An Algorithmic Approach to Identify Parcel Ownership with Publicly Available but Messy Data

Nicholas Polimeni (nicholas.polimeni@gatech.edu), Georgia Institute of Technology

Following the Great Recession, institutional investors have directed funds to the wholesale acquisition of single family homes, converting many into rental properties. Analysts, academics, and practitioners alike have had difficulty quantifying this new investment class due to the challenges of messy data that obfuscates the true ownership of parcels.

Current approaches utilize OpenRefine, a fuzzy matching tool, and additional manual inspection (An 2023) to cluster owners with similar names and addresses. The computational complexity of standard fuzzy clustering is $O(n^2)$, making it impractical for large datasets, without also considering the additional manual effort. Further, these algorithms are difficult to optimize for addresses; for instance, take two addresses that are identical except for their address number. Any string distance metric, the underlying calculation for fuzzy matching, will score such addresses as extremely similar.

This paper's method exploits the unique characteristics of addresses to develop an algorithmic, vectorized ($O(n)$) comparison method. Such a method demonstrates significant performance gains on large datasets, and avoids additional manual effort, making it an ideal tool for practitioners. In addition to the base algorithm, an extended algorithm is being developed to identify owners that use multiple addresses, and to supplement clustering with business registry or other data sources.

Table 1. Possible Cases of Ownership Data Inconsistencies for Same True Owners

Description	Cases	Name is a..., (case #)	Solution
Same Name, Same Address	Identical Name, Identical Address	Person's name (1)	High threshold fuzzy matching on address only
		Company's name (2)	
	Similar Name, Identical Address	Person's name (3)	High threshold fuzzy matching on address only
		Company's name (4)	
	Identical Name, Similar Address	Person's name (5)	High threshold fuzzy matching on address only
		Company's name (6)	
Same Name, Different Address	Identical Name, Different Address	Person's name (7)	Researcher must decide whether to assume that same name owners in given geography are equivalent.
		Company's name (8)	High threshold matching on company name
	Similar Name, Different Address	Person's name (9)	High risk of false positive
		Company's name (10)	High threshold matching on company name
Different Name, Same Address	Different Name, Identical Address	Person's name (11)	High threshold fuzzy matching on address only
		Company's name (12)	High threshold fuzzy matching on address only
	Different Name, Similar Address	Person's name (13)	High threshold fuzzy matching on address only
		Company's name (14)	High threshold fuzzy matching on address only
Different Name, Different Address	Different Name, Different Address	Person's name (15)	N/A - no way to capture
		Company's name (16)	Business registry data needed

Algorithm

The key component of the base algorithm is the address key (labeled ADDR KEY). This key is constructed to provide a fuzzy but highly accurate version of an address (in this case, the parcel owner's address). It does this by selecting for key features- the address number, the suite number, the zip code, and the last two letters of the longest substring in the street address.

The base algorithm accounts for cases 1-6 and 11-14 as described in Table 1 and ran in under 10 minutes on 50K records. Due to the complexity of the algorithm, this runtime should only linearly increase with data size, whereas other methods increase exponentially proportionally to the square of the data size. The expected execution time for 5M records is 16 hours.

An extended version of the algorithm attempts to match names in the case that their ADDR KEY was not sufficient (cases 7, 8, 10, 16). This secondary step applies more traditional fuzzy matching (Jaccard Similarity or a custom metric) to a reduced search space. However, this is a computationally intensive process that might not be deemed necessary in some cases.

Table 2. Example ADDR KEY Matching

Address 1	Zip 1	ADDR KEY 1	Address 2	Zip 2	ADDR KEY 2	Matched?
PO BOX 490734	30363	490734-0-OX-30363	P O. BOX 490734	490734-0-OX-30363	30363	Yes
3505 KOGER BLVD 400	30315	3505-400-ER-30315	3505 KOGER BLVD., SUITE 400	30315	3505-400-ER-30315	Yes
One Buckhead PL STE 300	30305	1-300-AD-30305	One Buckhead PL STE 325	2-0-IN-30303	1-325-AD-30305	No
5 PEACHTREE ST	30308	5-0-30308	5 PEACHTREE ST	30354	5-0-30354	No

Table 3. Comparison to OpenRefine Method (An 2023): Total # Properties by Largest Owners in Fulton County (2021)

Landlord	OpenRefine (OR)	OR + Manual	Base Algorithm	Net Diff (Base - OR)*	Match w/ Business Registry
Invitation Homes	554	708	692	+138	624
Progress / Front Yard Residential	556	760	702	+146	480
Starwood Capital	364	450	558	+194	365
Cerberus Capital	340	349	342	+2	324
Sylvan Reality (RNTR)	251	433	808	+557	805
Amherst Capital	716	750	715	-1	707
Tricon Residential	277	280	250	-27	250
SUM	3058	3730	4070	+1009 (33%)	3555

*Algorithm should be compared to OpenRefine with minimal manual work; such manual work becomes more difficult for large datasets. This approach is intended for big data.

Further Work & Limitations

While the extended algorithm considers owner name, a high emphasis is placed on owner address. The base, efficient algorithm generally appears to provide an accurate measure ownership scale. However, it requires extension or a small manual effort in querying the data structure to provide a more complete picture of investors that use multiple addresses.

We acknowledge that this method currently requires a high degree of programming expertise to execute. A Python library/tool is under development to improve ease of use. **Please contact Nicholas Polimeni with inquiries.**