A Problem with Distance Variables and Alternatives for their Use

Mid-Continent Regional Science Association 53rd Annual Conference

Thursday, June 8, 2023

Two uses of distance variables

- Close to store, school, park, work clearly affect home choice and price. Yet use of distance may be compromised in a fashion more serious than multicollinearity.
- Distances to, say, the Central Business District, is used to anchor a position in the study area.
  - The chief goal is to avoid omitted variable bias in parameter estimates of spatially distributed variables.
  - Estimate the direct influence of a landmark on an outcome of interest, such as a neighborhood park on home price
    - Crime, school quality, tree cover, employment center
- Critically most policy variables in the regional sciences are spatially distributed

Distance variables are NOT IDENTIFIED

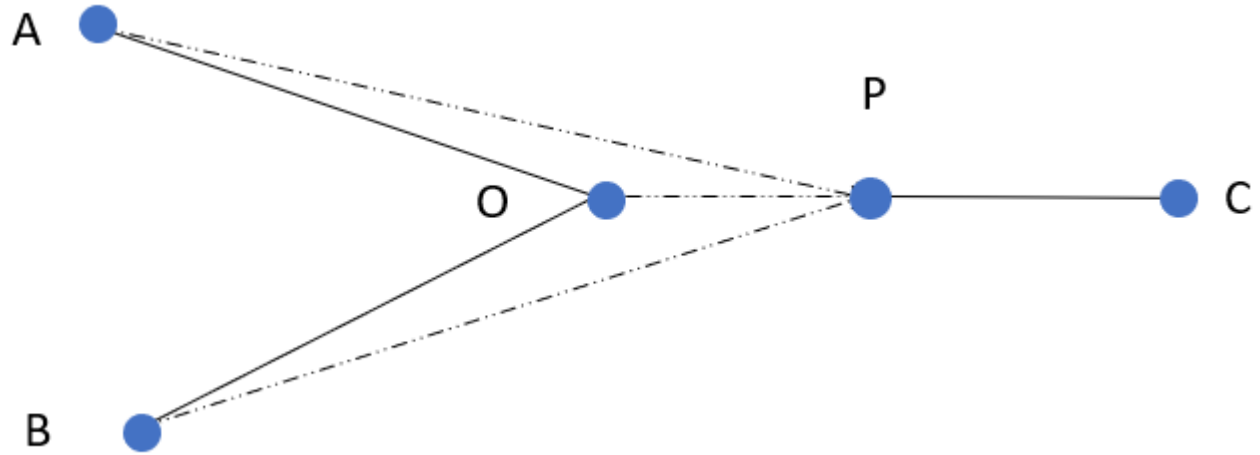
If a move to waste dump or to park, decreases of increases home value by $500:

- **How much is attributable to proximity to Park?**
- **How much is attributable to movement away from Dump?**

Home

It cannot be ascertained

Consider a multi-angled - an amenity examination



NO INDEPENDENT VARIATION !!!

No consistent way to measure effect using a distance variable

- Independent variables are only identified if there is true independent variation.

- Shifting distance to a site from observation to observation fully predetermines variation in all measured distance

Measured distance is directionless: Euclid's intention

$$\boldsymbol{Distance_{a\ to\ b} = \left[(long_a - long_b)^2 + (lat_a \quad lat_b)^2\right]^{1/2}}$$

What about its use as a control?

Use nested variables which are directional

$$\boldsymbol{Distance_{a\ to\ b}} = \left[(\boldsymbol{long_a} - \boldsymbol{long_b})^2 + (\boldsymbol{lat_a} - \boldsymbol{lat_b})^2\right]^{\boldsymbol{1/2}}$$

$$\boldsymbol{long_a} - \boldsymbol{long_b}$$
$$\boldsymbol{lat_a} - \boldsymbol{lat_b}$$
$$(\boldsymbol{long_a} - \boldsymbol{long_b})^2$$
$$(\boldsymbol{lat_a} - \boldsymbol{lat_b})^2$$

What about to anchor position?

- The obvious response if that distance is a ‘position free’ measure by design: Euclid established a pure measure.

- The square root (needed to employ the Pythagorean Theorem makes distance position free. That is why is works.

- So in our park and dump example, distance away from the home does not control for position; but confuses it. It assumes the park and dump have equal effect by distance to the house.

