Appendix E: ZTRAX Cleaning Decisions

As described in the manuscript, Zillow's ZTRAX database was made available through PLACES lab at Boston University (Zillow 2021). First, property sales were linked to property boundary polygons by PLACES, and those proximate to NHD lakes larger than 4 ha and met the medium-confidence transaction sale filter described in (Nolte et al. 2023) were selected. This filter varies by state and includes a range of different filter categories, such as document type, sale code, and intra-family transfers. This data was further filtered by state, such that if more than 67% of sales in a state met the high-confidence filter, only high-confidence sales were included. For the remaining states, medium and high confidence sales were selected.

We then merge the sales with the water quality data, keeping only those observations with a Secchi of chl-a sample on a lake within 2,500 m from the property within 5 years from the closest summer to the year of sale. After adjusting sales prices to 2021 $, we further exclude sale prices less than $10,000 (Gindelsky, Moutlton, and Wentland 2022) and remove the top 1% of sale prices in each State (Chun et al. 2021). To handle the missing property attributes problem, we consider a range of different sample selection criteria documented in Table 1 in the manuscript. For one sample, we do not include any property attributes and only perform additional data refinement related to water quality, (e.g. must meet a minimum number of sales within the lakefront property buffer). Samples that did not meet these criteria were dropped.

For samples with property attributes included, we cleaned the sample based on the selected attribute first removing all nonsensical values, such as negatives and zero values. We then identify outlier values in the 99^{th} percentile. For continuous property attributes, we also define the 1^{st} percentile as outliers. Then all outliers are removed from the sample. Finally, we

keep only properties classified as residential, single-family houses (RR000, RR101, RR102, RR999).

References

Chun, Yung, Stephanie Casey Pierce, and Andrew J. Van Leuven. 2021. "Are Foreclosure Spillover Effects Universal? Variation Over Space and Time." Housing Policy Debate 31(6):924–46. doi: 10.1080/10511482.2021.1882533.

Gindelsky, Marina, Jeremy G. Moutlton, and Scott A. Wentland. 2022. "Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata." in Big Data for Twenty-First-Century Economic Statistics. Vol. 79, Studies in Income and Wealth. National Bureau of Economic Research.

Nolte, Christoph, Kevin J. Boyle, Anita M. Chaudhry, Christopher Clapp, Dennis Guignet, Hannah Hennighausen, Ido Kushner, Liao Yanjun, Saleh Mamun, Adam Pollack, Jesse Richardson, Shelby Sundquist, Kristen Swedberg, and Johannes Uhl. 2021. "Data Practices for Studying the Impacts of Environmental Amenities and Hazards with Nationwide Property Data." Land Economics 102122-0090R. https://doi.org/10.3368/le.100.1.102122-0090R

Zillow. 2021. Zillow's Transaction and Assessment Database (ZTRAX). https://www.zillow.com/research/ztrax/