Geocoding – the Columbus way!

Rahul Bakshi

About the Research

- Part of Masters' Thesis
- Advisor: Craig Knoblock
- Other Committee members:

  Cyrus Shahabi and John Wilson
- Build a Geocoder with maximum accuracy

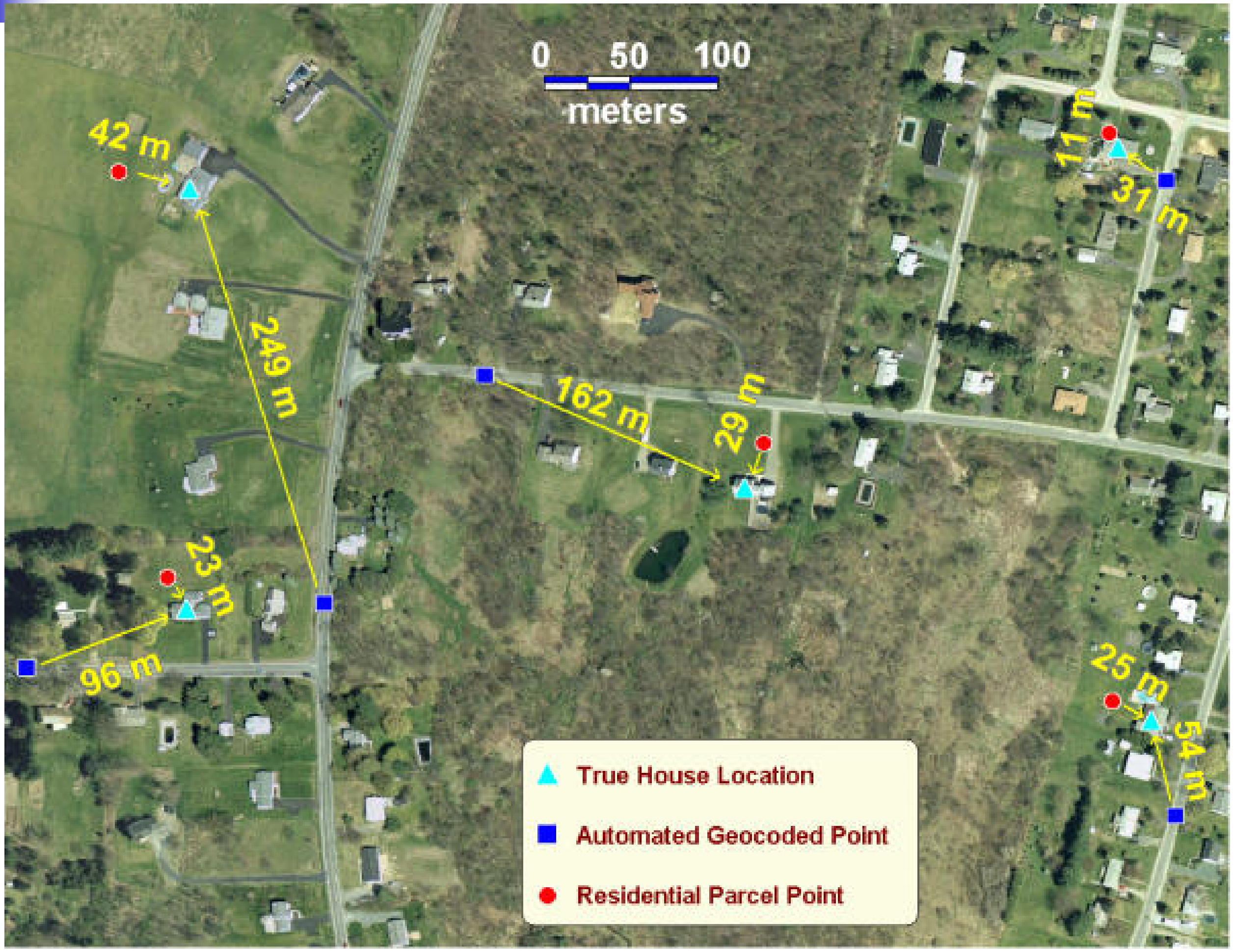
Thesis statement

- **The accuracy of the geocoded coordinates of a location can be significantly improved by exploiting online property-related data**

Motivating Problem

- Inaccuracies in the existing applications
- The error margins become critical in some applications:
  - Aligning Vector Data and Satellite Imagery
  - Environmental Health Studies
  - Urban Rescue and Recovery Operations

Positional Error Comparison



Reference: Cayo, M. R. and T. O. Talbot (2003). "Positional error in automated geocoding of residential addresses." International Journal of Health Geographics **2**(10).

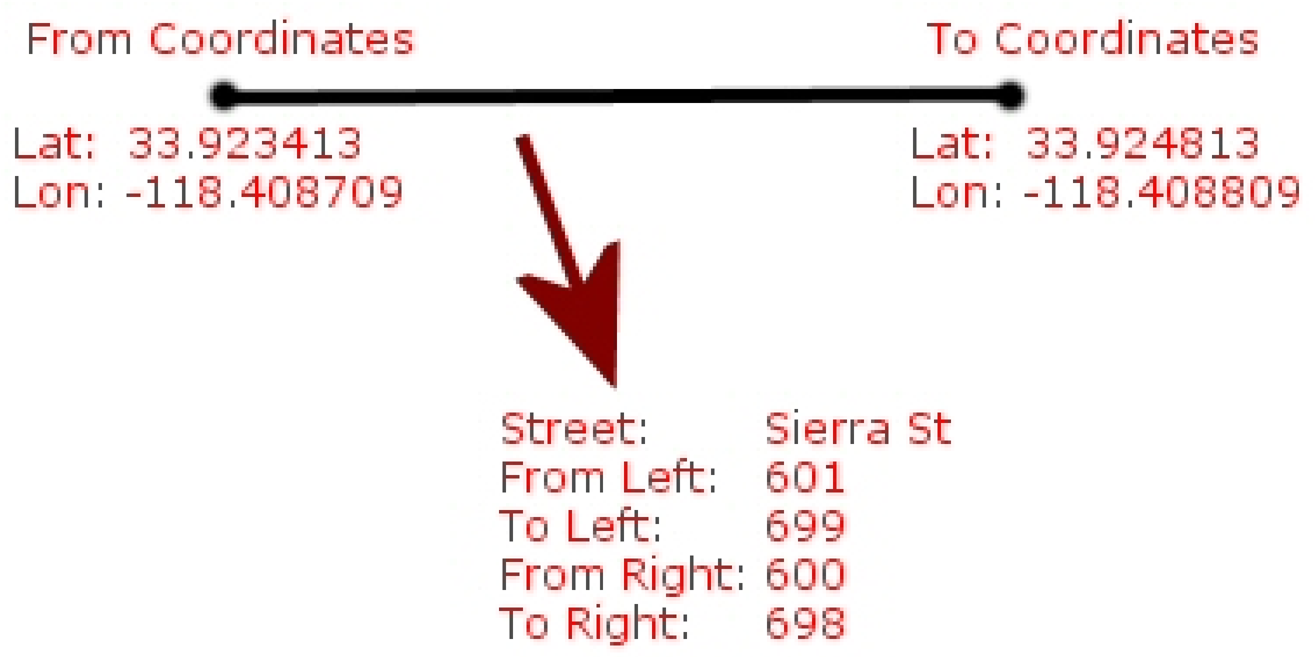
Street Data

- For the US, there are three main providers for street data
  - Geographic Data Technology (GDT)
  - Navigation Technologies (NavTech)
  - TIGER/Lines (Bureau of the Census)

Limitations of these sources

- Provide the address ranges and latitude/longitude information for the end points
- No data about number of addresses in a segment
- No data about the size of address/lots

Information in Street Sources



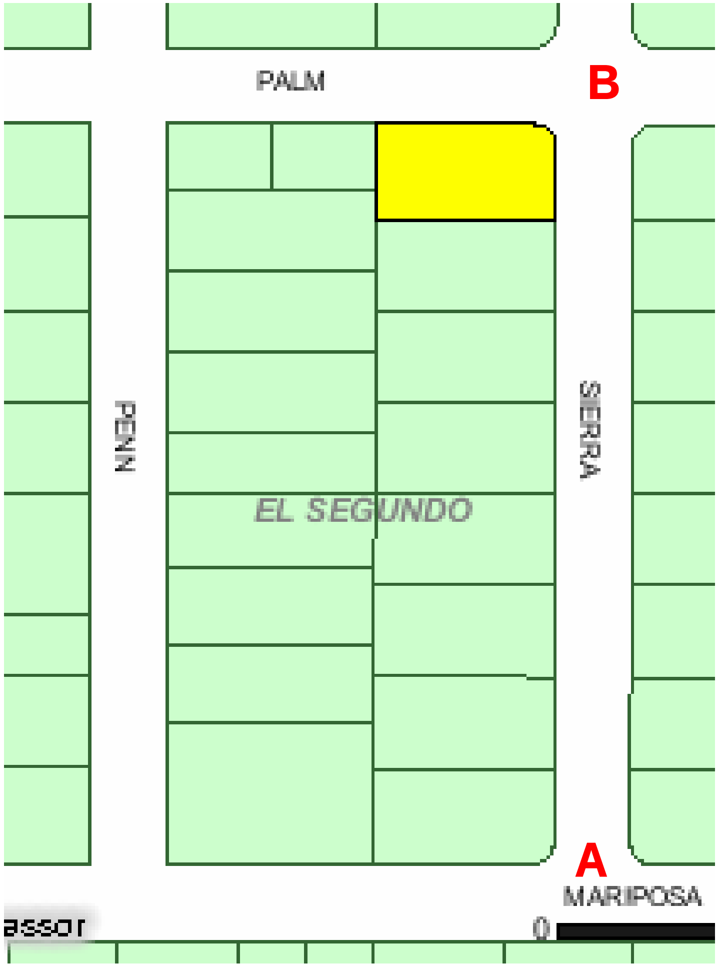
Lat: 33.923413	Lat: 33.924813
Lon: -118.408709	Lon: -118.408809

Street:	Sierra St
From Left:	601
To Left:	699
From Right:	600
To Right:	698

Existing Approach

- Address range method
- Get the street data from sources like NavTech, GDT, TigerLines
- Approximate the location based on information in the street data
- Example
  - Address to locate: 645 Sierra St, El Segundo, CA -90245

