Consumer Services

Response:

1. Location of Gas Stations throughout North Carolina

2021

NC Department of Transportation

Predictors:

1. County-specific highway and road mileages

2019 - 2023

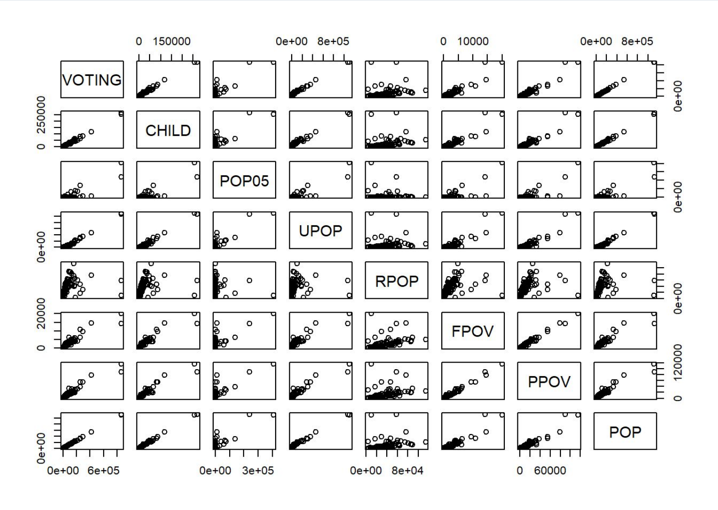
Census Bureau & American Community Survey

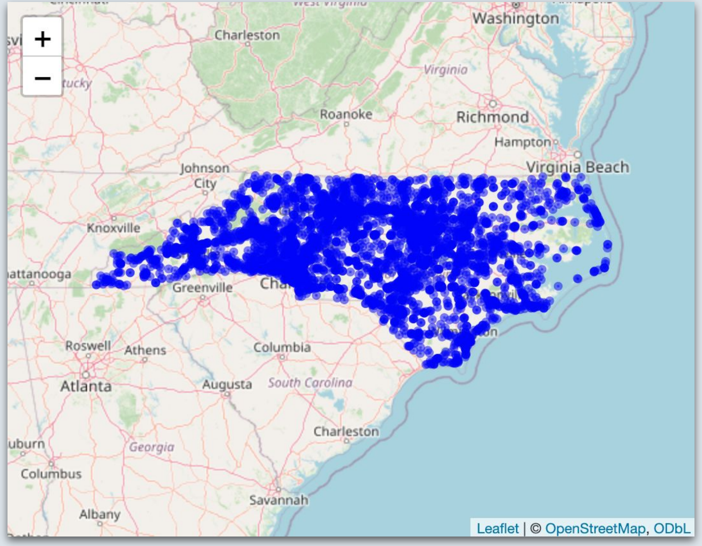
Predictors:

1. General Poverty Statistics (Individual and Household)

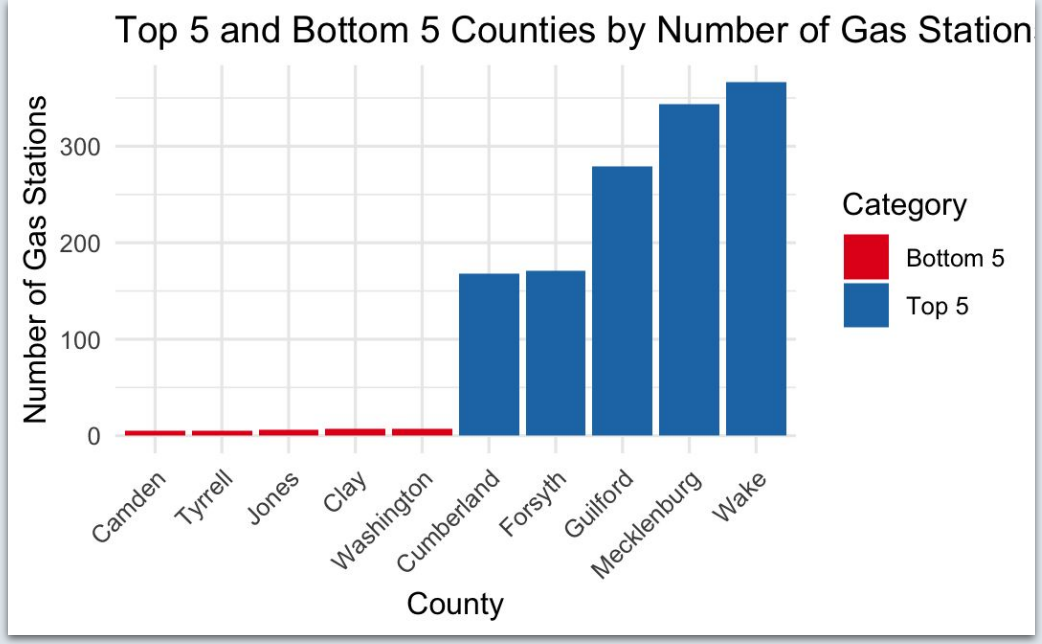
Collinearity Concerns and other data limitations

- Example of correlated predictors: Individuals below poverty & population
- Temporal mismatch in data sources
- Sample Size





Washington
Charleston
Virginia
Roanoke
Richmond
Hampton
Johnson City
Virginia Beach
Knoxville
Greenville
Roswell
Athens
Columbia
Atlanta
Augusta
South Carolina
Charleston
Georgia
Columbus
Savannah
Albany
Leaflet | © OpenStreetMap, ODbL

