Reverse geocoding and implications for geospatial privacy
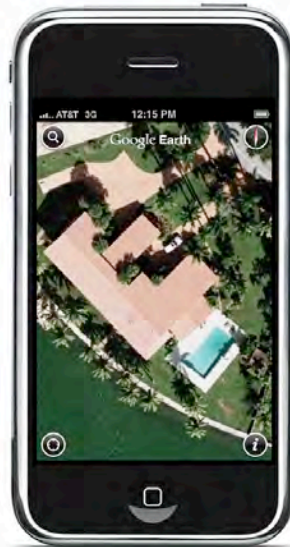
Paul Zandbergen

Department of Geography
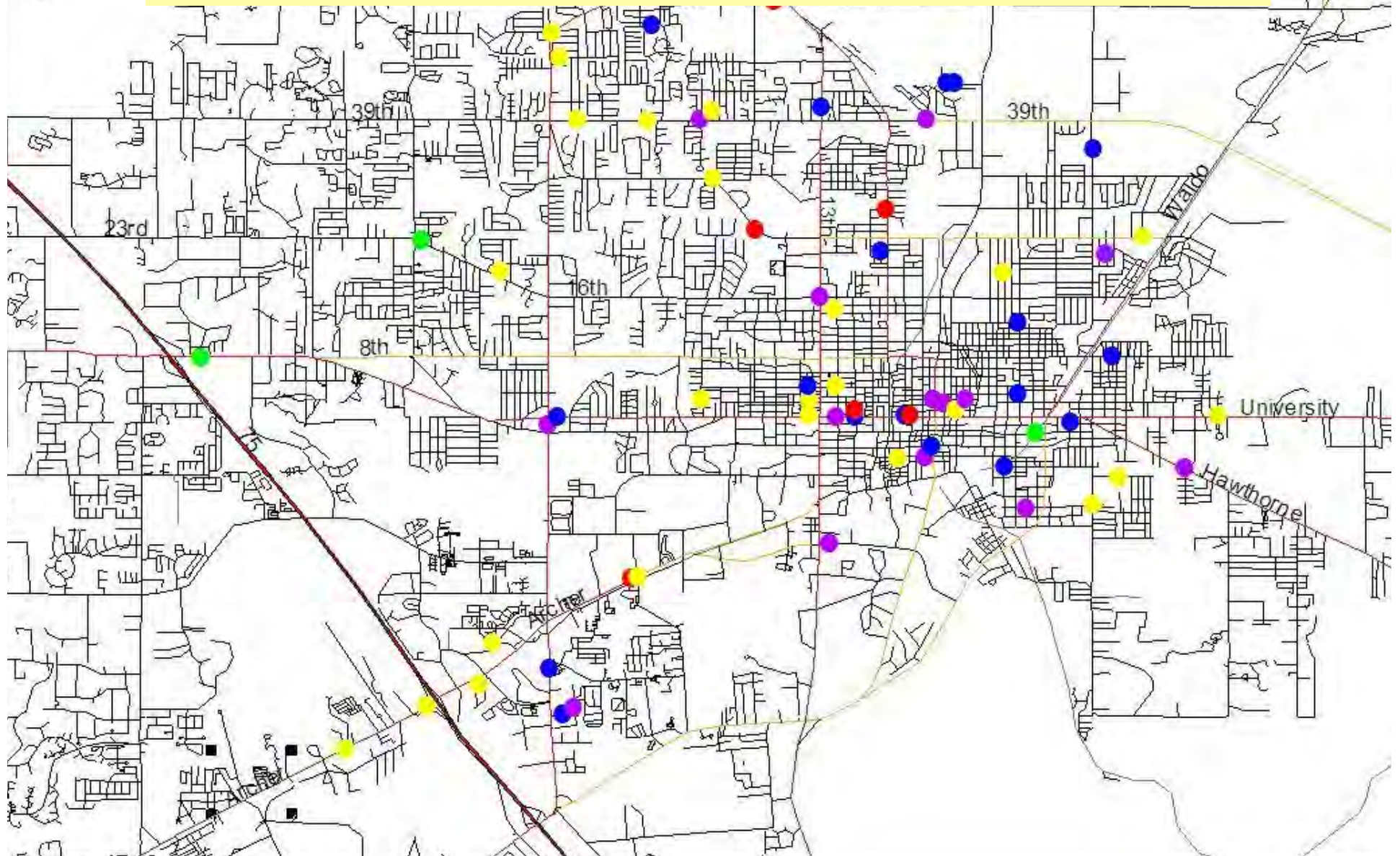
University of New Mexico

Outline

- Geospatial privacy
- Geocoding / reverse geocoding
- Experimental design
- Results and Conclusion