Example



Sierra St
From: A (33.923413, -118.408709)
To: B (33.924813, -118.408809)

Addresses on the Left: 601-699
Addresses on the Right: 600-698

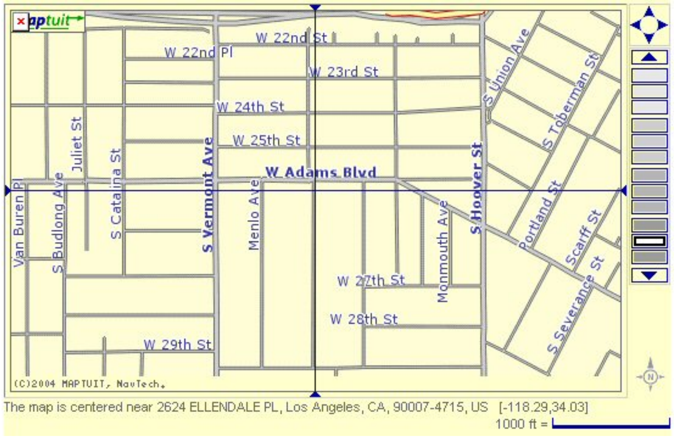
645: Left Side
22nd out of the 50 addresses on the left side

Interpolate the address on the street

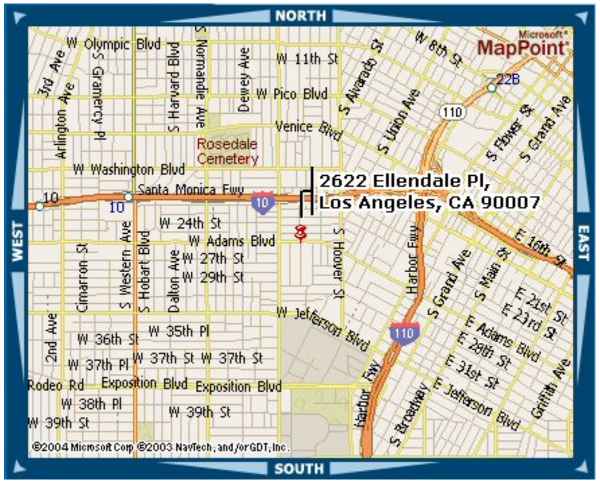
Limitations of the existing approach

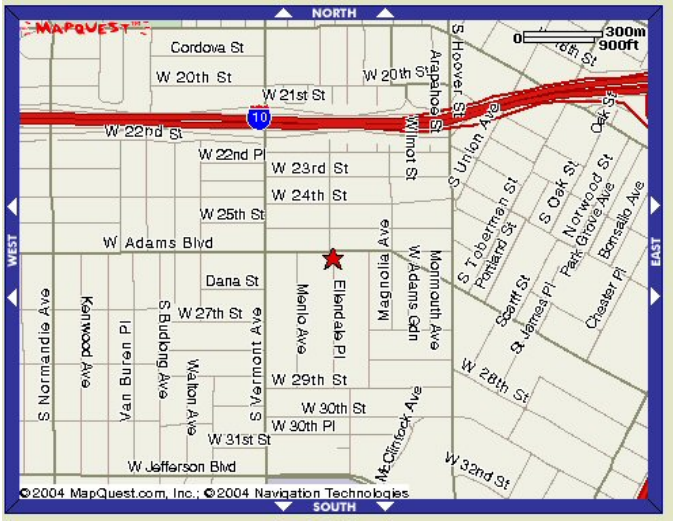
- Assumes all addresses are present in the given range – which is seldom the case
- Does not take into account the lot sizes
- Geocodes non-existent addresses as well
- E.g.: The following address **does not** exist - 2622 Ellendale Pl, Los Angeles, CA – 90007
- Lets see what do the existing services have to say...

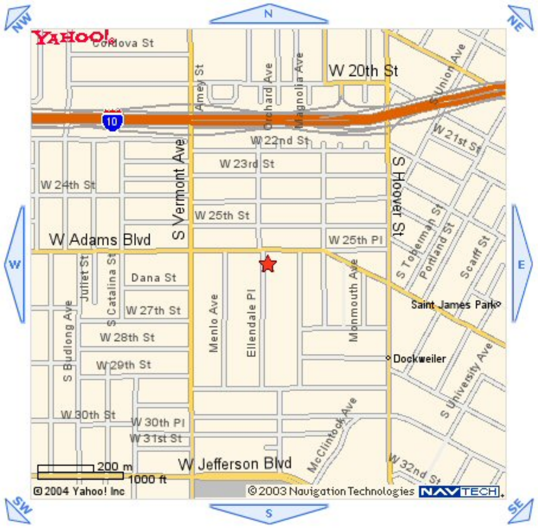
All of them geocode it !



The map is centered near 2624 ELLENDALE PL, Los Angeles, CA, 90007-4715, US [-118.29,34.03]







The Columbus approach

- Make use of the data already on the Internet
- Property tax sites – repository of information that one requires to make the interpolations more accurate
- Take the number of houses in account
- Take the lot sizes in account

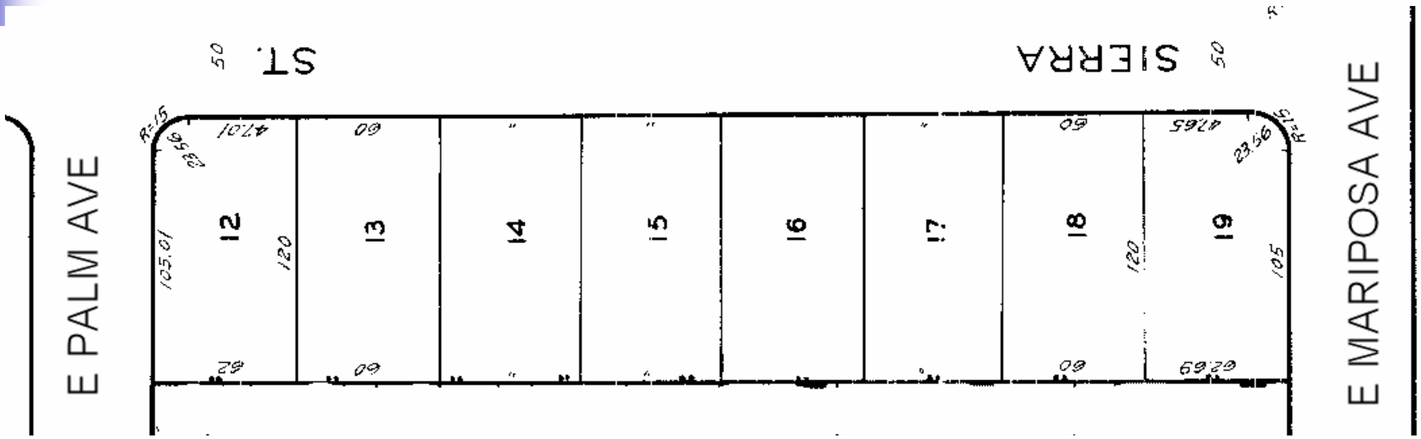
Uniform lot-size method

- Works when data source having information on the property parcels/addresses exists
- Exploits these sources to get the number of lots on the street segment
- Assumes all lots are equal in dimension

Outline of the method

- Get the information of the street segment from the street data source
- Query the property tax source to get the number of parcels before and after the current address
- Approximate the location of the address based on the new values