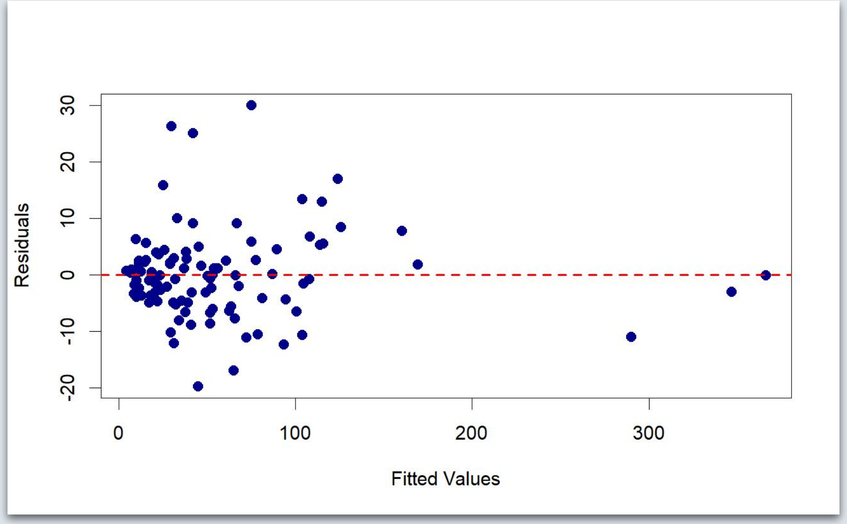


Top 5 and Bottom 5 Counties by Number of Gas Station
Number of Gas Stations
0
100
200
300
Camden
Tyrrell
Jones
Clay
Washington
Cumberland
Forsyth
Guilford
Mecklenburg
Wake
County
Category
Bottom 5
Top 5

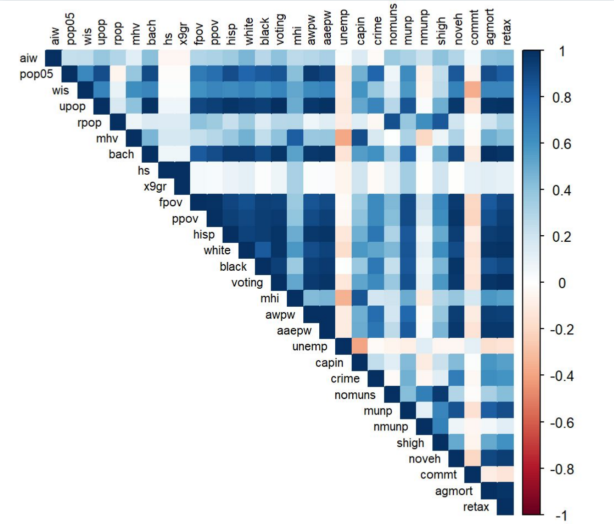
Baseline Model & Diagnostics

Ordinary Least Squares

- *pop*, *other*, *child*, *prpval* automatically dropped due to perfect multicollinearity
- Adjusted R^2 = 0.976 (but likely overfitting)
- Higher residuals in urban areas → Heteroscedasticity
- Many insignificant variables despite high R^2
- Unstable coefficients, low t-values → Multicollinearity



Residual variance increases with fitted values — poor fit in urban counties.



Severe predictor correlation → inflated standard errors, instability.

Multicollinearity Diagnostic with VIF

- Variance Inflation Factors (VIF) values well above the conventional threshold of 10 confirm severe multicollinearity
- Multicollinearity inflates standard errors and undermines coefficient interpretability

x9grr	**hs**	**voting**	upop	aaepw	bach	agmort	awpw	retax	white
179,703	**179,126**	**10,693**	7,411	1,385	1,336	1,321	1,034	911	785

shigh	ppov	noveh	black	fpov	nomuns	rpop	hisp	pop05
358	329	308	241	222	170	142	118	83

munp	crime	mhv	capin	nmunp	mhi	wis	aiw	commt	unemp
72	38	12	10	10	9.2	7.1	2.5	2.2	1.7

Stepwise Variable Reduction

- Removing the most collinear (and insignificant) variables at each step.
- 10-fold cross-validation on each iteration

