Online Appendix:

"Inaccuracies in Low Income Housing Geocodes: When and Why They Matter"

Contents

A Sample Characteristics

A.1 Comparison of subsamples and full dataset

The table below compares the full LIHTC data, the California subset of data used for the main analysis, as well as the national subset of data. The latter two datasets are summarized before the 35m cutoff for the discrepancy between the HUD and Google locations is applied. The California subset, therefore, contains all new development in California between 1999 and 2010. The national subset contains all developments between 2012 and 2020 that are proximate to our survey respondents' ZIP codes.

Of note, "0" counts for the number of units and the allocated amount of funding likely reflect errors in data entry, as some LIHTC developments will list multiple low-income units but zero units overall — a physical impossibility. While not the focus of our analysis, this additional measurement error underscores the need for scholars to analyze the quality of administrative datasets and disseminate recommendations as a form of public goods.

Table A.1: Comparison of subsamples used for analysis and full LIHTC dataset.

		Full LIHTC Data (N=50,567)	**California Subset (N=1,266)**	**National Subset (N=959)**
Year Placed in Service	*Min.*	1987	1999	2012
	Median	2003	2005	2014
	Mean	2002.7	2004.7	2015.0
	Max.	2022	2010	2020
Allocated Amount (USD Yearly)	*Min.*	0	10,000	0
	Median	340,265	667,328	961,646.5
	Mean	568,712.7	762,409.0	1,167,973.2
	Max.	120,000,000	3,850,000	19,748,284
Number of Units (Total)	*Min.*	0	2	0
	Median	47	72	66
	Mean	68.1	83.8	83.9
	Max.	2,025	665	1,238
Number of Low-Income Units	*Min.*	0	2	0
	Median	40	68	60
	Mean	60.4	76.6	74.0
	Max.	1,813	356	405