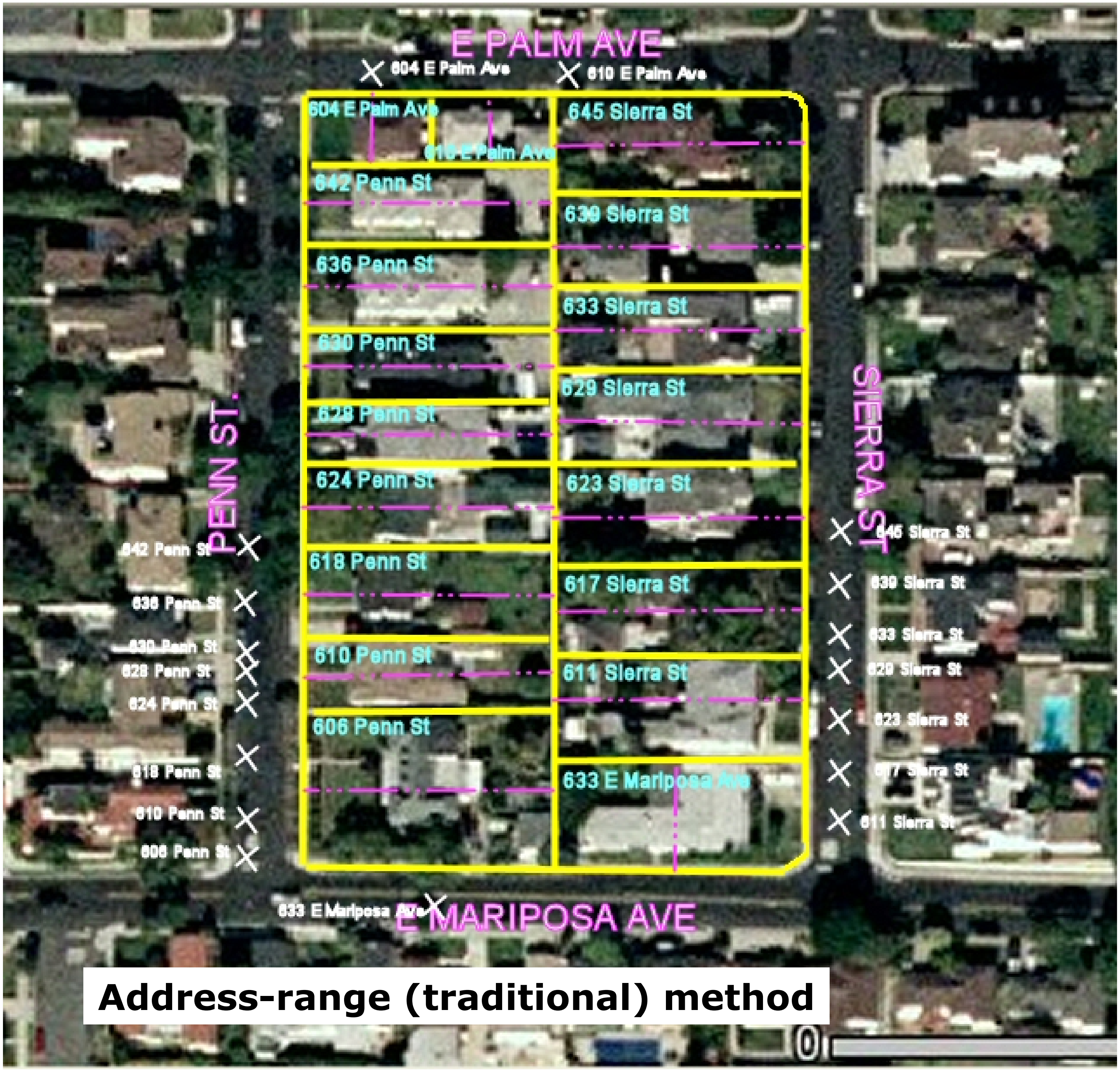
Corner lot problem



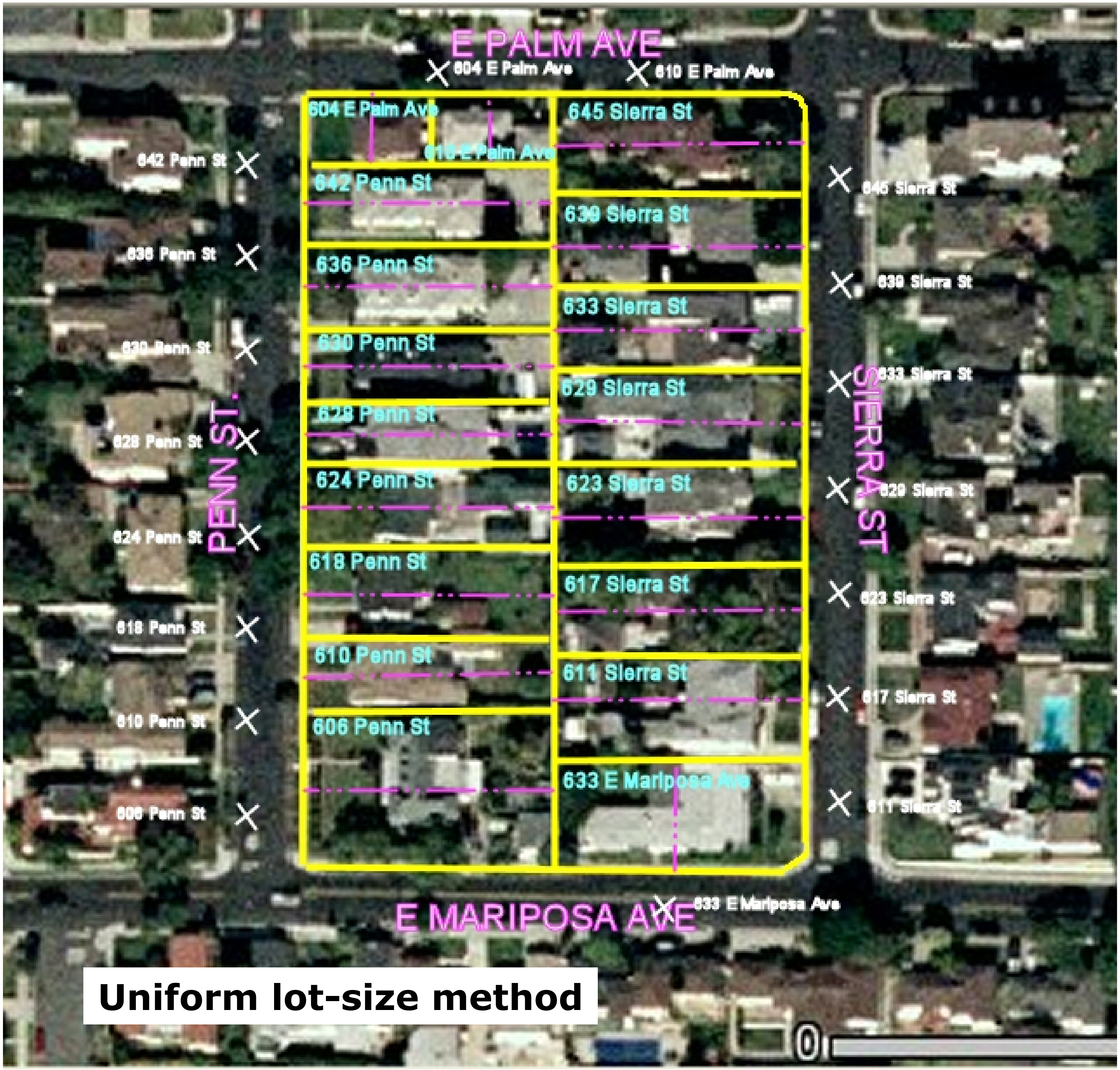
Number of dimensions on the street = number of lots on the street + corner lot

Algorithm

- Get the street data from the street-data-source
- Get number of lots before and after the current address from the property data source
- Add a corner lot
- Calculate the street length in terms of earth coordinates
- Calculate the lot size based on the street length and the number of lots on the street
- Interpolate the location of the address based on the average lot size



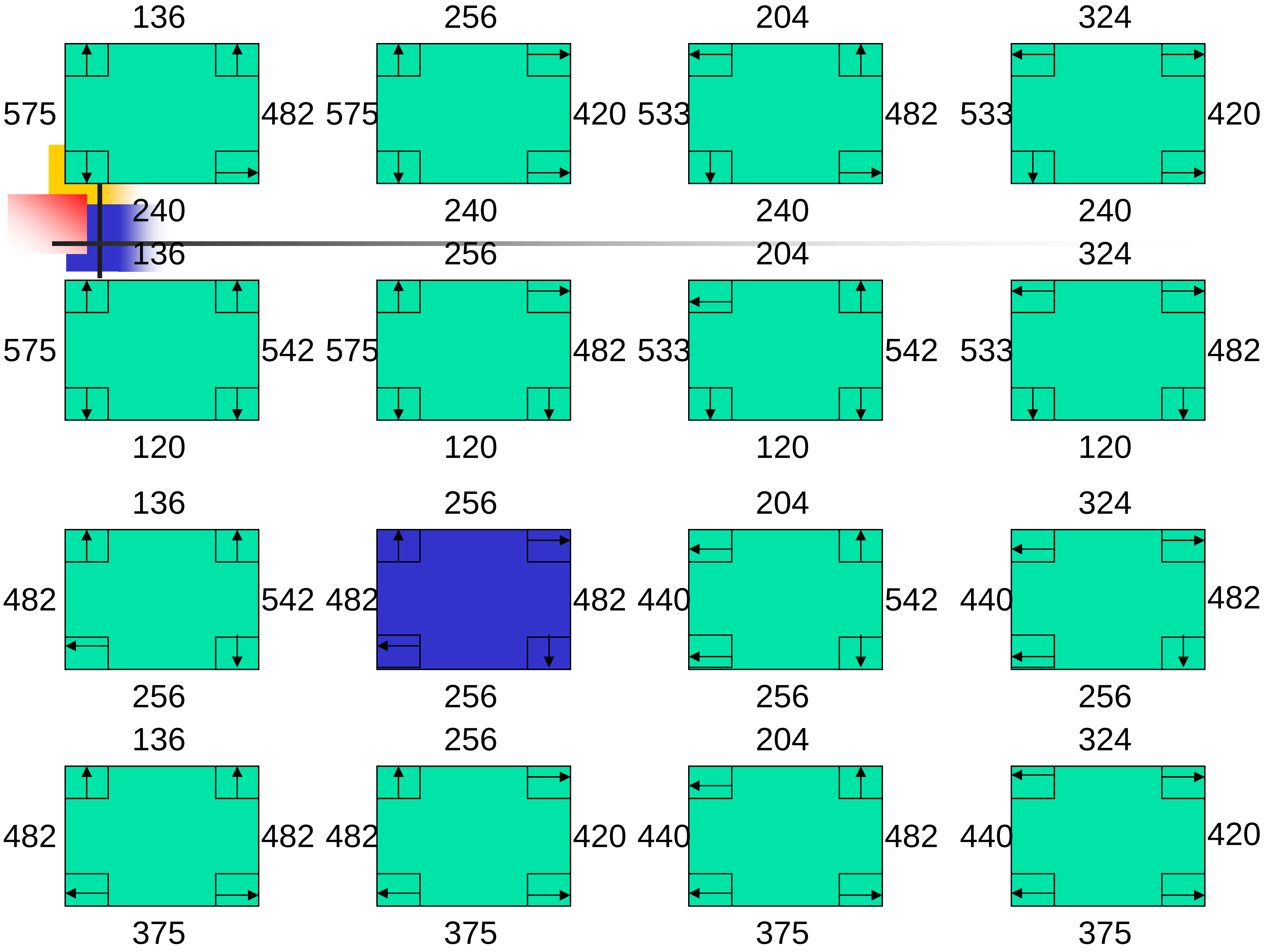
Address-range (traditional) method



Uniform lot-size method

Actual lot-size method

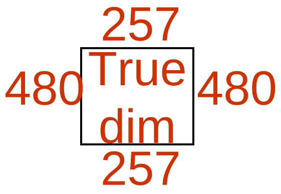
- The corner lot problem motivates us to optimize further
- Palm St, I do worse than traditional approach
- Possible only if the lot sizes available in the Property Tax sites
- Compute the sizes of each of the lots/streets and then run a matching algorithm
- Works on rectangular blocks


136
256
204
324
575
482
575
420
533
482
533
420
240
240
240
240
136
256
204
324
575
542
575
482
533
542
533
482
120
120
120
120
136
256
204
324
482
542
482
482
440
542
440
482
256
256
256
256
136
256
204
324
482
482
482
420
440
482
440
420
375
375
375
375

Finding the optimal layout

- Calculate the actual length and breadth (width) of the block using the information in the street data source

  [*length*, *width*]