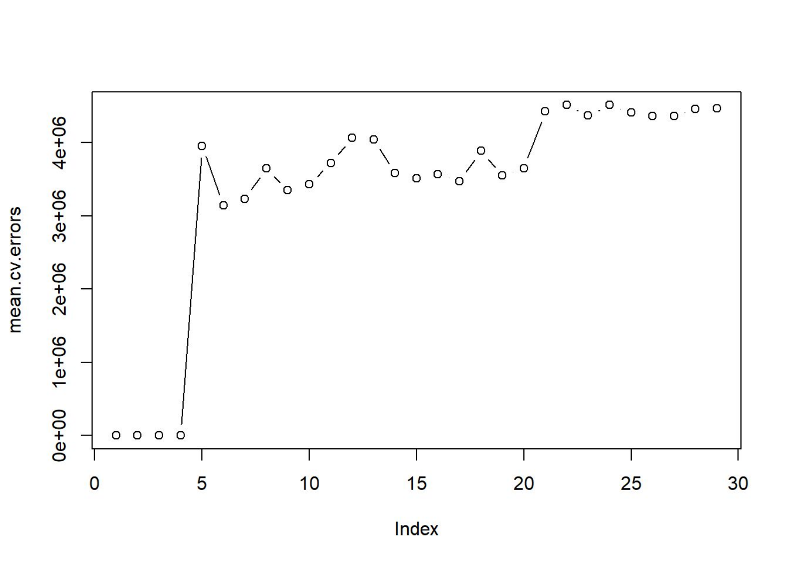
Model	Removed	No. of Predictors	Adjusted R^2	Residual SE	RMSE
Initial	-	33	0.976	9.474	2,088.88
w/o perfect multicollinearity	pop, child, other ppval	29	0.976	9.474	4,079.62
1st iteration	hs, x9gr	27	0.9766	9.365	39.94
2nd iteration	**voting, upop, bach**	**24**	**0.977**	**9.266**	**23.00**
3rd iteration	Awpw, retax, shigh, ppov, …	14	0.9631	11.75	52.47
4th iteration	Aaepw, capin, munp, mhi, mhv	9	0.8961	19.72	42.75
5th iteration	pop05	8	0.8956	19.76	47.13

Variable Selection Techniques

Best Subset & Stepwise Selection

- Highest Adjusted R^2 achieved with relatively large models (12-16)
- Model complexity penalties (BIC/Cp) favored smaller models (6-9)
- Repeated Variables across smaller models: <u>fpov</u>, <u>awpw</u>, <u>aaepw</u>, and <u>shigh</u>.

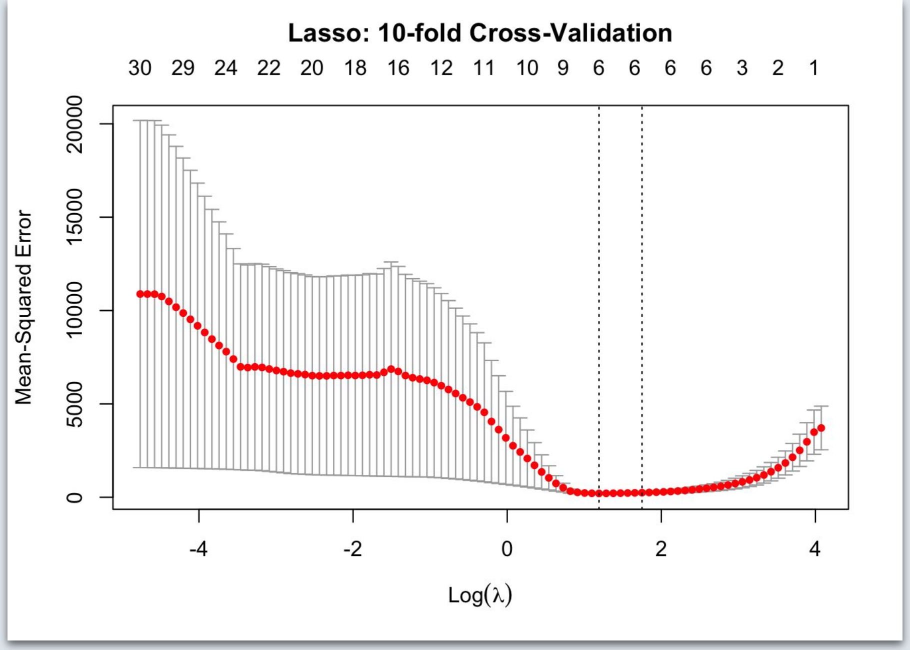
Selection Method	No. of Predictors	Adjusted R^2	BIC	Cp
Best Subset	8 (BIC/Cp) / 12 (Adj R^2)	0.9793 (12)	-353.13	-2.23
Forward Stepwise	9 (BIC) / 11 (Cp) / 16 (Adj R^2)	0.9789 (16)	-342.60 (9)	2.71 (11)
Backward Stepwise	6 (BIC) / 9 (Cp) / 15 (Adj R^2)	0.9792 (15)	-349.86 (6)	0.47 (9)
Cross-Validated Best Subset	4	0.9743	-347.3519	11.61



Test error decreases sharply up to a 4-variable model: <u>fpov</u>, <u>awpw</u>, <u>aaepw</u>, and <u>shigh</u>