A.2 Variation in HUD accuracy by sample characteristics

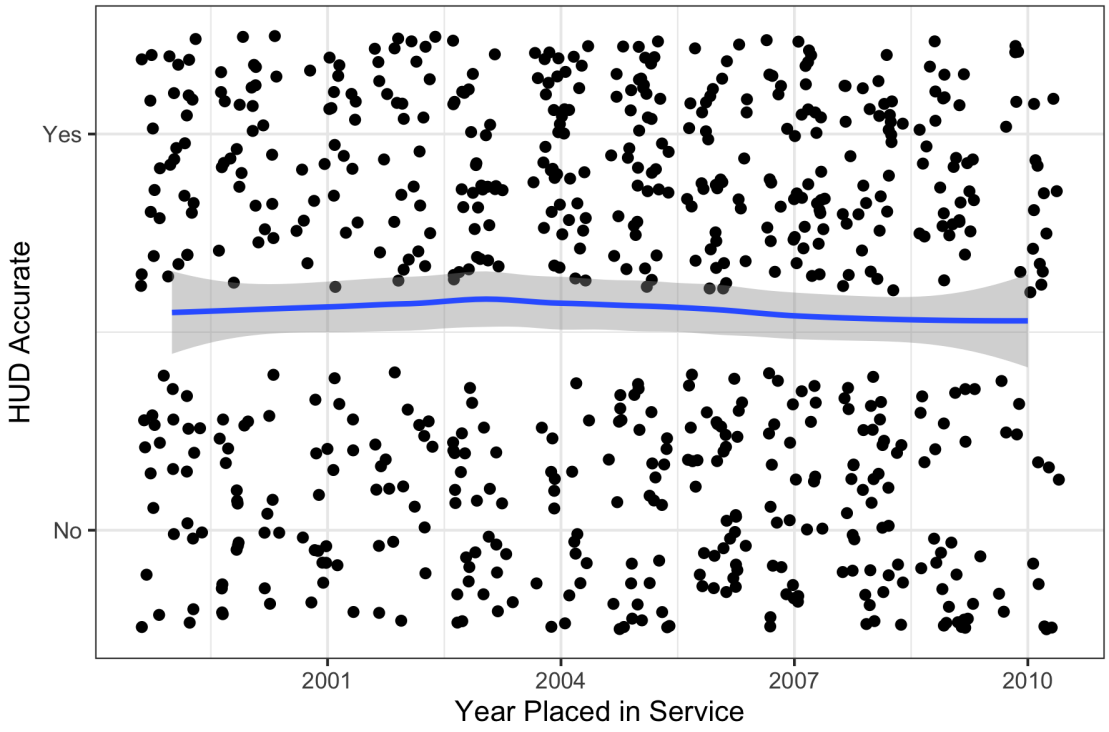

Figure A.1: A scatterplot showing the relationship between year placed in service and accuracy of the HUD geocode among manually checked observations ($N = 851$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

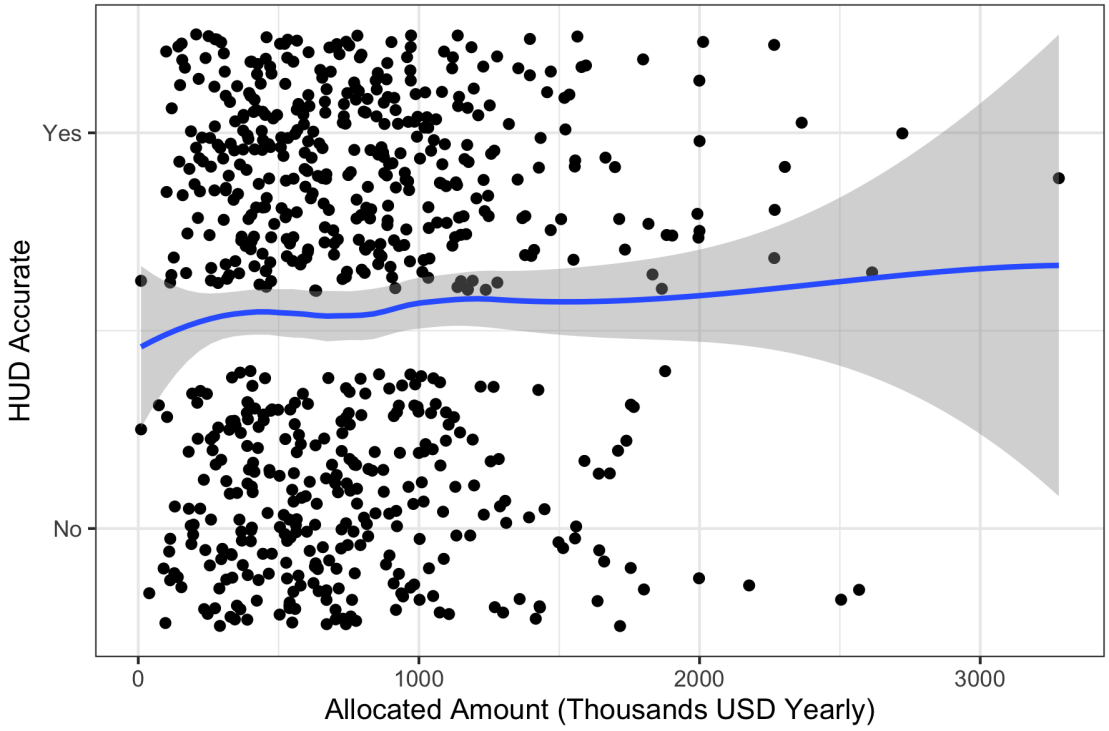

Figure A.2: A scatterplot showing the relationship between yearly allocated amount in thousands of USD and accuracy of the HUD geocode among manually checked observations ($N = 851$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