257

480 True dim 480

257

Finding the optimal layout

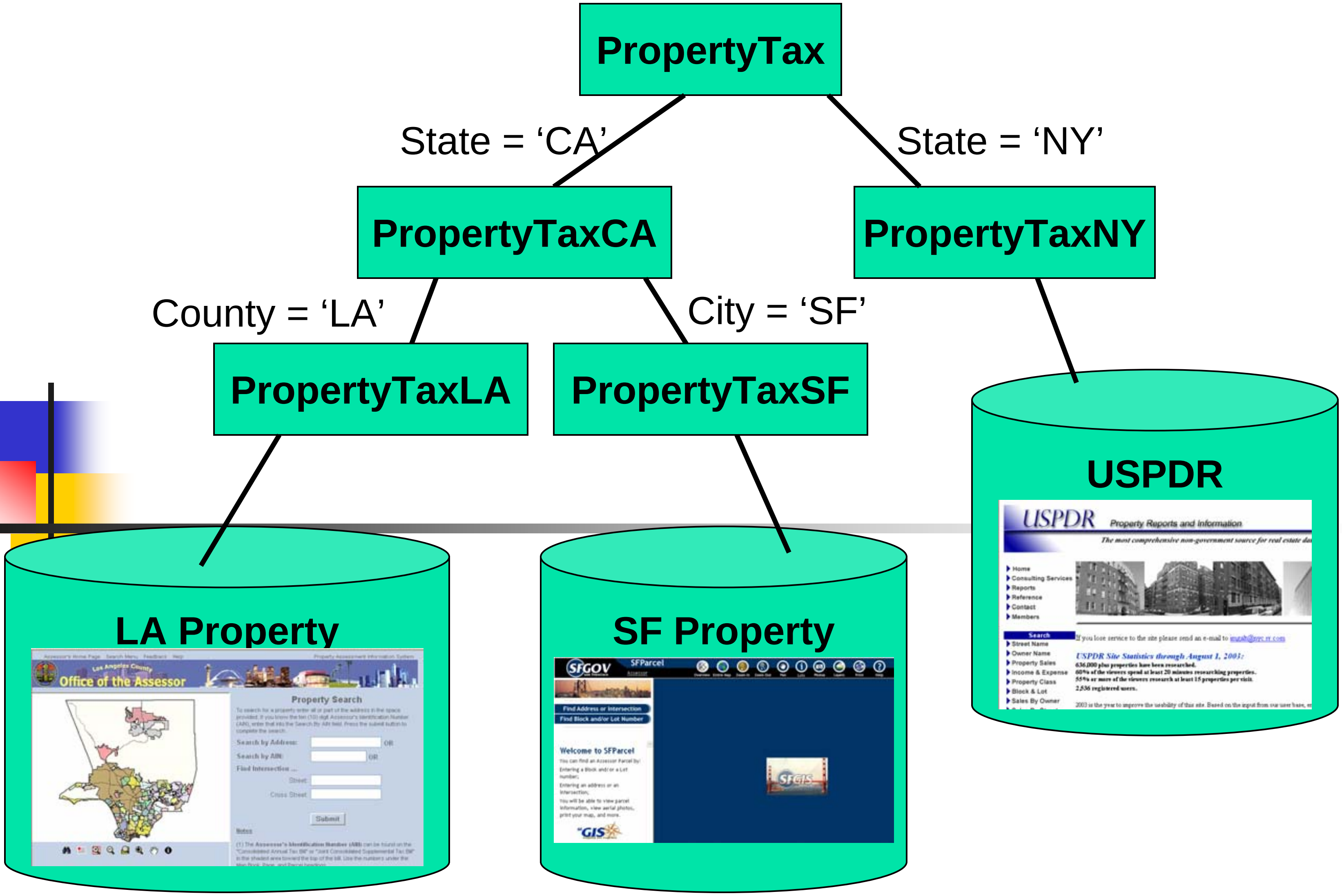
- Get the coordinates of the block from the street data source
- Query the property source and get the dimension of every lot on the block
- Compute the dimensions of the 16 possible orientations
- Compare these with the true dimension
- The layout that most closely matches / least error is chosen as the layout

Integrating data sources

- Unified Query Interface
  - Large number of property sites
  - Query a single relations
- Different property sources for different places
- New York: State, Los Angeles: County
- Disparate representations : structure and attribute names
- Street Data: organized by county or states

Source Descriptions

- Describe the Source as view over Domain description
  - A single property relation
- Three types of Sources
  - Property Tax
  - Property Tax with details of dimensions
  - Street Data Sources



Welcome to SFParcel

You can find an Assessor Parcel by:

Entering a Block and/or a Lot number;

Entering an address or an intersection;

You will be able to view parcel information, view aerial photos, print your map, and more.

The most comprehensive non-government source for real estate da

If you lose service to the site please send an e-mail to

USPDR Site Statistics through August 1, 2003:

636,000 plus properties have been researched.
60% of the viewers spend at least 20 minutes researching properties.
55% or more of the viewers research at least 15 properties per visit.
2,536 registered users.

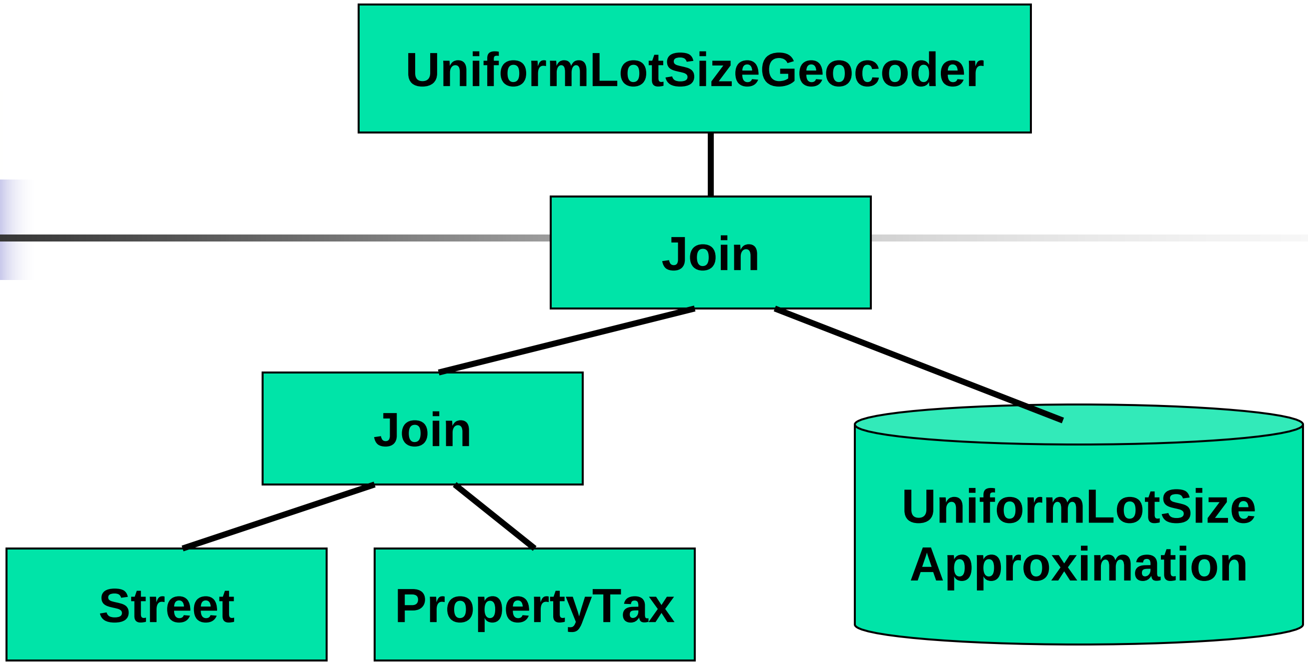
2003 is the year to improve the usability of this site. Based on the input from our user base,

LAProperty(sa, ci, st, zi, fraddr, fraddl, toaddr, toaddl, before, after) :-

PropertyTax(sa, ci, co, st, zi, fraddr, fraddl, toaddr, toaddl, before, after, lotwidth, lotdepth)^

(co = ‘Los Angeles’)^

(st = ‘CA’)



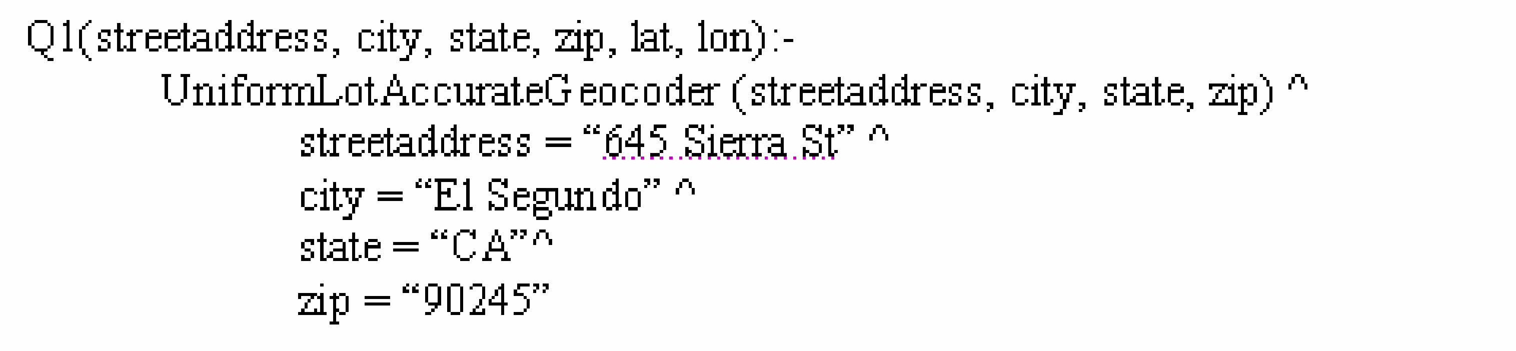
UniformLotSizeGeocoder(sa, ci, co, st, zi, lat, lon):-

Street(sa, ci, co, st, zi, frlat, frlon,tolat, tolon, fename,
fetype, zipl, zipr, fraddr, fraddl, toaddr, toaddl)^

PropertyTax(sa, ci, co, st, zi, fraddr, fraddl, toaddr, toaddl,
before, after,lotwidth, lotdepth)^

UniformLotApproximation(frlat, frlon, tolat, tolon, before,
after, lat, lon)