- To use this as an anchor – to capture all other partial effects by definition complicates the space – missing all estimates of all other spatially distributed variables

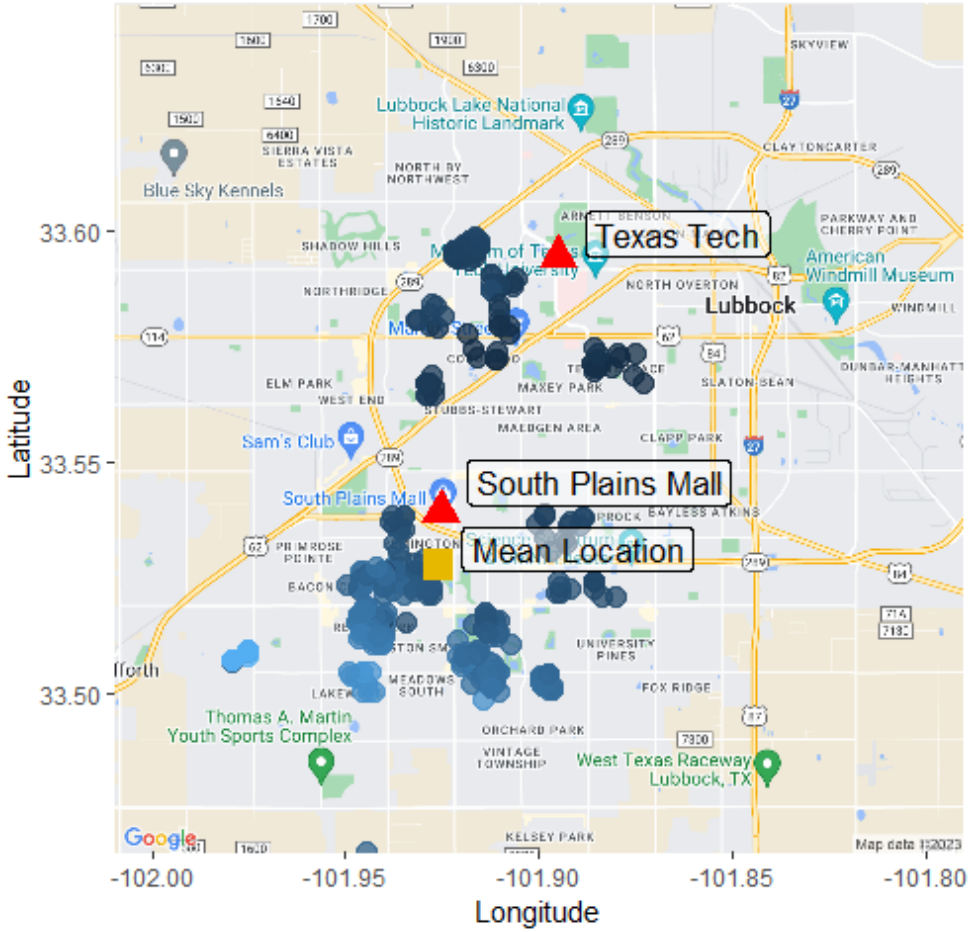
A good anchor position must capture a varying change in value from position in every direction

- A position five miles to the northeast must be able to efficiently differ from a position five miles to the southwest.

  - It must capture all other relevant impacts on value change in dependent variable for each position – as a varying longitudinal/latitudinal array.

  - Otherwise crime rates at a given position that affect home value would be entangled in other features at that position, which can be expected to differ at another position.

Empirical application: Lubbock, TX



The study space is also much smaller. Demographics across the study space are more uniform than those in Columbus.