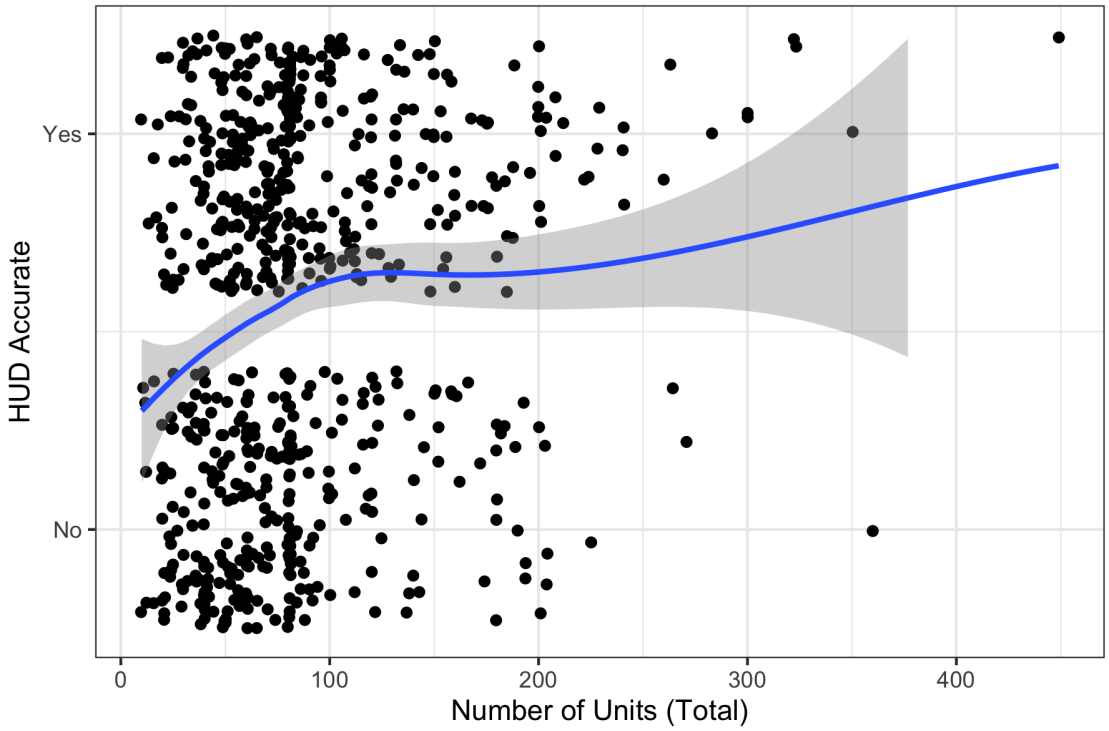

Figure A.3: A scatterplot showing the relationship between total number of units and accuracy of the HUD geocode among manually checked observations ($N = 851$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

