Heirs’ property research brainstorm

This text discusses an econometric estimation to understand how land tenure, specifically heirs’ property, affects forest health in North Carolina. The study uses land parcel data on ownership and property characteristics, including an indicator for heirs’ property. High-resolution NDVI data and forest inventory data are also available. The NDVI data covers the entire area at a 1-meter resolution, whereas the forest inventory data is sampled based on plots. Although the forest inventory provides a richer set of forest characteristics, it is sparse relative to the known location of heirs’ properties.

**Data Structure**

We have:

1. Land parcel data with ownership structure (including heirs property indicator)
2. High-resolution NDVI data (1m resolution, wall-to-wall coverage)
3. Forest inventory data (plot-based, sparse but rich in characteristics)

**Potential Estimation Approaches**

**1. Direct NDVI Analysis**

Since the NDVI data has complete coverage, we could directly compare NDVI values on heirs properties versus other ownership types. This would give us a simple measure of vegetation density/health.

**2. Two-Stage Approach**

Given the sparsity of forest inventory data relative to heirs’ properties, we might consider:

* First stage: Use Forest inventory plots to establish relationships between NDVI and more detailed forest health metrics
* Second stage: Apply these relationships to predict detailed forest health for all properties, including heirs’ properties

**3. Matching Methods**

To address potential selection bias (since heirs’ properties likely aren't randomly distributed):

* Propensity score matching to compare heirs’ properties with similar non-heirs’ properties
* Nearest neighbor matching based on property characteristics, geographic location, etc.

**4. Spatial Econometric Models**

Since forest health likely exhibits spatial autocorrelation:

* Spatial lag models to account for spillover effects
* Spatial error models to address spatially correlated unobservables

**Important Controls to Consider**

* Property size
* Time since acquisition/inheritance
* Adjacent land uses
* Topography and soil characteristics
* Historical land use
* Distance to markets/urban areas
* Socioeconomic characteristics of owners

**Potential Challenges**

* Endogeneity concerns (properties becoming heirs’ properties might have preexisting differences)
* Missing data or measurement error in heirs’ property identification

After intersecting the NDVI data with all the parcels in our dataset, including both heirs and non-heirs, I found that the NDVI of heirs’ properties is approximately 14% higher compared to non-heirs’ properties. To assist in filtering out urban and coastal areas, I categorized the counties in North Carolina into two regions (both rural): an eastern region and a western region.

**Refining Our Analysis**

**Regional Differences**

* Does the 14% difference hold consistently across both eastern and western regions?
* These regions likely have different forest types and topography which could influence NDVI values

**Control Variables**

To ensure the observed difference is truly due to heirs’ property status:

* Property size (heirs’ properties might be systematically larger/smaller)
* Distance to roads/development
* Topographic features (slope, aspect, elevation)
* Parcel attributes
  + Owner occupied
  + Market value
  + Others?

**Statistical Approaches**

1. **Multiple Regression**:

Where Heirs is your binary indicator, and Region distinguishes eastern/western areas

1. **Propensity Score Matching**:
   * Calculate propensity scores based on observable characteristics
   * Match heirs’ properties to similar non-heirs’ properties
   * Compare NDVI differences within matched pairs
2. **Spatial Lag Models**:
   * Account for spatial autocorrelation in NDVI values
   * Add a spatially lagged dependent variable to your model

**Exploring Mechanisms**

If you want to understand why heirs’ properties have higher NDVI:

1. **Forest Structure Analysis**:
   * Use forest inventory plots that fall within heirs/non-heirs’ properties
   * Compare metrics like tree density, size distribution, species composition
2. **Management Practices**:
   * If available, collect data on timber harvesting history, conservation programs
   * These could be mediating variables between ownership structure and forest condition

Both directions you're considering are excellent approaches for your analysis. Let me outline how to pursue each one.

**Approach 1: Multiple Regression**

For the multiple regression approach, I'd suggest the following steps:

**Step 1: Data Preparation**

* Aggregate NDVI values at the parcel level (mean, median, percentiles) [DONE]
* Create dummy variables for region (eastern=1, western=0) [DONE]
* Generate interaction terms (heirs × region)
* Prepare control variables (property size, distance metrics, etc.)

**Step 2: Base Model Specification**

This gives you the main effect of heirs’ property status controlling for region and other factors.

**Step 3: Interaction Model**

This allows the effect of heirs’ property to differ between eastern and western regions.

**Step 4: Robustness Checks**

* Try different aggregations of NDVI (mean vs. median)
* Include spatial fixed effects at different levels (county, township)
* Test for spatial autocorrelation and consider spatial error models if present

**Approach 2: Two-Stage Analysis**

**Stage 1: Establish NDVI-Forest Health Relationship**

1. Identify forest inventory plots with corresponding NDVI data
2. Estimate relationships between NDVI and forest metrics:

Where ForestMetric could be basal area, species diversity, carbon storage, etc.