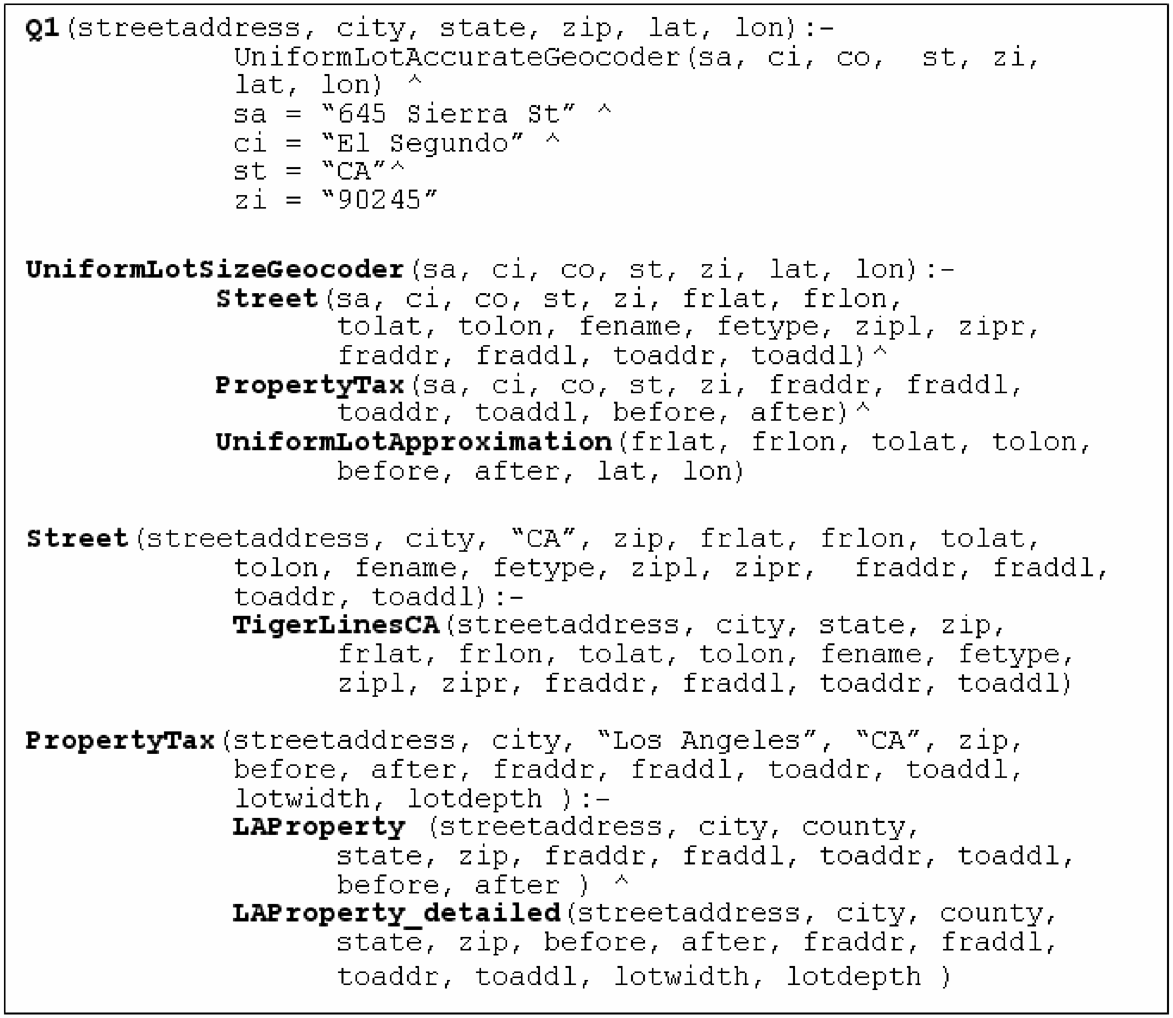
Query



```
Q1(streetaddress, city, state, zip, lat, lon):-
        UniformLotAccurateGeocoder (streetaddress, city, state, zip) ^
                streetaddress = "645 Sierra St" ^
                city = "El Segundo" ^
                state = "CA"^
                zip = "90245"
```

- Inverse the source descriptions
- Generate datalog program to solve the query

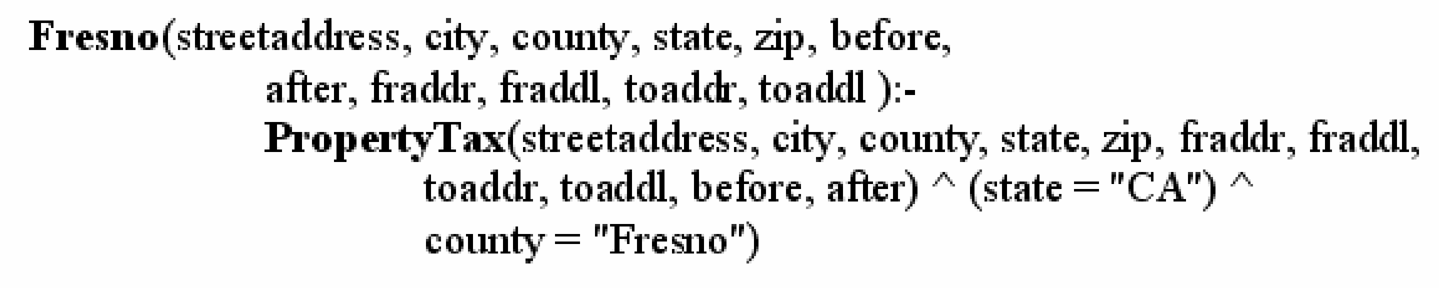
Datalog program generated



```
Q1(streetaddress, city, state, zip, lat, lon):-
            UniformLotAccurateGeocoder(sa, ci, co,  st, zi,
            lat, lon) ^
            sa = "645 Sierra St" ^
            ci = "El Segundo" ^
            st = "CA"^
            zi = "90245"

UniformLotSizeGeocoder(sa, ci, co, st, zi, lat, lon):-
           Street(sa, ci, co, st, zi, frlat, frlon,
                  tolat, tolon, fename, fetype, zipl, zipr,
                  fraddr, fraddl, toaddr, toaddl)^
           PropertyTax(sa, ci, co, st, zi, fraddr, fraddl,
                  toaddr, toaddl, before, after)^
           UniformLotApproximation(frlat, frlon, tolat, tolon,
                  before, after, lat, lon)

Street(streetaddress, city, "CA", zip, frlat, frlon, tolat,
            tolon, fename, fetype, zipl, zipr,  fraddr, fraddl,
            toaddr, toaddl):-
            TigerLinesCA(streetaddress, city, state, zip,
                  frlat, frlon, tolat, tolon, fename, fetype,
                  zipl, zipr, fraddr, fraddl, toaddr, toaddl)

PropertyTax(streetaddress, city, "Los Angeles", "CA", zip,
            before, after, fraddr, fraddl, toaddr, toaddl,
            lotwidth, lotdepth ):-
            LAProperty (streetaddress, city, county,
                  state, zip, fraddr, fraddl, toaddr, toaddl,
                  before, after ) ^
            LAProperty_detailed(streetaddress, city, county,
                  state, zip, before, after, fraddr, fraddl,
                  toaddr, toaddl, lotwidth, lotdepth )
```

Advantage of this model

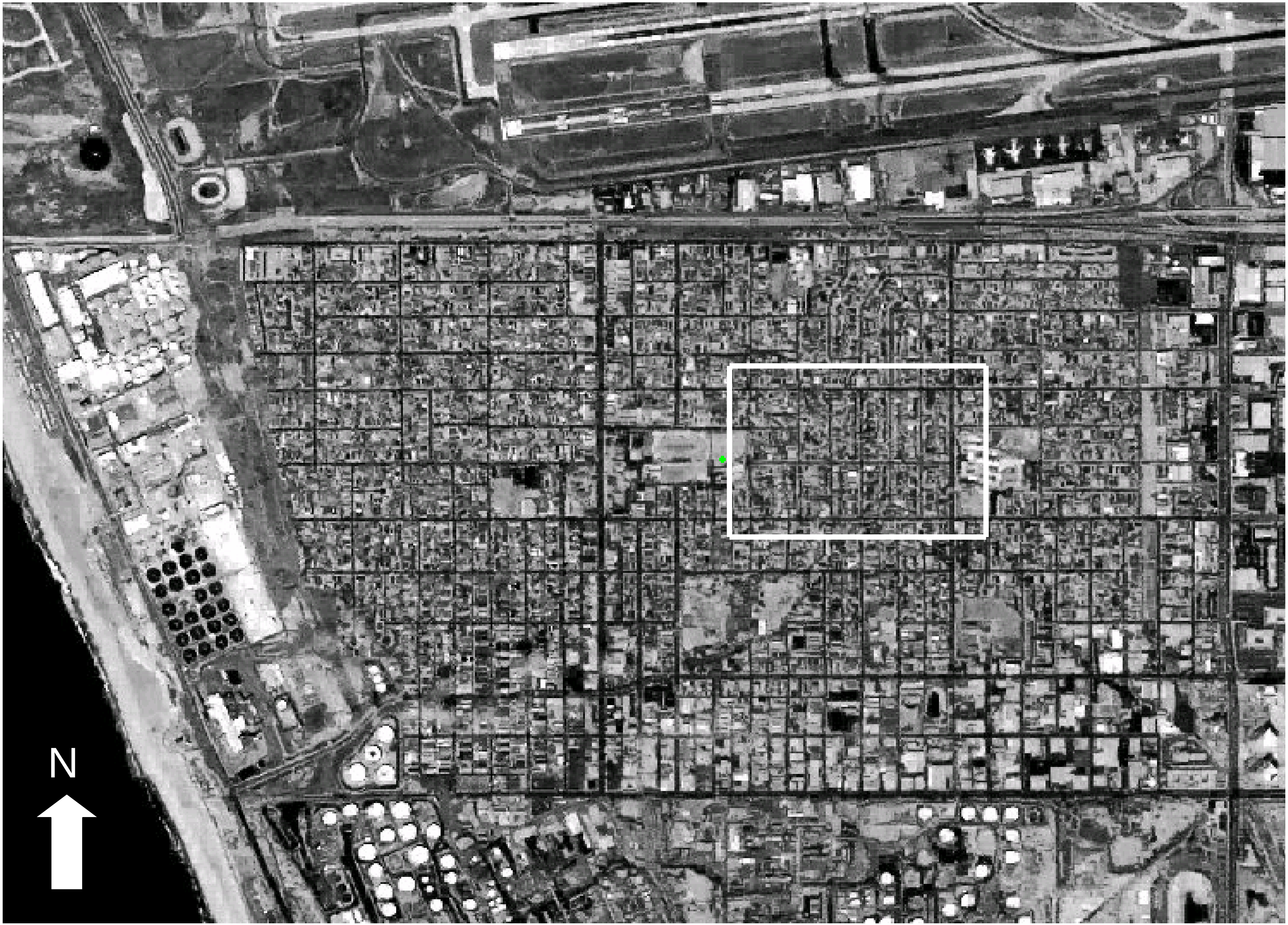
- GLAV (Global-Local as View)
- Easy to add new sources