Regularization Techniques

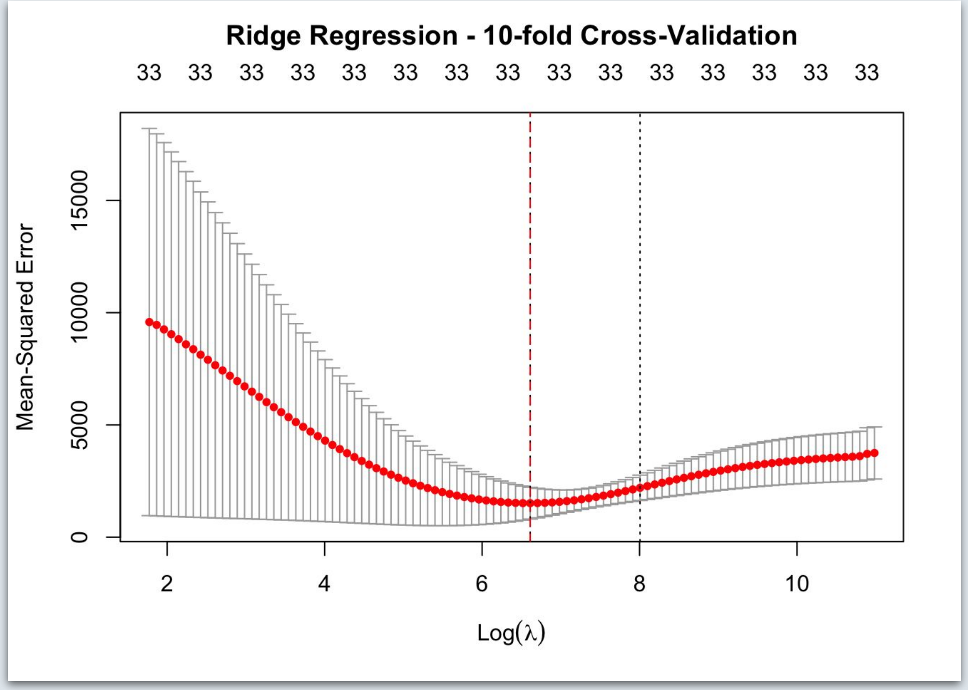
LASSO



Statistics	Original LASSO	With Interaction Terms	With Polynomials
Lambda(λ)	3.29	3.29	3.29
RMSE	11.16	11.30	11.14
R^2	0.963	0.963	0.966

6 Key Predictors (Original Model)	Coefficient
Families below Poverty	10.048804
People Below Poverty Line	24.328007
White Population Percentage	12.166976
State Municipal Primary Road Length	2.261565
State Highway Length	6.594529
Real Estate Taxes on Mortgaged properties	3.607343
Intercept	58.04

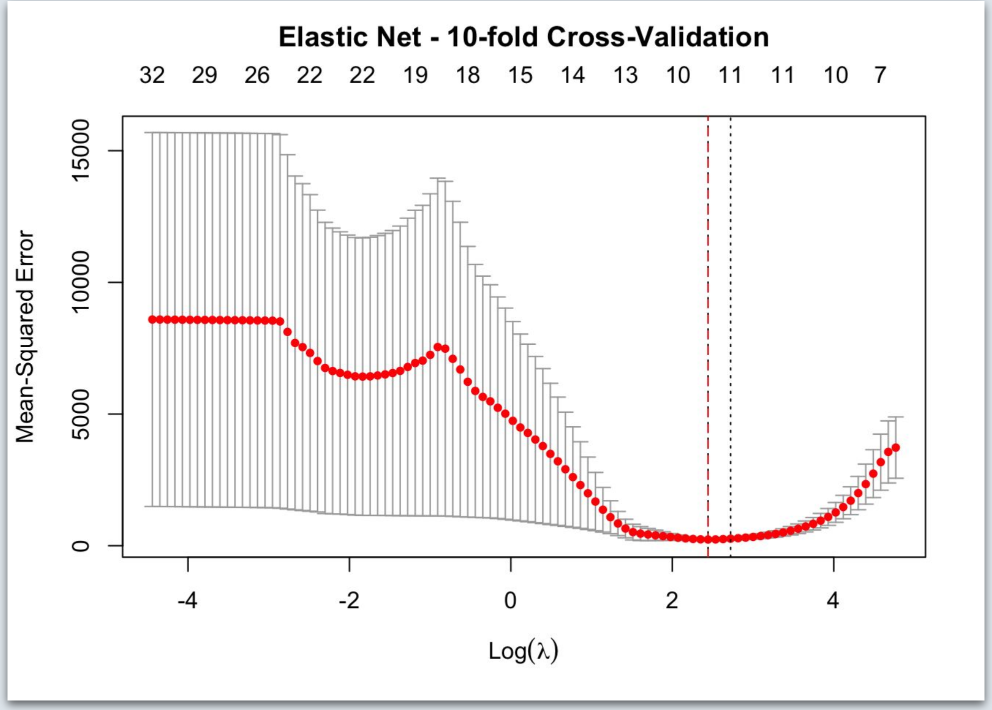
Ridge



Lambda (λ)	RMSE	R^2
743.07	28.41	0.782

- All variables retained
- Lower R^2 than LASSO (0.996)
- No further modifications due to lower R^2 and higher RMSE value
- 20 variables accounted with coefficients above 1 and below -1
- Limited coefficient value range (-1, 2) → Intercept: 58.04

Elastic Net



Lambda (λ)	RMSE	R^2
11.51	13.04	0.954

11 Key Predictors	Coefficient
Families Below Poverty	10.9581189
People Below Poverty	10.7380196
Total Population	2.6218477
White Population	5.6876633
Voting Age Population	3.3234355
Childhood Population	0.5649902
State Municipal Primary Road Length	4.0647695
State Highway Length	6.7714996
House Without Motor Vehicles	4.0647695
Real Estate Taxes	5.5029393
Other Race Population	2.4568653
Intercept	58.04

Alpha = 0.5

R^2 between Ridge and Lasso

Higher RMSE than LASSO

**Greater emphasis on demographic variables as seen by selected Predictors

Commonalities with Lasso:

