Do Homeowners Care About Air Quality? Estimating the Effect of Poor Air Quality on Home Prices Using Wildfire Smoke Data

Model and Data

Tristan Misko

17 November 2021

1 Model and Design

Table 1: Effect of Wildfire Smoke on Housing Prices

	(1)	(2)
(Intercept)	183687.00***	73976.901
	(1100.8998)	(1113.0914)
n.score	−0.6631	
	(0.6412)	
treat.k5.t50		−1484.0340***
		(168.9167)
unemp	1114.7302***	1563.9206***
	(27.7190)	(48.1914)
County Fixed Effects	Yes	Yes
Time Fixed Effects	Yes	Yes
Observations	234600	45655
R^2	0.9804	0.98367

The main idea of the paper is to determine whether housing prices respond to changes to in air quality. The naive regression of housing prices on air quality may suffer from significant endogeneity. If homeowners respond to air quality, then it is likely that taste-based sorting occurs, producing a selection bias. To overcome endogeneity, we use wildfire

Table 2:

Statistic	N	Mean	St. Dev.	Min	Max
X	234,600	117,300.500	67,723.330	1	234,600
year	234,600	2,014.696	2.773	2,010	2,019
month	234,600	6.652	3.434	1	12
zhvi.score	234,600	135,041.500	79,504.160	22,372	1,646,548
SizeRank	234,600	1,328.630	787.017	1	3,071
STATEFP	234,600	32.497	13.459	10	55
COUNTYFP	234,600	114.463	117.166	1	840
n.light	234,600	2.748	4.662	0.000	30.714
n.medium	234,600	0.584	1.625	0.000	25.995
n.heavy	234,600	0.130	0.550	0	12
n.score	234,600	25.749	55.318	0.000	820.000
period	234,600	58.000	33.196	1	115
unemp	234,600	5.908	2.505	1.100	21.648
post43	234,600	0.626	0.484	0	1
post48	234,600	0.583	0.493	0	1
post55	234,600	0.522	0.500	0	1
post60	234,600	0.478	0.500	0	1
post67	234,600	0.417	0.493	0	1
m.s.pre43	234,600	28.083	15.841	5.602	61.858
m.s.post43	234,600	24.355	15.515	2.782	75.967
m.s.delta43	234,600	−3.729	10.237	−36.251	30.297
m.s.pch43	234,600	−0.059	0.403	−0.820	1.634
m.s.pre48	234,600	25.528	14.411	5.129	56.054
m.s.post48	234,600	25.907	16.633	2.911	81.352
m.s.delta48	234,600	0.378	10.704	−31.451	39.691
m.s.pch48	234,600	0.108	0.493	−0.796	2.159
m.s.pre55	234,600	26.099	14.801	4.814	60.081
m.s.post55	234,600	25.428	15.898	2.941	79.808
m.s.delta55	234,600	−0.671	8.164	−25.744	29.185
m.s.pch55	234,600	0.031	0.338	−0.754	1.295
m.s.pre60	234,600	24.754	14.237	4.637	57.425
m.s.post60	234,600	26.834	16.644	2.963	83.996
m.s.delta60	234,600	2.081	8.551	−22.197	34.476
m.s.pch60	234,600	0.156	0.387	−0.728	1.536
m.s.pre67	234,600	26.489	15.616	4.153	66.977
m.s.post67	234,600	24.716	14.460	3.395	74.525
m.s.delta67	234,600	−1.773	6.198	−19.957	17.780
m.s.pch67	234,600	−0.012	0.265	−0.683	0.992

smoke as a source of exogenous variation in air quality and exploit the heterogeneous increase in wildfire smoke since 2015 to obtain casual estimates.

2 Description of Data

The data take the form of panel data, with monthly observations at the county level of the number of days in each month in which the county is covered by wildfire smoke plumes, the mean air quality index (AQI) over the month, and the level of the Zillow housing price index in that month. I also have associated to each observation a set of controls for unemployment level and housing characteristics. The data range from June 2010 to July 2019 and contain counties outside of the geographic west of the United States (such counties may suffer potentially significant confounding from wildfires themselves). I am currently working on expanding the smoke dataset to include all months in 2019 and 2020.