Fresno(streetaddress, city, county, state, zip, before,
after, fraddr, fraddl, toaddr, toaddl):-
PropertyTax(streetaddress, city, county, state, zip, fraddr, fraddl,
toaddr, toaddl, before, after) ^ (state = "CA") ^
county = "Fresno")

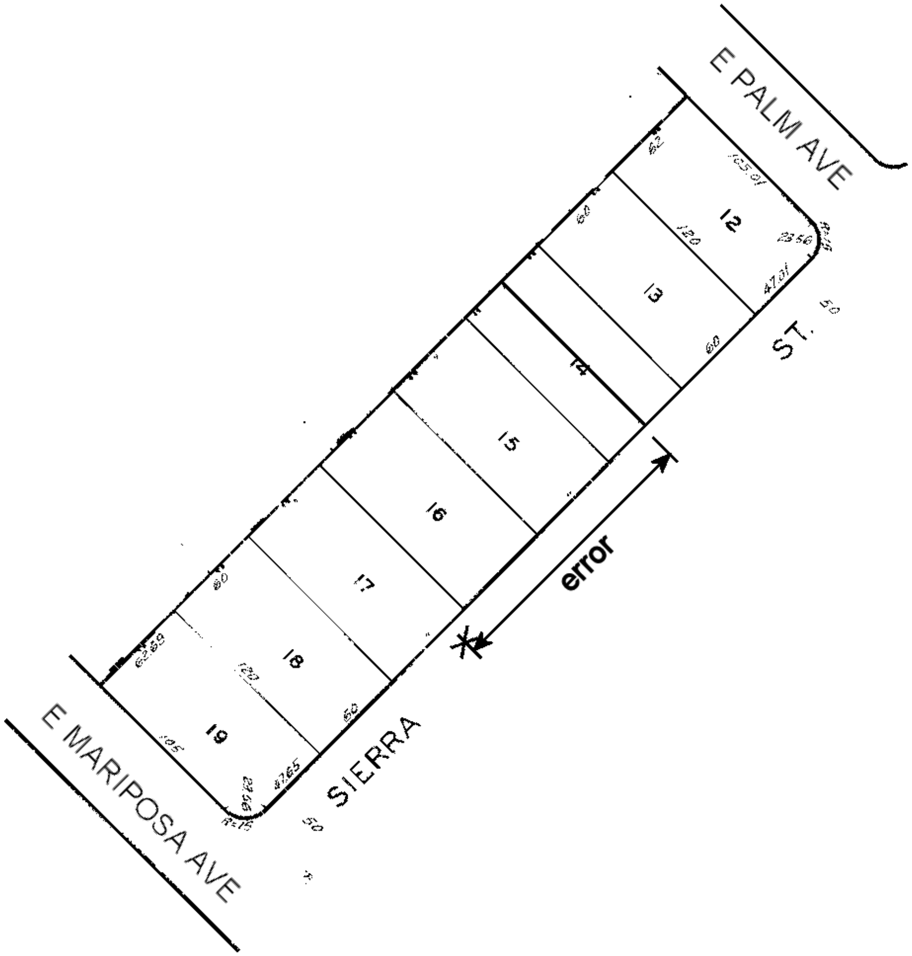
Results

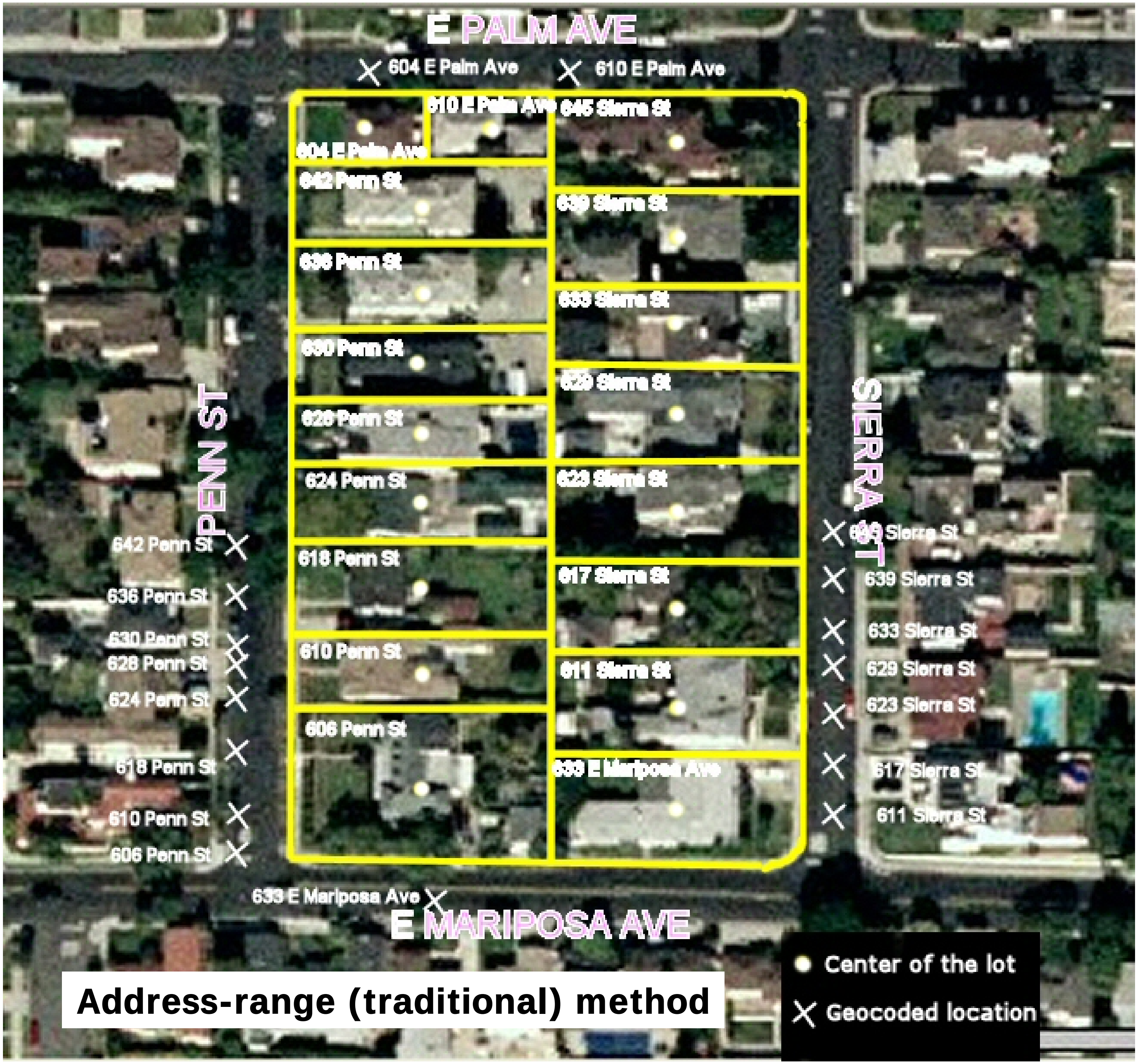
- Chosing a region
  - El Segundo
- Data Source
  - Conflated TIGER/Lines
- Fetch Agent Platform to convert website data into XML
- Prometheus 2.0 information mediator
- Geocoded 267 addresses spanning 13 blocks
- Actual lot-size method could not be applied to 58 addresses
- None of the methods could be applied to one address
- Results based on the remaining 208 addresses