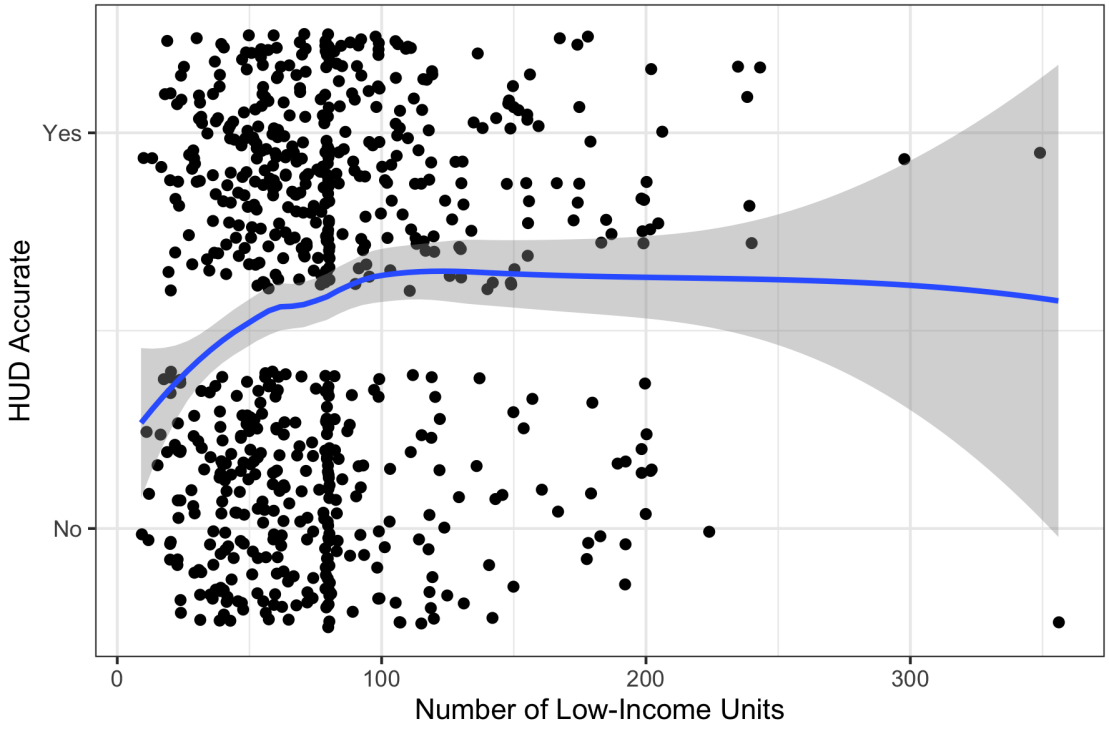

Figure A.4: A scatterplot showing the relationship between the number of low income units and accuracy of the HUD geocode among manually checked observations ($N = 851$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

A.3 Variation in HUD accuracy by population density

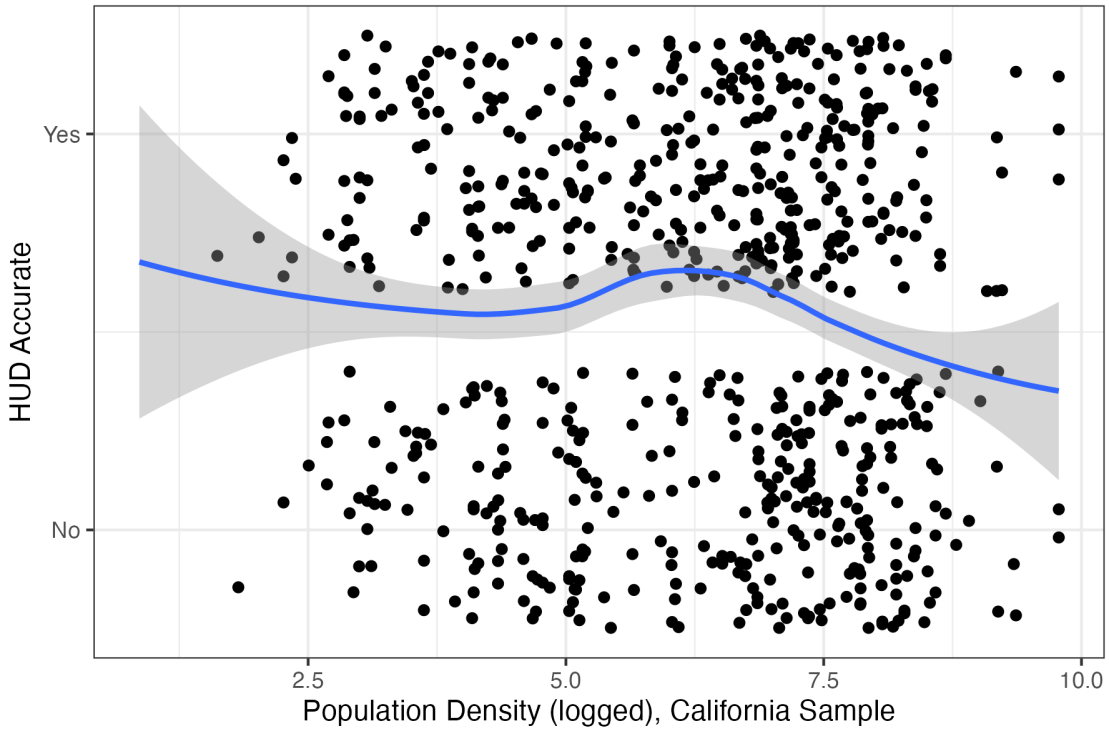

Figure A.5: A scatterplot showing the relationship between the census tract population density (logged) and accuracy of the HUD geocode among manually checked observations in the California sample (N = 851). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