Families Below Poverty

People Below Poverty

State Highway length

Real Estate Taxes

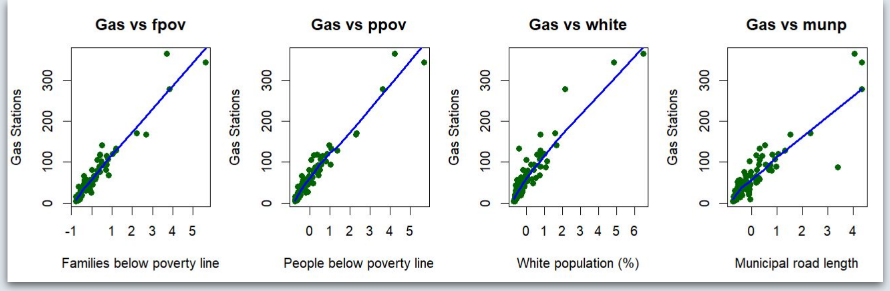
State Municipal Primary Road Length

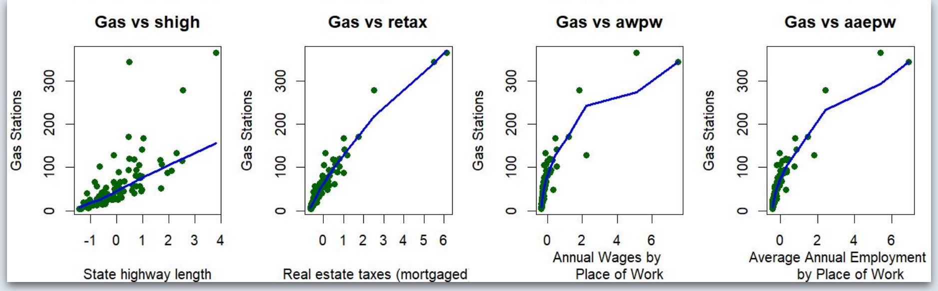
Model Performance Comparison

OLS Results

- OLS on the LASSO and CV Best Subset models → CV Best subset shows superior predictive accuracy

Model	No. of predictors	Adjusted R^2	Residual SE	AIC	RMSE
VIF-Selected	24	0.977	9.266	-	23.00
LASSO	6	0.9681	10.92	770.7408	13.85
CV Best Subset	**4**	**0.9743**	**9.798**	**747.092**	**10.49**



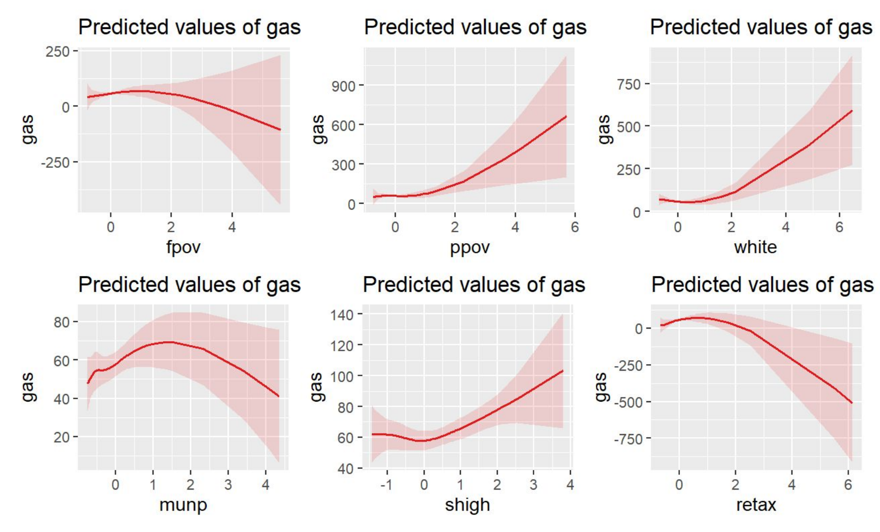


- High-leverage urban counties with high gas station counts → Non-linearities
- awpw, retax, and white show curved or nonlinear relationships, especially at upper values.

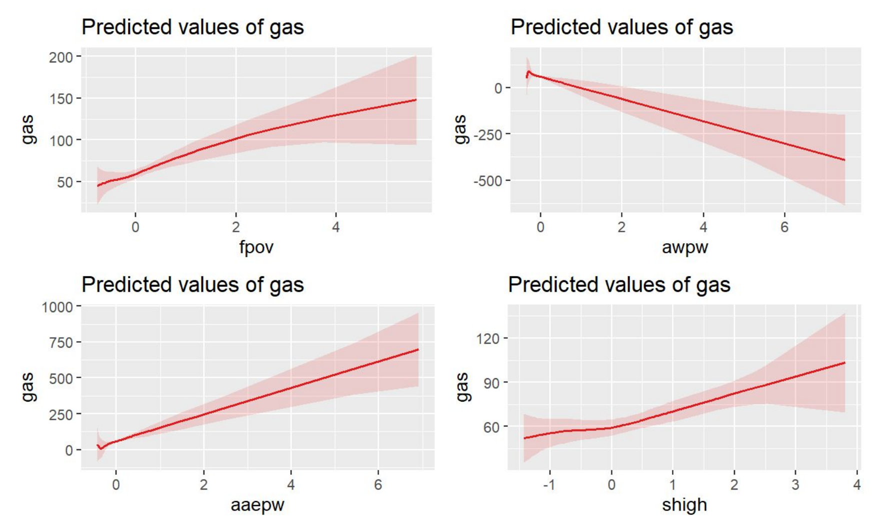
Non-Linear Modeling Approaches

Natural Splines

- Fit Natural Splines with 4 degrees of freedom on LASSO and Best Subset models
- Significant non-linear effects detected for: fpov, awpw, aaepw, shigh, retax
- Spline terms often significant at higher degrees (2–4)
- Distortion driven by high-leverage counties at extreme values