Empirical application: Lubbock, TX

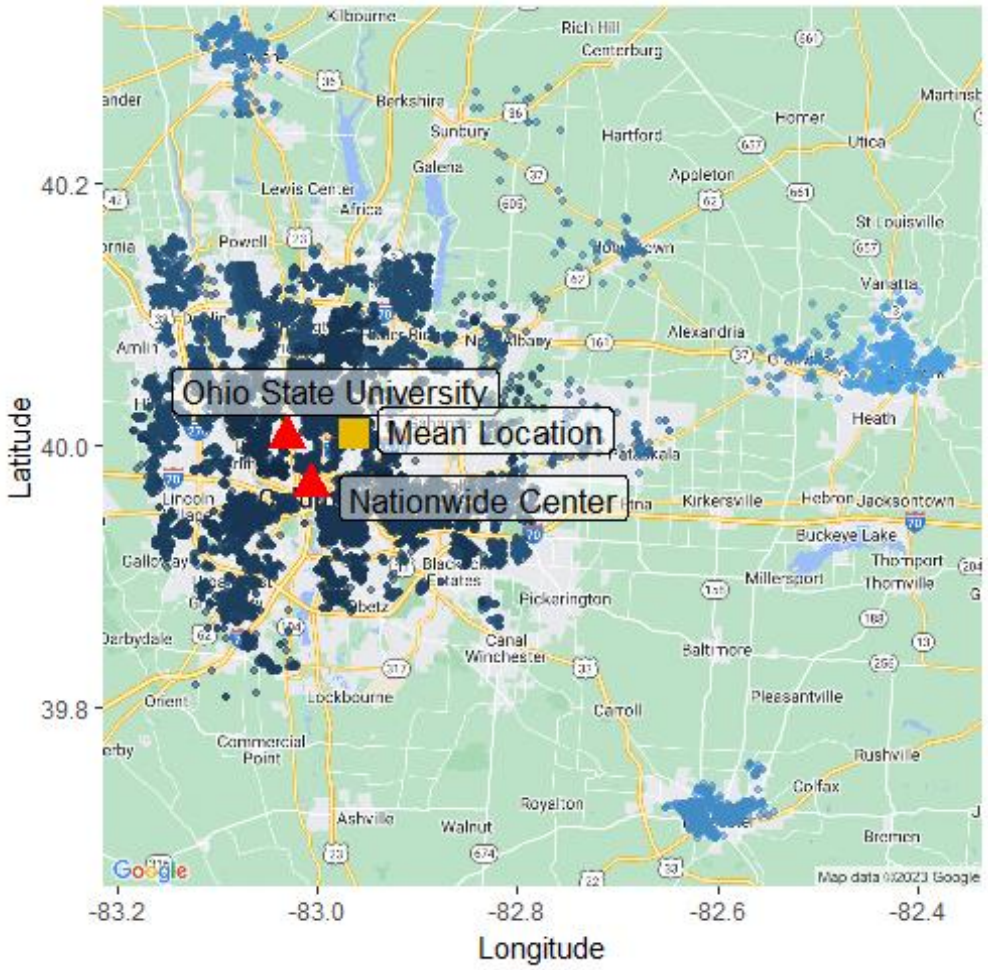
	Baseline	Distance (Tech)	Distance (Mall)	Distance: Mall&Tech	Fixed Point 1 (Tech)	Fixed Point 2 (Mall)
Square Foot	**86.95**** (4.47)**	**86.16**** (4.33)**	**87.11**** (4.42)**	**86.18**** (4.34)**	**83.23**** (4.05)**	**83.23**** (4.05)**
Lot size (Sq. yard)	1.97**** (0.38)	1.98**** (0.37)	1.97**** (0.38)	1.98**** (0.37)	2.05**** (0.34)	2.05**** (0.34)
House Age (Years)	-1837.61*** (249.83)	-2444.6**** (271.58)	-1625.00**** (257.00)	-2336.77*** (308.87)	-2221.41**** (250.30)	-2220.80**** (250.28)
Garage (0/1)	21375.78*** (7100.4)	24089.37**** (6900.7)	23466.00**** (7064.5)	24462.2**** (6923.1)	26373.60**** (6422.01)	26375.00**** (6422.30)
Env. Proxy birdsXspecies	**657.87**** (128.57)**	**308.89** (143.21)**	**531.09**** (134.53)**	**300.88*** (143.71)**	**128.45 (135.50)**	**128.44 (135.47)**
Income ($1000)	0.14* (0.087)	0.16** (0.084)	0.14* (0.087)	0.15* (0.085)	0.12 (0.079)	0.12 (0.079)
Dist. (Tech) (decimal degree)	-	**-10542.40*** (2140.41)**	-	**-9707.07*** (2425.05)**	-	-
Dist. (Mall) (decimal degree)	-	-	**10575.00*** (3651.5)**	**2974.55 (4049.1)**	-	-
Long (Tech) (decimal degree)	-	-	-	-	57699.00**** (6511.8)	-
Lat (Tech) (decimal degree)	-	-	-	-	-4822.7** (1793.5)	-
Long^2 (Tech) (decimal degree)	-	-	-	-	7345.86**** (797.17)	-
Lat^2 (Tech) (decimal degree)	-	-	-	-	-197.37**** (29.78)	-
Long (Mall) (decimal degree)	-	-	-	-	-	-13898.70**** (2810.1)
Lat (Mall) (degree)	-	-	-	-	-	-6050.13**** (1720.2)
Long^2 (Mall) (decimal degree)	-	-	-	-	-	7345.13**** (797.0)
Lat^2 (Mall) (degree)	-	-	-	-	-	-195.90**** (29.65)
LogLik	**-4465.09**	**-4452.35**	**-4460.00**	**-4452.70**	**-4424.11**	**-4424.11**
AIC$_c$	**8946**	**8922**	**8939**	**8925.5**	**8872**	**8872**
Adj. R^2	**0.85**	**0.85**	**0.85**	**0.85**	**0.86**	**0.86**