3 Empirical Strategy

A county is categorized as treated if the difference between its post-2014 monthly mean smoke score and its pre-2014 monthly mean smoke score is greater than some threshold value α. Likewise, control counties are determined by finding those counties whose difference is less than some threshold β. This leaves a group of "boundary" counties which are neither treatment nor control counties. The coefficient of interest is δ, which we expect to be small and to have a negative sign. This basic model is given by

$$\text{price}_{c,t} = \delta \cdot \text{smoke}_{c,t} + D_c + T_t + \gamma \cdot \text{Unemp}_{c,g(t)} + \zeta \cdot \text{HC}_{c,t} + \epsilon_{c,t}.$$

Variable Descriptions:

- $\text{price}_{c,t}$ (*numeric variable*): The Zillow Home Value Index value, a smoothed indicator of housing prices in county d and time period t.
- $\text{smoke}_{c,t}$ (*dummy variable*): A time-dependent treatment variable determined from the smoke score, which a weighted sum of the number of light, medium and heavy smoke days over the month within each county. The dummy is one if we are in the

post treatment period (beginning 2015) and the county experiences a change in mean monthly smoke score above a threshold value across the pre- and post-treatment period.

- D_c (*dummy variable*): A set of dummy variables for the county fixed effects.
- T_t (*dummy variable*): A set of dummy variables for the time fixed effects.
- $\text{Unemp}_{c,g(t)}$ (*numeric variable*): The quarterly unemployment rate at the county level. Here, g denotes the mapping of months to quarters.
- $HC_{c,t}$ (*numeric variable*): A set of controls for the characteristics of the average home in a county

3.1 Further Models

One extension that I am still setting up is the instrumental variables model, which uses smoke treatment level instrument for Air Quality Index (AQI) in the first stage model, then estimates the causal effect of air quality on housing prices in the second stage model.

Since my treatment variable has differing dosage levels, I am looking into the literature on estimating Two-Way Fixed Effects models to employ on my data. For now, I will estimate some extensions of the basic model, I will use multiple dummies encoding different buckets of smoke score values. See the below for a discussion of robustness checks of these arbitrary threshold values.

3.2 Robustness Checks

There are a number of robustness checks that remain to be done for my paper. The first and most important will be parallel trends in ZHVI for the difference-in-differences estimation strategy. Checking that ZHVI trends do not differ significantly before treatment is crucial for the estimates to be unbiased. Robustness to perturbing the time of treatment activation will also be checked.

The basic model also depends on the selection of threshold values for the smoke score difference to determine inclusion in the treatment or control group. The regression estimates will be recorded in graphs in which these threshold values are varied continuously to