(Lasso)



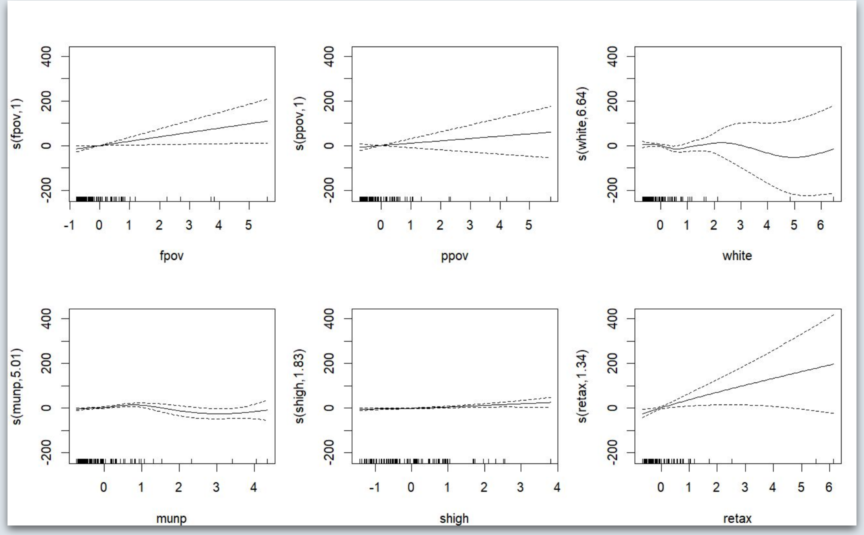
(Best Subset)

- Improved in-sample fit, but slightly worse test error → risk of overfitting

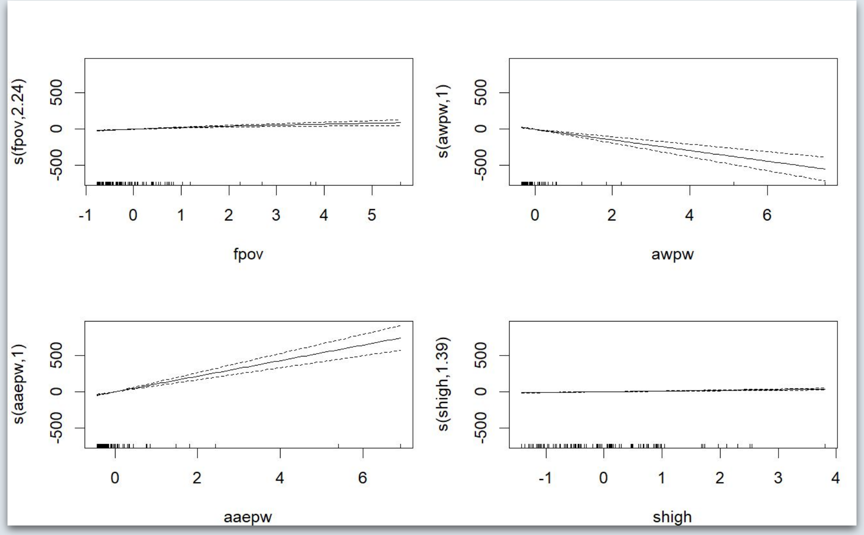
Model	Significant Nonlinear Terms	Adjusted R^2	Residual SE	AIC	RMSE
LASSO + Splines	retax, white, awpw	0.9714	10.33	774.12	14.69
CV Best Subset + Splines	**fpov, awpw, aaepw, shigh, retax**	**0.9753**	**9.61**	**753.61**	**13.72**

General Additive Models (GAM)

- GAMs capture both linear & non-linear effects flexibly
- Lasso → non-linear patterns primarily in white population, munp, and retax with most effects staying close to linear.
- Best Subset → simple model captures most patterns effectively with only modest non-linearity detected.



(Lasso)



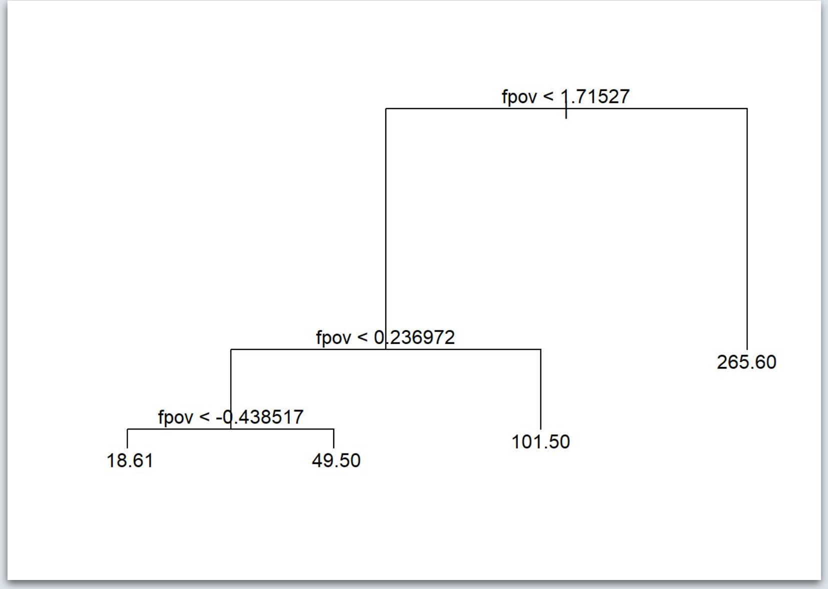
(Best Subset)