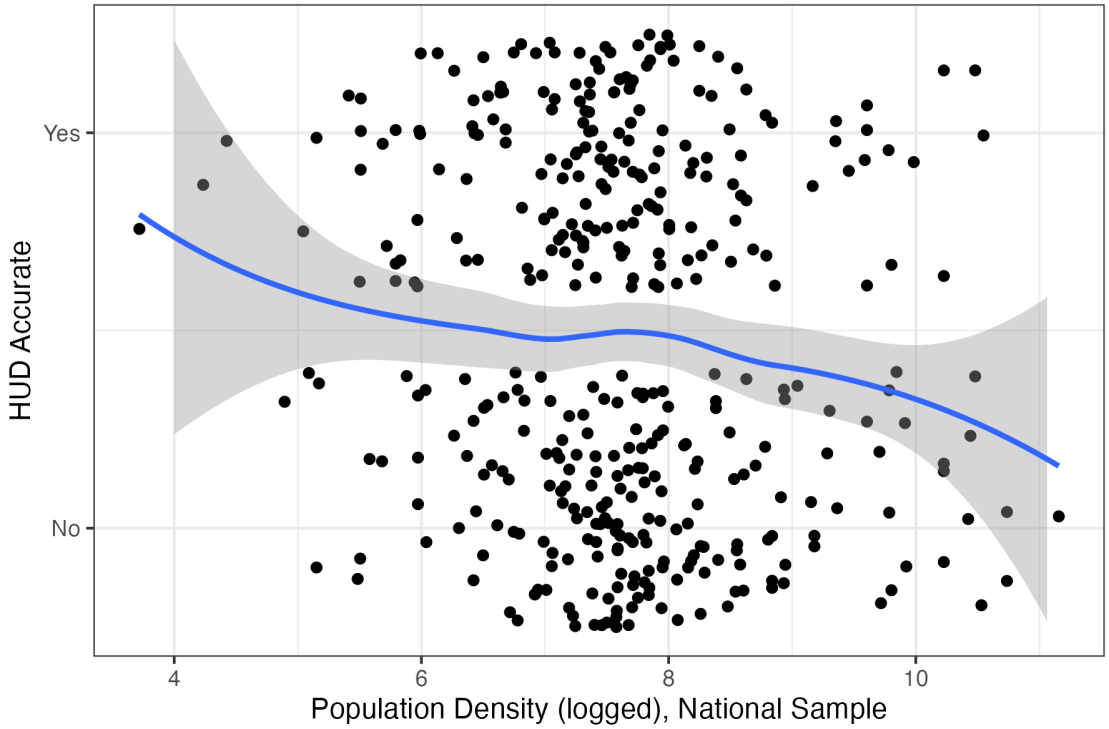

Figure A.6: A scatterplot showing the relationship between the census tract population density (logged) and accuracy of the HUD geocode among manually checked observations in the national sample (N = 477). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