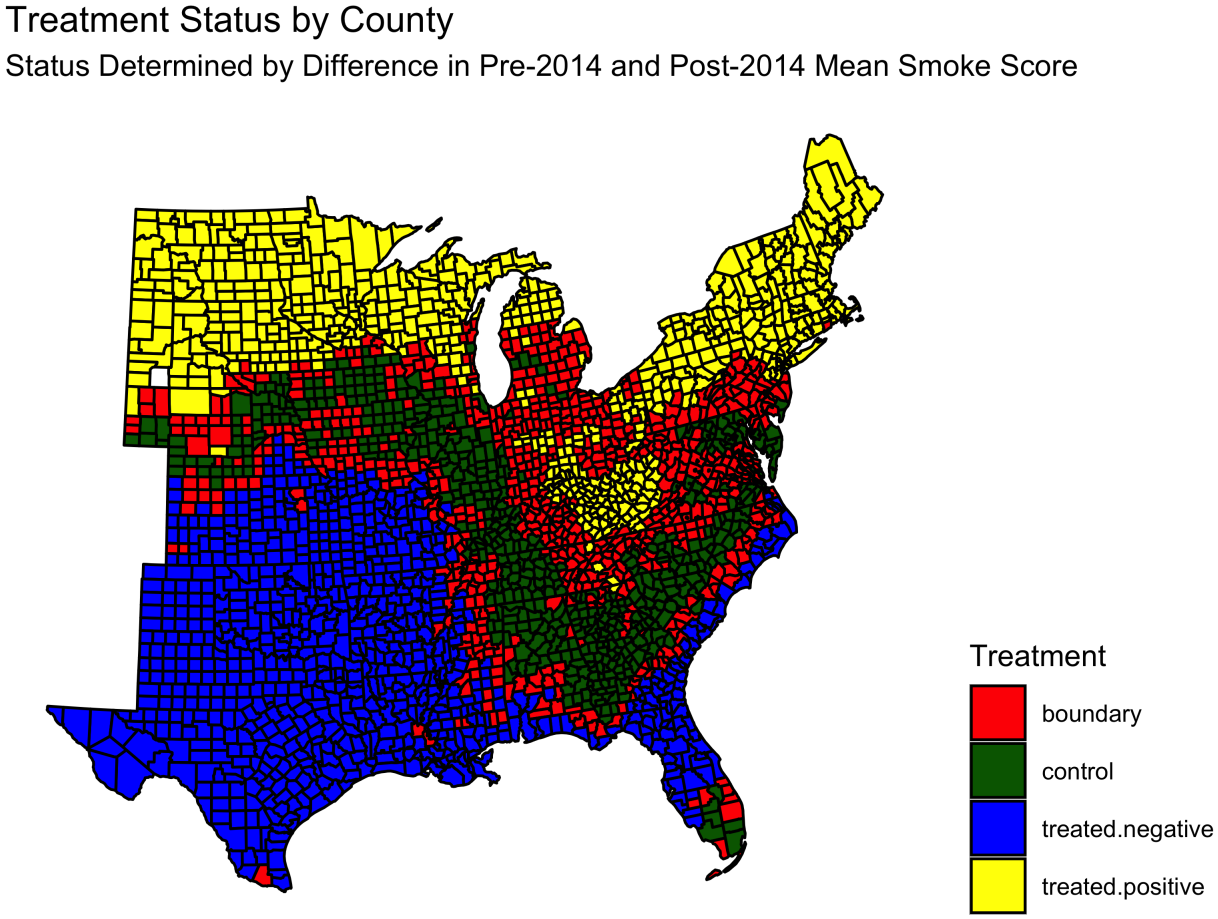
demonstrate the sensitivity of the results to the thresholds. In addition, we will reestimate all equations using different weights within the aggregation scheme of the smoke dataset to ensure that the final results are not sensitive to weighting decisions.

4 Data Summary Table

The summary statistics for the main data are displayed in the table below. The variable `zhvi.score` encodes housing prices. The variables `n.light`, `n.medium`, and `n.heavy` encode the number of smoke days in the month at a given smoke intensity. `n.score` records the smoke score. The control datasets are still being cleaned and processed, so they have not yet been matched to the main data frame, but I do have the .csv files on my machine.

Table 3: Summary Statistics of Main Variables

Statistic	N	Mean	St. Dev.	Min	Median	Max
zhvi.score	234,600	135,041.500	79,504.160	22,372	115,794	1,646,548
n.light	234,600	2.748	4.662	0.000	0.316	30.714
n.medium	234,600	0.584	1.625	0.000	0.000	25.995
n.heavy	234,600	0.130	0.550	0	0	12
n.score	234,600	25.749	55.318	0.000	1.671	820.000
period	234,600	58.000	33.196	1	58	115
post2014	234,600	0.522	0.500	0	1	1


Treatment Status by County
Status Determined by Difference in Pre-2014 and Post-2014 Mean Smoke Score
Treatment
boundary
control
treated.negative
treated.positive