- ppov was redundant with fpov → removed for parsimony (Lasso)
- Best Subset GAM most robust model overall

Model	Predictors	Adjusted R^2	Deviance Explained	AIC	RMSE
LASSO GAM	fpov, ppov, white, munp, shigh, retax	0.978	98.2%	742.29	25.19
LASSO GAM (without ppov)	fpov, white, munp, shigh, retax	0.978	98.2%	741.32	19.75
CV Best Subset GAM	**fpov, awpw, aaepw, shigh**	**0.9753**	**97.7%**	**753.61**	**10.76**

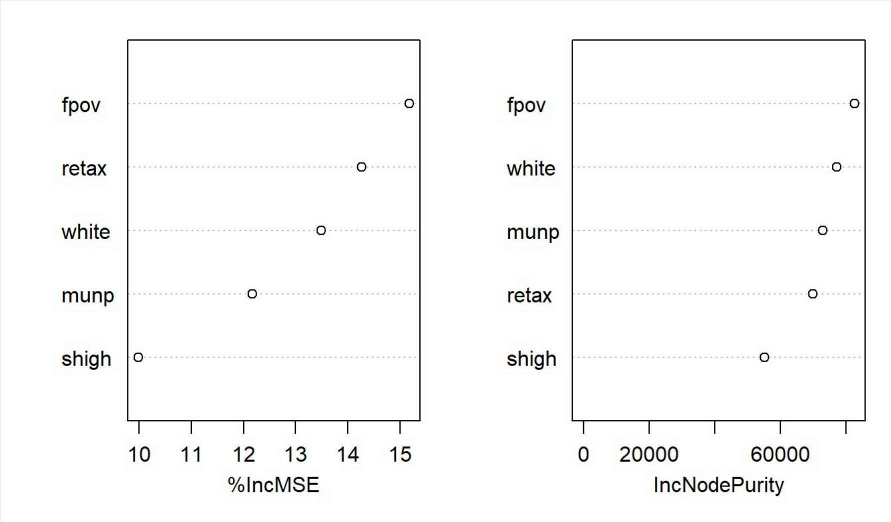
Regression Trees

- Cross-validated RMSE: 22.56
- fpov revealed as the sole splitting variable
- Tree identifies four poverty-based groups with distinct gas station averages
- Higher poverty levels → more gas stations

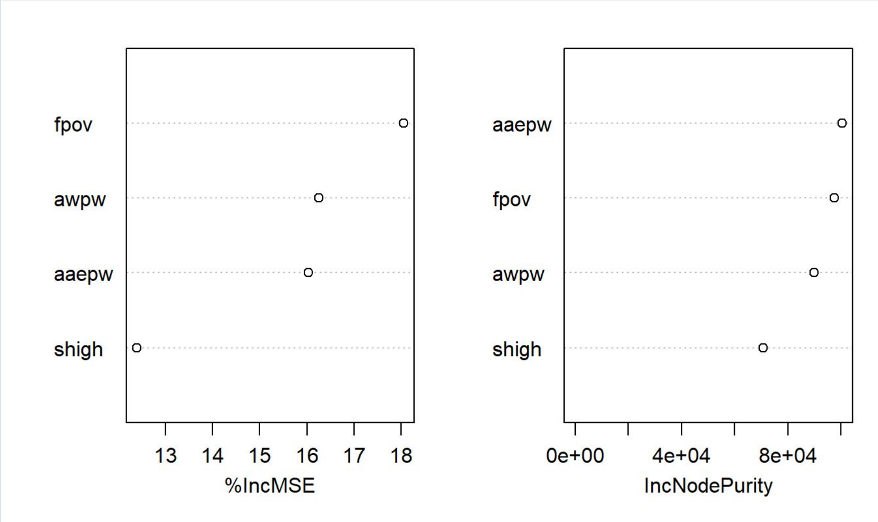


Random Forests

- Captures complex interactions and non-linearities
- Competitive performance (RMSE = 8.85, R^2 = 89.9%)
- Variable importance confirms fpov as dominant predictor
- Other key predictors: retax, white, munp, aaepw