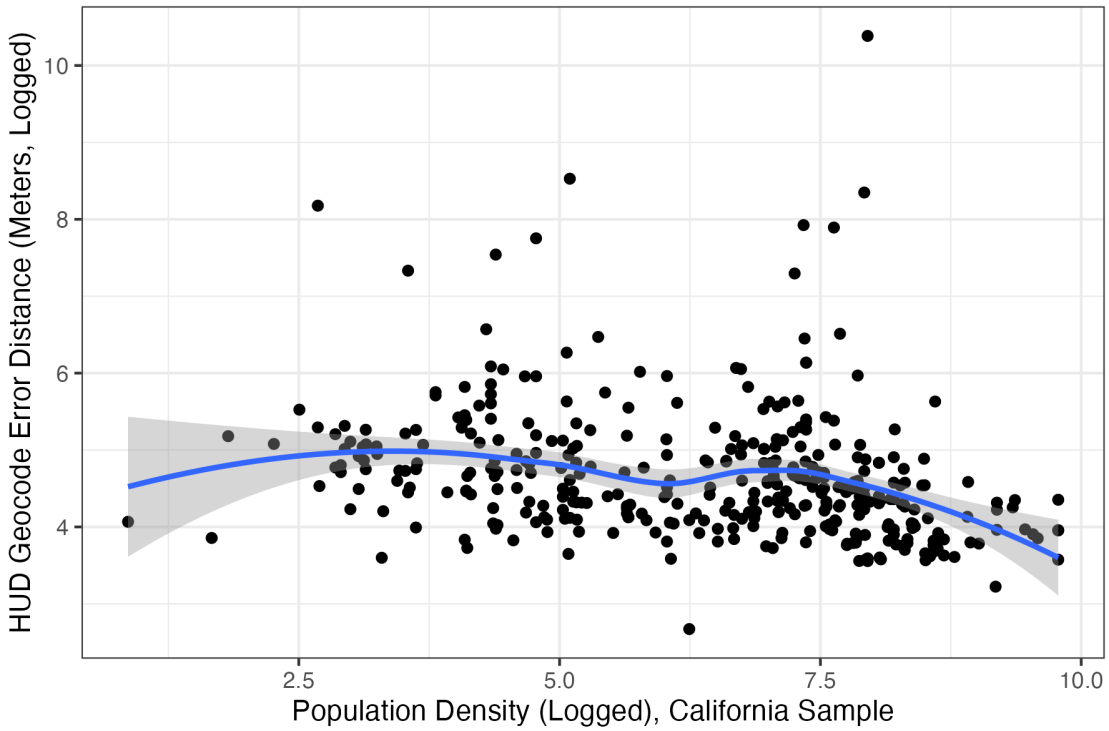

Figure A.7: A scatterplot showing the relationship between the census tract population density (logged) and the distance between the inaccurate HUD geocode and the true LIHTC location among manually checked observations in the California sample ($N = 380$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

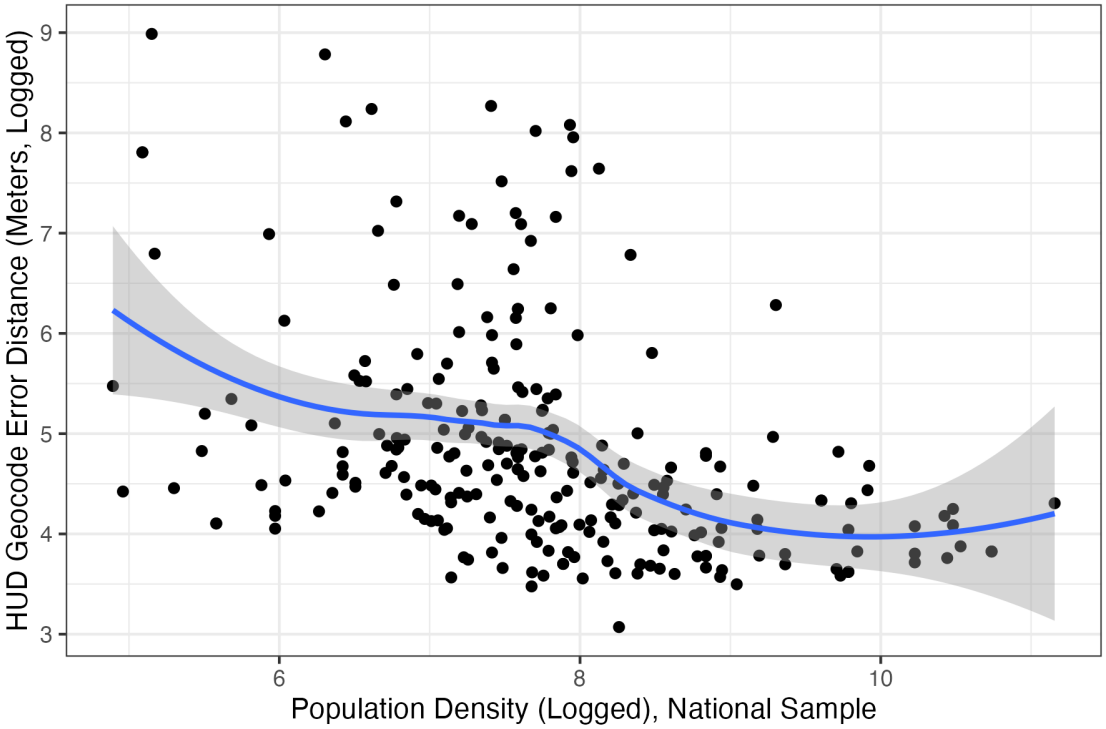

Figure A.8: A scatterplot showing the relationship between the census tract population density (logged) and the distance between the inaccurate HUD geocode and the true LIHTC location among manually checked observations in the national sample ($N = 252$). Points are jittered to better show the distribution. The blue line is a loess curve fit to the data with a 95% confidence interval shaded in gray.