**Stage 2: Predict Forest Health for All Parcels**

1. Use coefficients from Stage 1 to predict forest metrics for all parcels:
2. Then estimate the heir’s property effect:

**Implementation Strategy**

For Approach 1, you could start with:

Pseudocode for Multiple Regression Approach

```python

import statsmodels.api as sm

Prepare the data

X = df[['heirs\_property', 'eastern\_region', 'heirs\_eastern\_interaction',

'property\_size', 'road\_distance', 'elevation', ...]]

X = sm.add\_constant(X) # Add a constant term to the model

y = df['mean\_ndvi']

Perform the regression analysis

model = sm.OLS(y, X).fit(cov\_type='HC3') # Use robust standard errors

print(model.summary())

```

For Approach 2:

Stage 1: Forest Inventory Plots Analysis

First, we prepare the data for forest inventory plots by selecting relevant features such as 'ndvi', 'elevation', and 'slope'. Additionally, we add a constant to the dataset:

```python

X\_stage1 = forest\_plots[['ndvi', 'elevation', 'slope', ...]]

X\_stage1 = sm.add\_constant(X\_stage1)

y\_stage1 = forest\_plots['basal\_area'] # or other forest metric

```

Next, we fit an Ordinary Least Squares (OLS) regression model to the dataset:

```python

model\_stage1 = sm.OLS(y\_stage1, X\_stage1).fit()

```

Stage 2: Prediction for All Parcels

We proceed by using the fitted model to predict basal area for all parcels. Relevant features and a constant are prepared similarly:

```python

X\_all = all\_parcels[['ndvi', 'elevation', 'slope', ...]]

X\_all = sm.add\_constant(X\_all)

all\_parcels['predicted\_basal\_area'] = model\_stage1.predict(X\_all)

```

Stage 3: Analyzing the Effect of Heirs Property

Finally, we analyze the effect of heirs property by preparing a new dataset with additional features such as 'heirs\_property' and 'eastern\_region'. We then fit another OLS regression model:

```python

X\_stage2 = all\_parcels[['heirs\_property', 'eastern\_region', ...]]

X\_stage2 = sm.add\_constant(X\_stage2)

y\_stage2 = all\_parcels['predicted\_basal\_area']

model\_stage2 = sm.OLS(y\_stage2, X\_stage2).fit()

print(model\_stage2.summary())

```

I'd be happy to provide the equivalent R code for both approaches. Here's how you could implement them:

**Approach 1: Multiple Regression in R**

# Load necessary packages

library(tidyverse)

library(sandwich) # For robust standard errors

library(lmtest) # For coeftest function

# Assuming your data is in a dataframe called 'parcels\_data'

# with columns like 'mean\_ndvi', 'heirs\_property', 'eastern\_region', etc.

# Create interaction term

parcels\_data <- parcels\_data %>%

mutate(heirs\_eastern\_interaction = heirs\_property \* eastern\_region)

# Base model

base\_model <- lm(mean\_ndvi ~ heirs\_property + eastern\_region +

property\_size + road\_distance + elevation + slope,

data = parcels\_data)

# Print summary with robust standard errors

coeftest(base\_model, vcov = vcovHC(base\_model, type = "HC3"))

# Interaction model

interaction\_model <- lm(mean\_ndvi ~ heirs\_property + eastern\_region +

heirs\_eastern\_interaction + property\_size +

road\_distance + elevation + slope,

data = parcels\_data)

# Print summary with robust standard errors

coeftest(interaction\_model, vcov = vcovHC(interaction\_model, type = "HC3"))

# Optional: For spatial considerations, consider using 'spdep' package

# library(spdep)

# Create neighbors list from your spatial data

# nb <- dnearneigh(coordinates(parcels\_spatial), 0, 5000) # 5km threshold

# Then test for spatial autocorrelation and implement spatial models if needed

**Approach 2: Two-Stage Analysis in R**

# Stage 1: Establish NDVI-Forest Health Relationship using forest inventory plots

library(tidyverse)

library(caret) # For cross-validation

# Assuming 'forest\_plots' dataframe has both NDVI and forest metrics

stage1\_model <- lm(basal\_area ~ ndvi + elevation + slope + aspect,

data = forest\_plots)

summary(stage1\_model)

# Optional: Cross-validation

# set.seed(123)

# train\_control <- trainControl(method = "cv", number = 10)

# cv\_model <- train(basal\_area ~ ndvi + elevation + slope + aspect,

# data = forest\_plots,

# method = "lm",

# trControl = train\_control)

# print(cv\_model)

# Stage 2: Predict forest metrics for all parcels

# Assuming 'all\_parcels' has the same predictor variables as in stage1\_model

all\_parcels <- all\_parcels %>%

mutate(predicted\_basal\_area = predict(stage1\_model, newdata = all\_parcels))

# Analyze heirs property effect on predicted forest metric

stage2\_model <- lm(predicted\_basal\_area ~ heirs\_property + eastern\_region +

property\_size + road\_distance,

data = all\_parcels)

summary(stage2\_model)

# For spatial considerations

# library(spdep)

# Implement spatial regression if spatial autocorrelation is detected

**Additional R Packages to Consider**

Depending on your specific needs, you might also want to explore:

1. For matching methods:

library(MatchIt) # For propensity score matching

m.out <- matchit(heirs\_property ~ property\_size + elevation + road\_distance,

data = parcels\_data, method = "nearest")

matched\_data <- match.data(m.out)

1. For spatial models:

library(spatialreg) # For spatial regression models

# After creating a spatial weights matrix 'listw'

spatial\_lag\_model <- lagsarlm(mean\_ndvi ~ heirs\_property + eastern\_region +

property\_size, data = parcels\_data,

listw = listw)

1. For robust regression:

library(MASS)

robust\_model <- rlm(mean\_ndvi ~ heirs\_property + eastern\_region +

property\_size, data = parcels\_data)