Parameter estimates of non-spatially distributed variables such as square footage or presence of a second story relatively stable across comparison models.

Parameter estimates of spatially distributed variables fully stabilize only under models with fixed position controls (models 5 and 6), especially the policy variables.

Measures of efficiency $(Adjusted\ R^2, AIC, Log\ Likelihood)$ change very little across all six models

Empirical application: Columbus, OH



Reflects a more common scale of regional science examination: much larger space, with far more observations and a more diverse population

Empirical application: Columbus, OH

	Baseline	Distance (OSU)	Distance (NWD)	Distance: OSU&NWD	Fixed Point 1 (OSU)	Fixed Point 2 (NWD)
Square Foot	102.09**** (1.72)	102.09**** (1.72)	102.09**** (1.72)	102.08**** (1.73)	101.785**** (1.73)	101.785**** (1.73)
Lot size (Sq. yard)	0.17**** (0.029)	0.18**** (0.031)	0.18**** (0.031)	0.18**** (0.031)	0.1645**** (0.031)	0.1645**** (0.031)
House Age (Years)	154.70**** (28.67)	132.903**** (28.69)	128.27**** (28.75)	137.58**** (28.95)	149.12**** (28.91)	149.12**** (28.91)
Second Story (0/1)	20861.52*** (1844.05)	20638.08**** (1838.62)	20550.28**** (1839.20)	20719.1**** (1839.83)	20264.29**** (1855.14)	20264.29**** (1855.14)
Income ($1,000)	1035.15**** (49.71)	923.23**** (51.08)	942.26**** (50.70)	912.199**** (51.90)	933.635**** (52.23)	933.635**** (52.23)
Offenses per District	4.75 (13.75)	-33.28** (14.31)	-29.91** (14.27)	-33.27** (14.32)	-26.54* (14.27)	-26.54* (14.27)
Pct White (%)	205.20**** (39.78)	304.747**** (41.15)	284.28**** (37.69)	294.05**** (42.11)	325.63**** (42.81)	325.63**** (42.81)
Dist. (OSU) (decimal degree)	-	-59072.7**** (6515.273)	-	-105061*** (38976)	-	-
Dist. (NWD) (decimal degree)	-	-	-58928.72*** (6744.519)	48275.56 (40338.58)	-	-
Long (OSU) (decimal degree)	-	-	-	-	-20551.57*** (8444.746)	-
Lat (OSU) (decimal degree)	-	-	-	-	-2982.95 (8266.658)	-
Long^2 (OSU) (decimal degree)	-	-	-	-	-94210.55*** (18972.2)	-
Lat^2 (OSU) (decimal degree)	-	-	-	-	-224564.1*** (41848.9)	-
Long (NWD) (decimal degree)	-	-	-	-	-	-18453.85*** (8105.454)
Lat (NWD) (decimal degree)	-	-	-	-	-	-17846.22** (8725.727)
Long^2 (NWD) (degree)	-	-	-	-	-	-94210.55*** (18972.2)
Lat^2 (NWD) (degree)	-	-	-	-	-	-224564.1*** (41848.9)
LogLik	-171260.5	- 171219.5	- 171222.4	- 171218.7	- 171228.1	- 171228.1
AIC_c	342537	342457	342463	342458	342480	342480
Adj. R^2	0.43	0.43	0.43	0.43	0.43	0.43

Parameter estimates of non-spatially distributed variables such as square footage or presence of a second story relatively stable across comparison models.

Parameter estimates of spatially distributed variables fully stabilize only under models with fixed position controls (models 5 and 6), especially the policy variables.

Measures of efficiency $(Adjusted\ R^2, AIC, Log\ Likelihood)$ change very little across all six models