Investor Ownership of Twin Cities Rental Properties: Technical Appendix

How is *investor ownership* defined?

This tool relies on the MetroGIS Regional Parcel Dataset, which compiles tax/real estate data assembled from assessors' offices in Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington counties. As we do not have information on whether a property is an investment property, we use a unit's homesteaded status as a proxy. For a given cutoff—e.g., 10 properties—we classify an entity as an investor if it owns 10 or more *non*-homesteaded residential properties in the seven-county Twin Cities area.

In this tool, a residential property, or home, includes single-family homes, townhomes, and condominiums. We exclude other residential properties, such as apartments and duplexes that are always intended for rental purposes.

How are the data prepared for analyses?

The tool uses *use classifications*, the classifications that determine the tax rate for each property, to filter tax data to only single-family homes, townhomes, and condominiums. A property's owner can report up to four use classifications in the dataset. For example, the owner of one parcel in Hennepin County reported its use classifications as (1) residential, (2) farm, (3) farm—non-productive land, and (4) commercial. We exclude a property if any of its use classifications suggest the property may not be residential. Thus, we would exclude the Hennepin County parcel mentioned above. A small share of records include no use classifications. (See the Metro Regional Parcel Dataset Attributes for more information on how complete each attribute is.) In those cases, we look at the style of home and dwelling type, if available, to determine the structure's classifications for our purpose. The home style and dwelling type are also used to identify properties that may not be residential or single-family (e.g., garages or duplexes). This process also excludes any property that has more than one use and was not already captured in our examination of use classifications, as detailed above (e.g., a single-family home that is also a convenience store).

Preparing this dataset for analysis required dealing with inconsistent spellings and abbreviations of tax payer and owner names and addresses. To overcome these issues, we did fuzzy matching on names and addresses separately. There are many different fuzzy-matching methodologies, each of which could lead to slightly different results. Here we describe our chosen methodology in more detail.

Methodology

1. We removed any special characters or punctuation (e.g., commas or hyphens).
2. Starting with a list of unique names, we created 3-character-grams for each of the names. We also padded the names with a white space on both sides to capture instances where parts of the name are in a different order (e.g., "Jane A Smith" versus "Smith Jane A"). For example, the 3-character-grams for the name "Jane A Smith" are " Ja", "Jan", "ane", "ne ", "e A", " A ", "A S", " Sm", "Smi", "mit", "ith", and "th ".
3. Then we created a TF-IDF (Term Frequency-Inverse Document Frequency) matrix based on those 3-character grams. The TF-IDF value for a given 3-character-gram in a given name/address is calculated as follows:

   $$\text{tf-idf} = \text{term-frequency} \times \text{inverse-document-frequency}$$
   $$\text{term-frequency} = \text{the number of times a given 3-character-gram appears in the name/address}$$
   $$\text{inverse-document-frequency} = \log\left(\frac{1 + \text{total number of names/addresses}}{1 + \text{number of names/addresses with 3-character-gram}}\right) + 1$$

   For example, for "Jane A Smith" and "Jay A Doe," part of their TF-IDF matrix would look like this:

	" Ja"	"Jan"	"Jay"	"ane"	" A "	"Doe"	"ith"	…
Jane A Smith	0.1	0.2	0.0	0.3	0.1	0	0.6	…
Jay A Doe	0.1	0.0	0.2	0.0	0.1	0.5	0.0	…
…	…	…	…	…	…	…	…	…

4. Finally, we identified potential matches based on cosine similarity index, which is defined as:

   $$\text{cosine similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

   where A and B are the TF-IDF vector representations of tax payer names (or addresses) from step 3; higher values suggest higher similarity. We use an initial cutoff of 0.7 similarity score to identify potential matches.
5. To better pick a cutoff, we randomly selected 1,000 pairs and manually labeled them according to whether they are correct. We then ran a logistic regression where the similarity score is the explanatory variable and the manual label (1 if the pair is correct, and 0 otherwise) is the response variable. Then we picked the similarity-score cutoff based on the point at which the predicted probability of the pair being correct is at or above 50 percent. This gave us a similarity-score cutoff of 0.80.
6. Even with a relatively high cutoff, there are many cases where names or addresses meet the cutoff but should not be matched. For example, "Jane A Smith" and "Jane E Smith", or "100 Main St" and "101 Main St". To reduce the number of false matches, we required that middle name initials, if present, have to match for the names to be considered a matched pair. For addresses, we required that all numbers—including street number or suite number—match.

About 3.6 percent of entities share the same address but have different names, or vice versa. We considered these entities to be the same entity if they share either the same name or the

same address, or both. For instance, if entities "A" and "B" list their addresses as "123 Main Street," we listed them in our final dataset as the first name that shows up alphabetically, or "A."

The MetroGIS data include newly constructed or remodeled homes and condominium units that are on the market and are thus still owned by their construction company or developer. We identified these entities based on their names and excluded any properties associated with them from the sample.

How is investor type determined?

For each non-homesteaded property, we classify its ownership as one of three types:

- *Bank-owned*: This includes properties that are owned by banks, mortgage servicers, or any financial companies, identified based on their names. These properties may be in real estate ownership after a foreclosure.
- *Public/community-owned*: This includes properties that are owned by nonprofit or public entities such as Twin Cities Habitat for Humanity or a local housing and redevelopment authority, also identified based on their names.
- *Neither*: This includes all properties that do not fall into one of the two types above. These properties include those on the rental market and owned by for-profit entities.