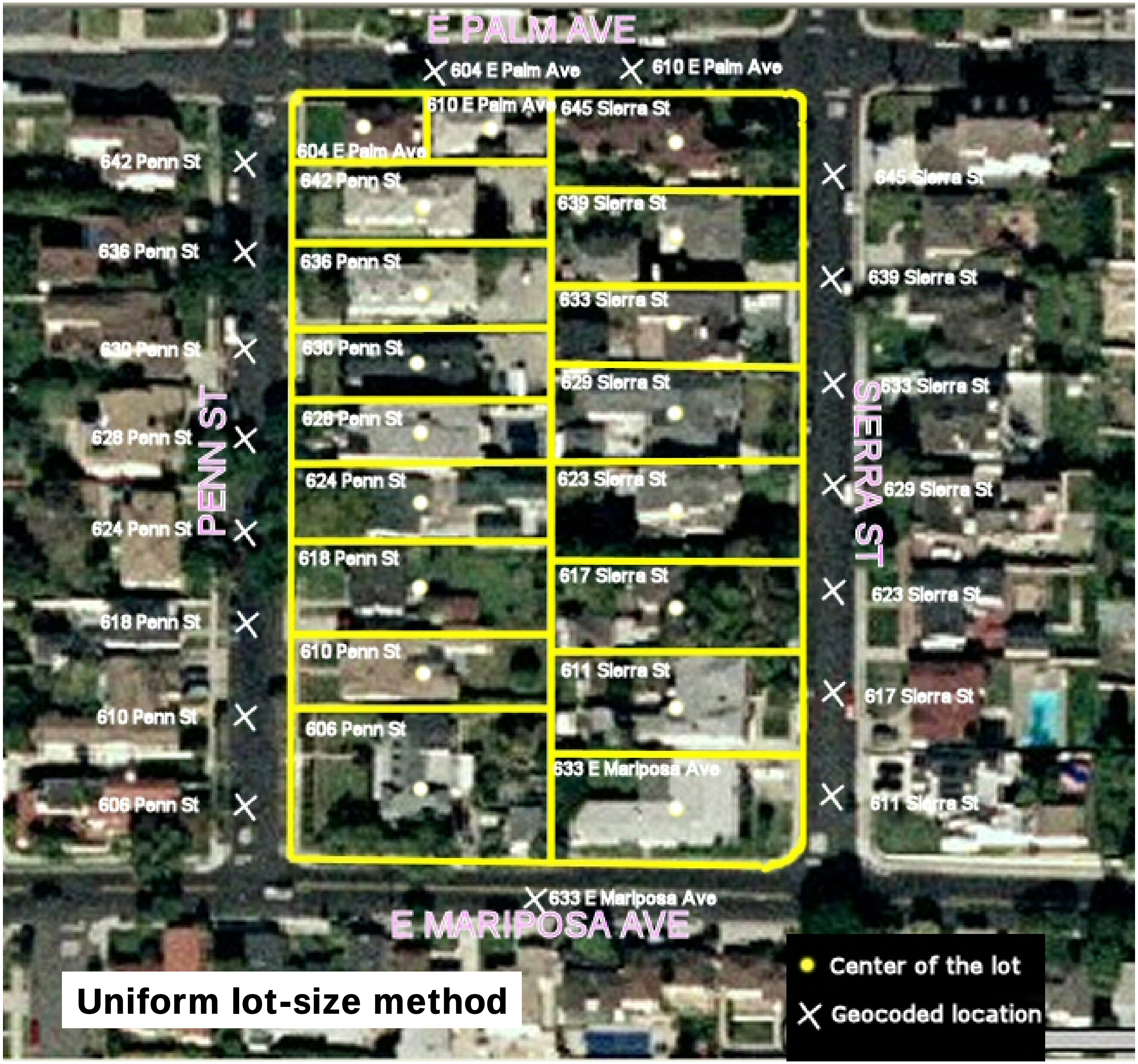
Chosen area for goecoding

Driving distance

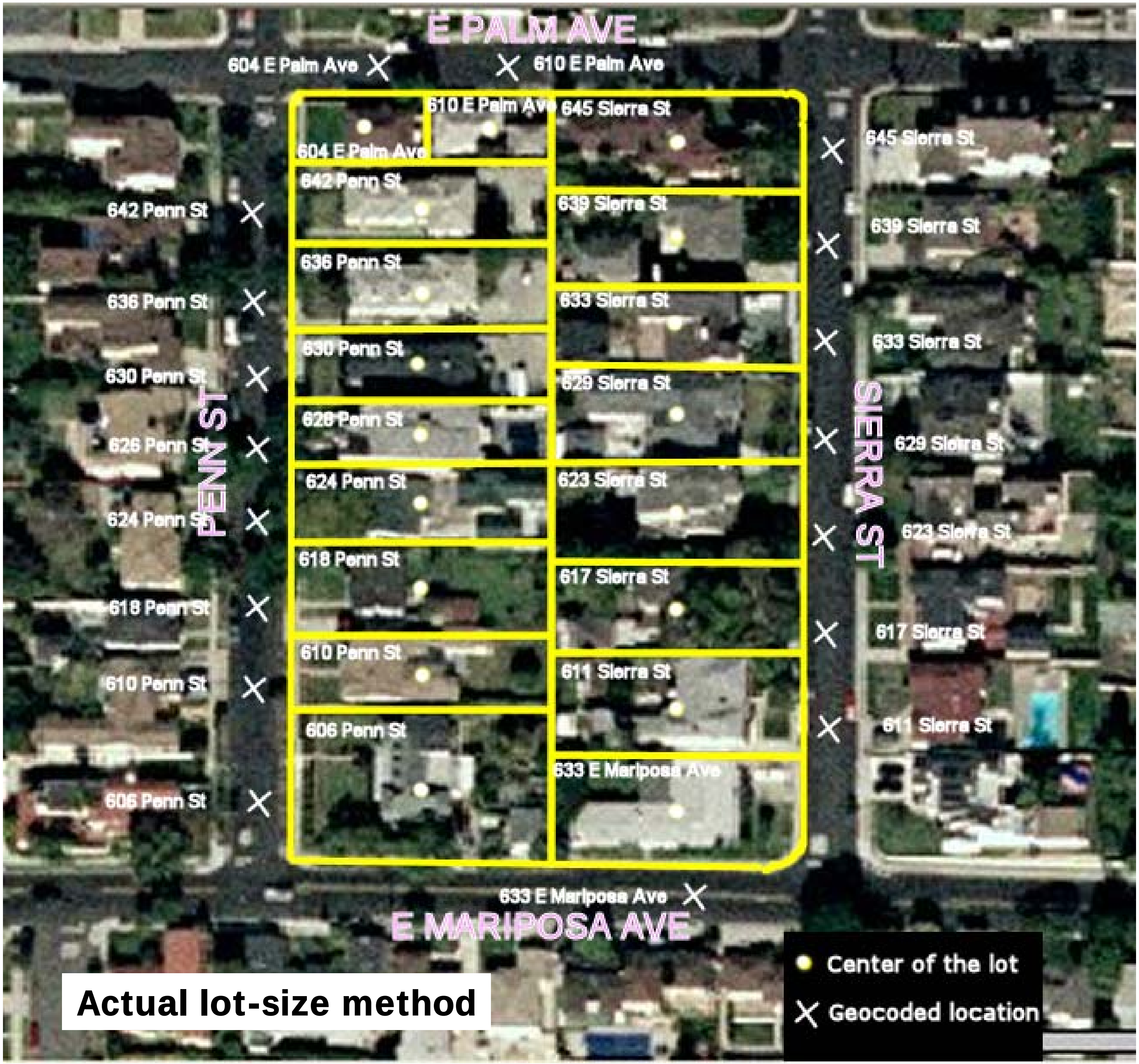




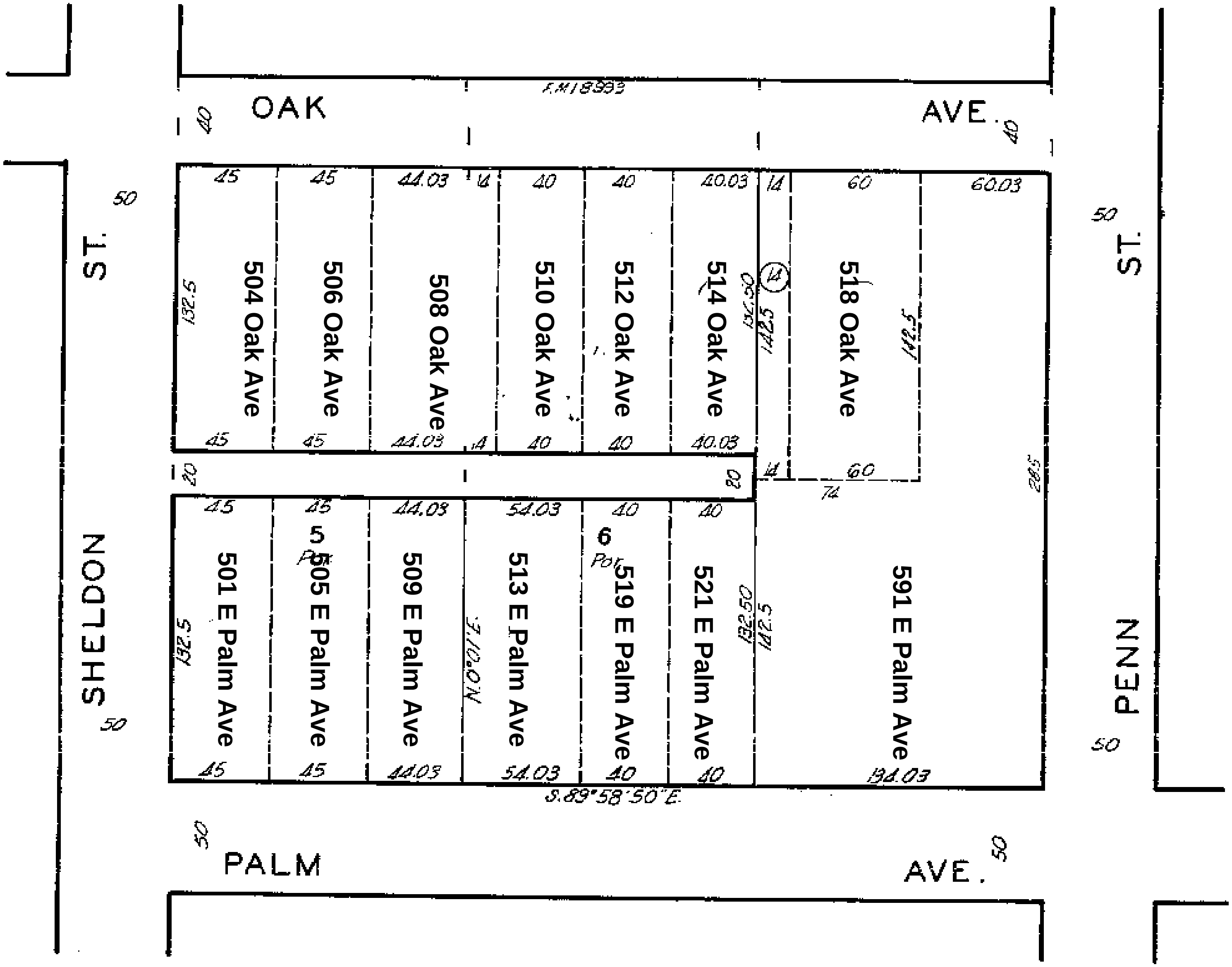
Address-range (traditional) method

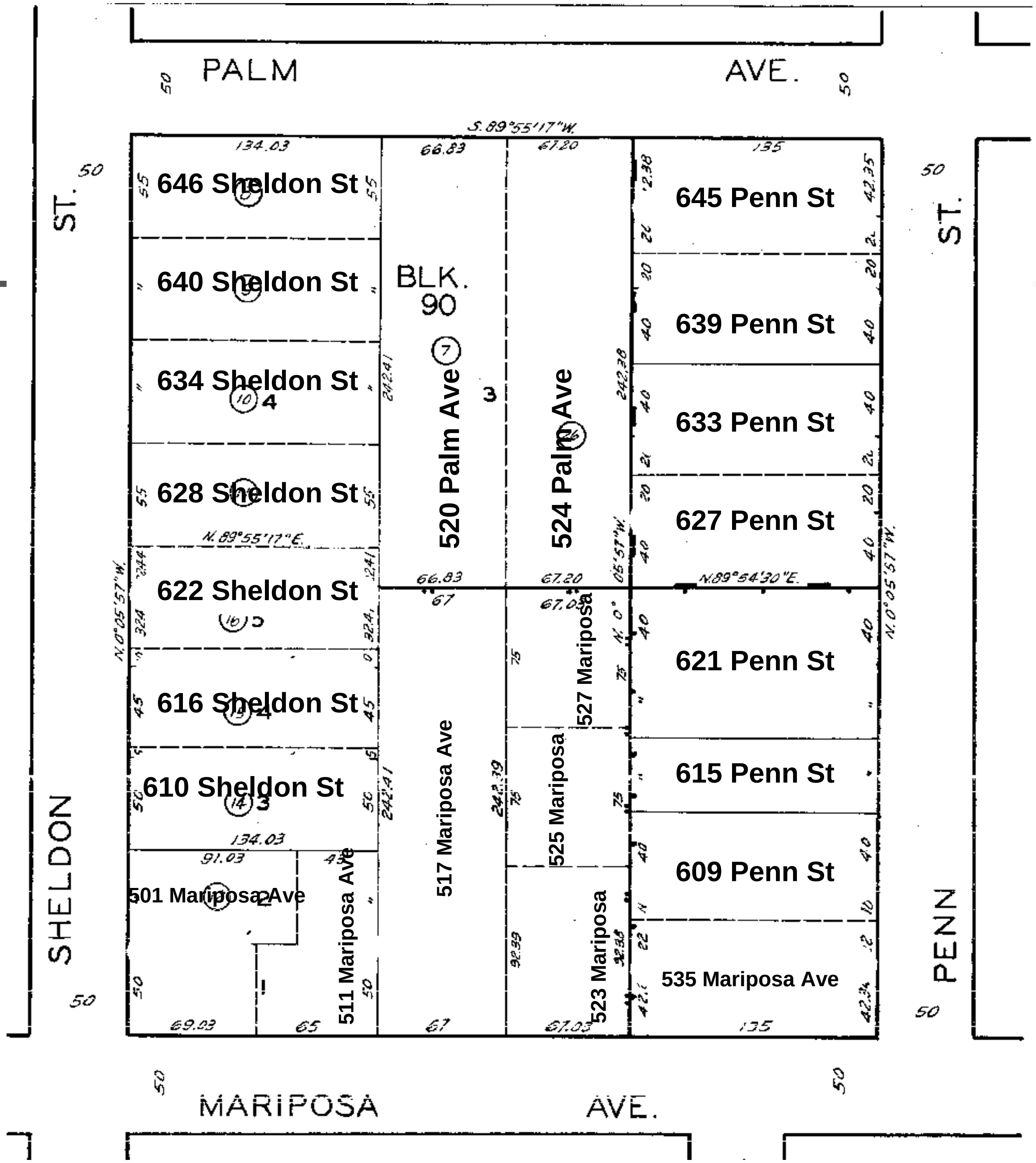


Uniform lot-size method



Actual lot-size method

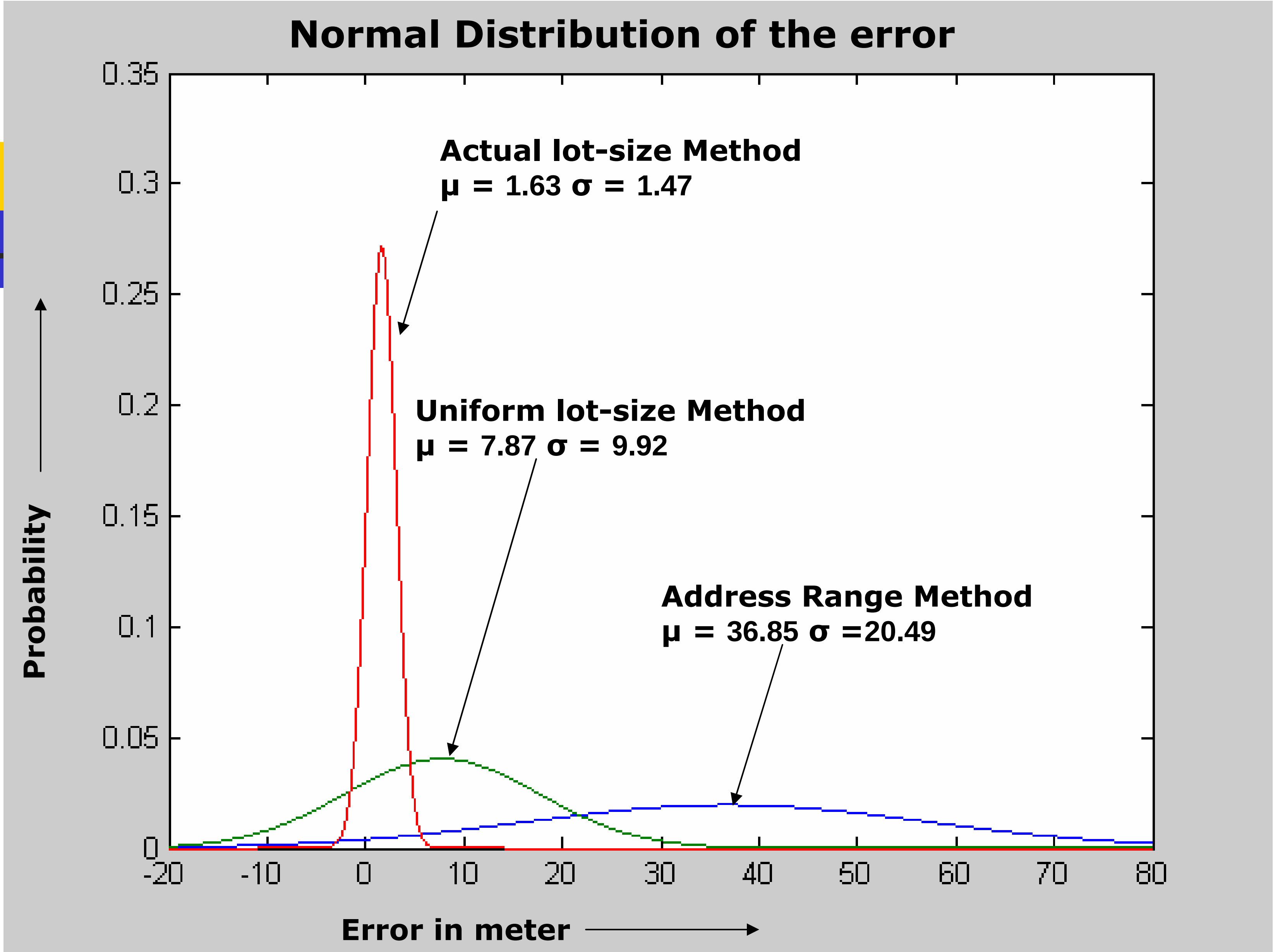

OAK AVE.
F.M.18993
40
SHELDON ST.
PENN ST.
PALM AVE.
50
504 Oak Ave
506 Oak Ave
508 Oak Ave
510 Oak Ave
512 Oak Ave
514 Oak Ave
518 Oak Ave
501 E Palm Ave
505 E Palm Ave
509 E Palm Ave
513 E Palm Ave
519 E Palm Ave
521 E Palm Ave
591 E Palm Ave
5
Por.
6
Por.
14
45
44.03
40
40.03
60
60.03
132.5
136.50
142.5
20
74
285
54.03
132.50
134.03
N.0°01'E.
S.89°58'50"E.


PALM AVE.
MARIPOSA AVE.
SHELDON ST.
PENN ST.
BLK. 90
S.89°55'17"W.
N.89°55'17"E.
N.89°54'30"E.
N.0°05'57"W.
646 Sheldon St
640 Sheldon St
634 Sheldon St
628 Sheldon St
622 Sheldon St
616 Sheldon St
610 Sheldon St
501 Mariposa Ave
511 Mariposa Ave
517 Mariposa Ave
520 Palm Ave
524 Palm Ave
527 Mariposa
525 Mariposa
523 Mariposa
645 Penn St
639 Penn St
633 Penn St
627 Penn St
621 Penn St
615 Penn St
609 Penn St
535 Mariposa Ave
50
134.03
66.83
67.20
135
242.41
242.38
91.03
69.03
65
67
67.03

Comparison of Results

(all errors are in meters)	**Address-range**	**Uniform lot-size**	**Actual lot-size**
Average Error	36.85359	7.87149	1.62993
Standard Deviation	20.49335	9.92361	1.46958