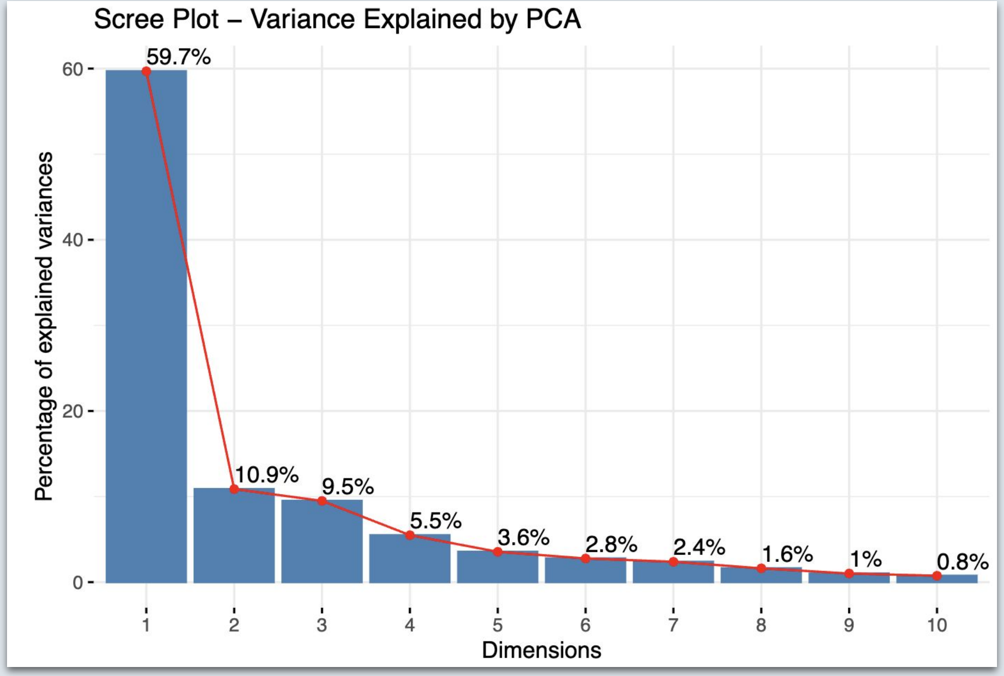
(Lasso)



(Best Subset)

Dimensionality Reduction Analysis

Principal Component Analysis (PCA)



Number of PC's	R^2
2	0.9439063
3	0.9434983
4	0.9454956
5	0.95089885

Best Model: 2 PC's

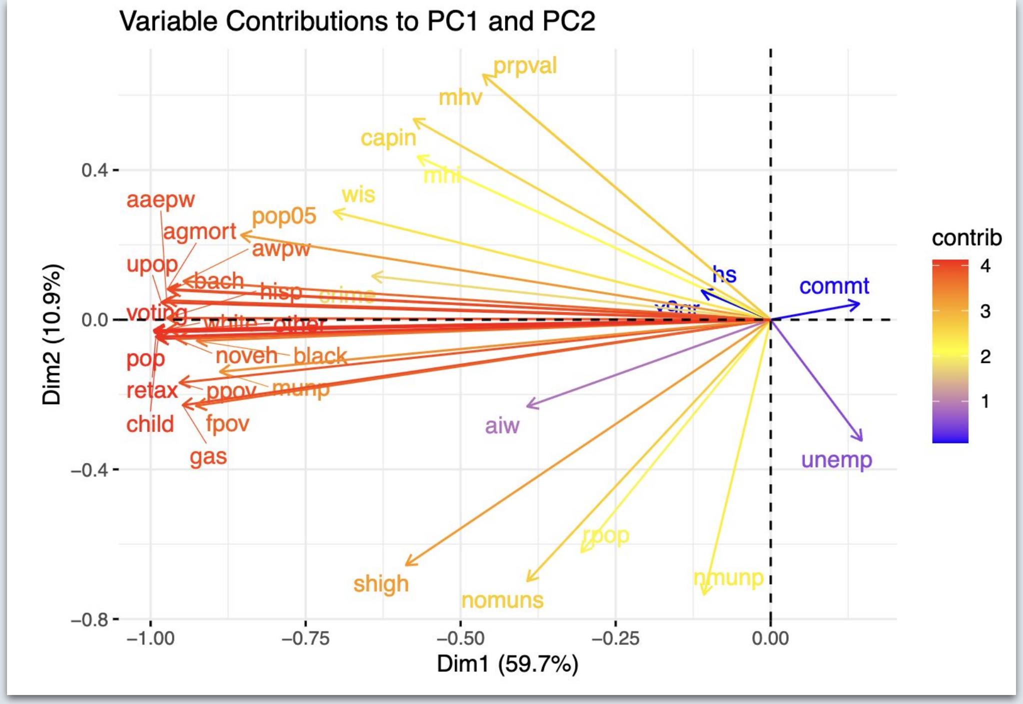
- Little variation from 2 to 5 PC's
- Best model with least complexity

Highest R^2: 0.9509 with 5 PC's

**5 PC's account for 85.561% of total variance

- PC1 Accounts for 59.7% of variance alone

PCA (cont...)



*PCA was important in reinforcing insight of regression approaches.

- Highlights variations due to population size, wealth, and rural/urban characteristic
- Any additional PC's would complicate and define too many that would result in overfitting

PC1	**PC2**
Captures factors pertaining to general size of county	Captures contrast between urban and rural classified counties
Uniform loading across all variables between -0.25 to - 0.15	<u>Highest Loadings</u> State Non-Municipal Secondary Paved Total Miles: 0.346 Total State highway Miles: 0.346 Rural Population: 0.325 Unemployment Rate: 0.170
Counties with more negative association (i.e. Mecklenburg) were large in terms of population and economic activity.	<u>Lowest Loadings:</u> Median House Value: -0.279 Per Capita Money Income: -0.279 Median Household Income: -0.227

Final Thoughts

- GAM applied to the Best Subset variables provided the best overall balance, maintaining strong predictive accuracy, interpretability, and flexibility to capture non-linear effects.
- Importantly, across all models, the same set of key variables kept showing up: FPOV, AWPW, AAEPW, and SHIGH — consistently driving gas station density across North Carolina counties
- Future research may benefit from stratifying counties by urban-rural classification or incorporating more granular, spatially explicit data to further improve forecasting across diverse regions.