B Comparison of HUD and Google

B.1 HUD accuracy by distance to Google

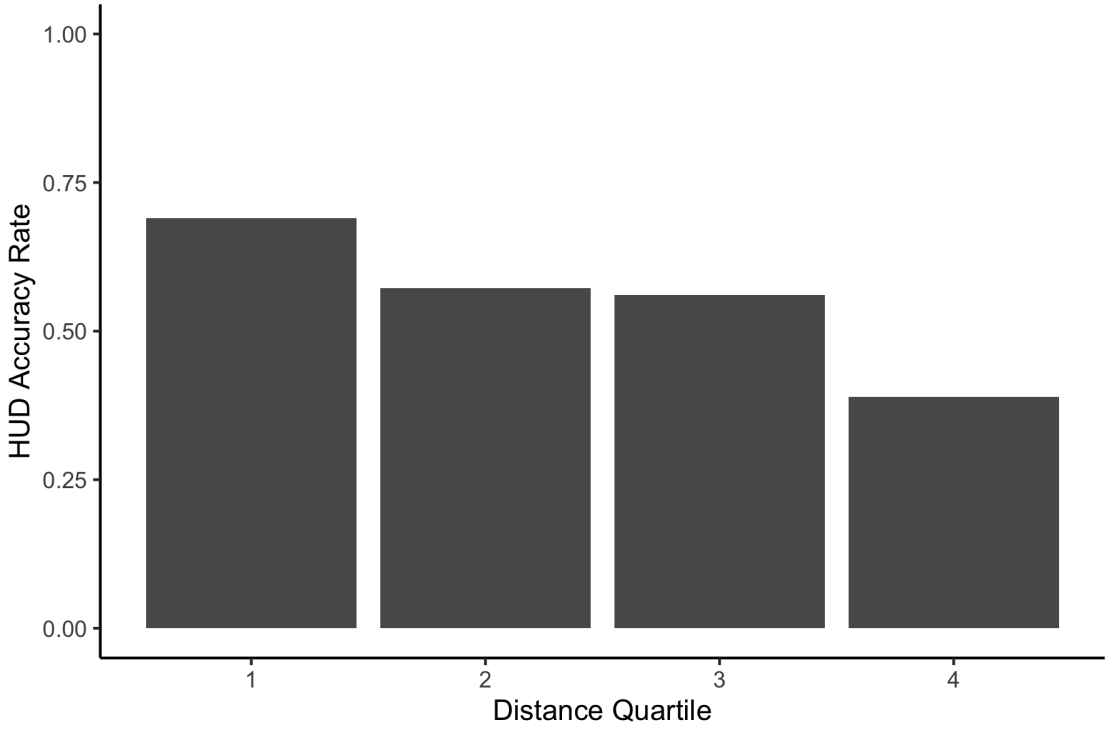

Figure B.1: A bar plot showing the relationship between the distance between the HUD and Google coordinates for a given facility and accuracy of the HUD geocode among manually checked observations ($N = 851$). Quartile 1 is between 35 and 52.67 meters, quartile 2 is between 52.67 and 76.15 meters, quartile 3 is between 76.15 and 121.77 meters, and quartile 4 is greater than or equal to 121.77 meters.

B.2 How accuracy is affected by directionals

Table B.1: Accuracy of HUD and Google coordinates, by whether there is a directional (e.g.,"N", "S", "E", "W") in the facility address

	HUD	**Google**
WITH direction in address	50.7%	92.5%
WITHOUT direction in address	56.8%	95.1%