Minimum Error	0.86578	0.07086	0.03487
Maximum Error	73.80526	56.64072	7.80242

- Average percentage of improvement over traditional approach
  - Uniform lot-size method: 78.65%
  - Actual lot-size method: 95.59%


Normal Distribution of the error
Actual lot-size Method
μ = 1.63 σ = 1.47
Uniform lot-size Method
μ = 7.87 σ = 9.92
Address Range Method
μ = 36.85 σ =20.49
Probability
Error in meter
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-20
-10
0
10
20
30
40
50
60
70
80

Related Work

- Cayo, M. R. and T. O. Talbot (2003) Positional error in automated geocoding of residential addresses
- Ratcliffe (2001) On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units
- Krieger et al. (2001) Evaluating the accuracy of geocoding in public health research
- Gupta, Marciano et al.(1999) Integrating GIS and Imagery through XML-Based Information Mediation

Conclusion & Future Work

- More accurate geocoding achieved
- Integrating other sources to get property data
- Solved the address-validating problem
- Extend the actual lot size method to non-rectangular blocks
- Integrate more property tax data sources

Acknowledgements

- Thanks to Craig for his valuable guidance, Snehal for help with the algorithms and implementation,

  Shou-de for the calculations in the actual lot size method
- Thanks to Cyrus Shahabi and John Wilson

Questions / Comments