Geospatial Privacy

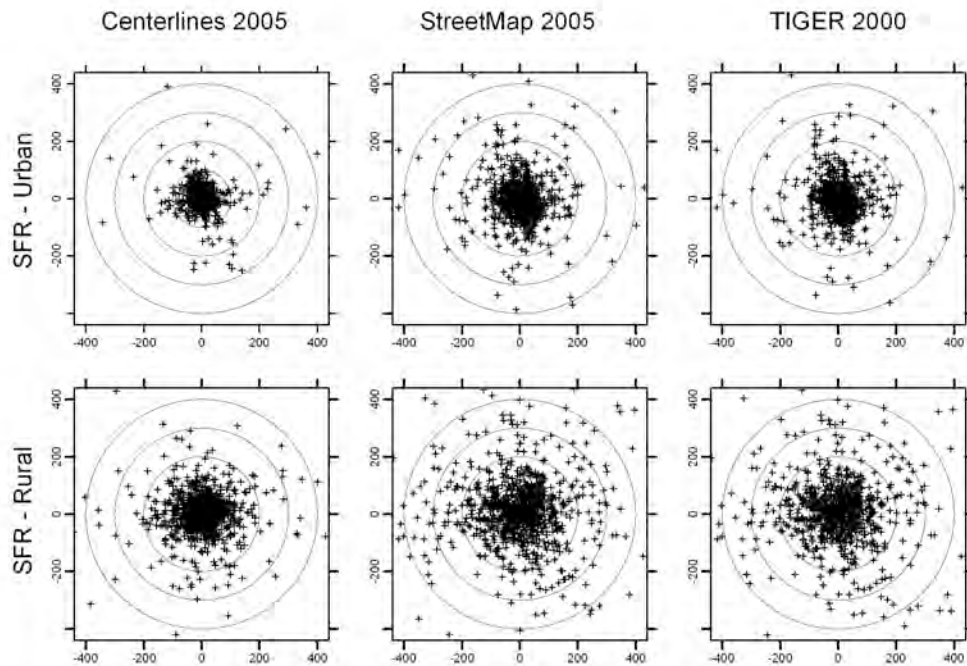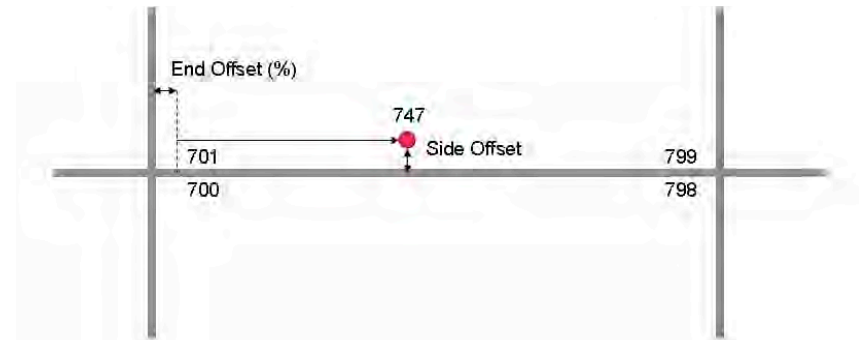- Under increasing threat from new technologies, high resolution data and easy-to-use tools

How easy is it to hack this map?

Geocoding

- Widely employed
- Well understood
- Substantial errors

Tool of a hacker: Reverse geocoding
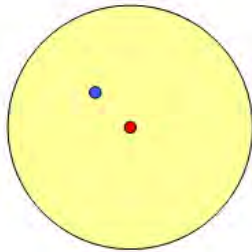
- Geocoding in reverse
- Relative easy, relatively new
- Key tool for "hacking" published maps
- Not well understood

ID	Address	ZIP
101	123 Main St	12345
102	456 Central Ave	12346
…	…	…

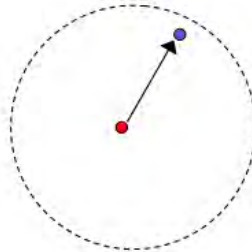# How to protect spatial confidentiality?

- Geographic masking

- But how to do this most effectively?

- Need better understanding of reverse geocoding

| random point within circle | random point on circle | random distance and direction | donut masking | donut masking with exclusion |
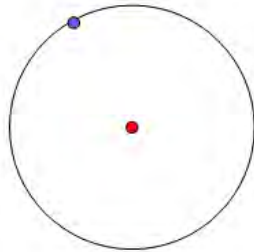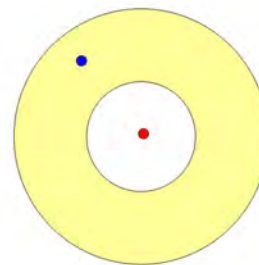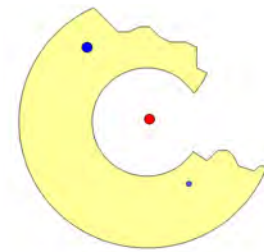
Review of the State of the Art

PUTTING PEOPLE ON THE MAP

PROTECTING CONFIDENTIALITY WITH LINKED SOCIAL-SPATIAL DATA

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES
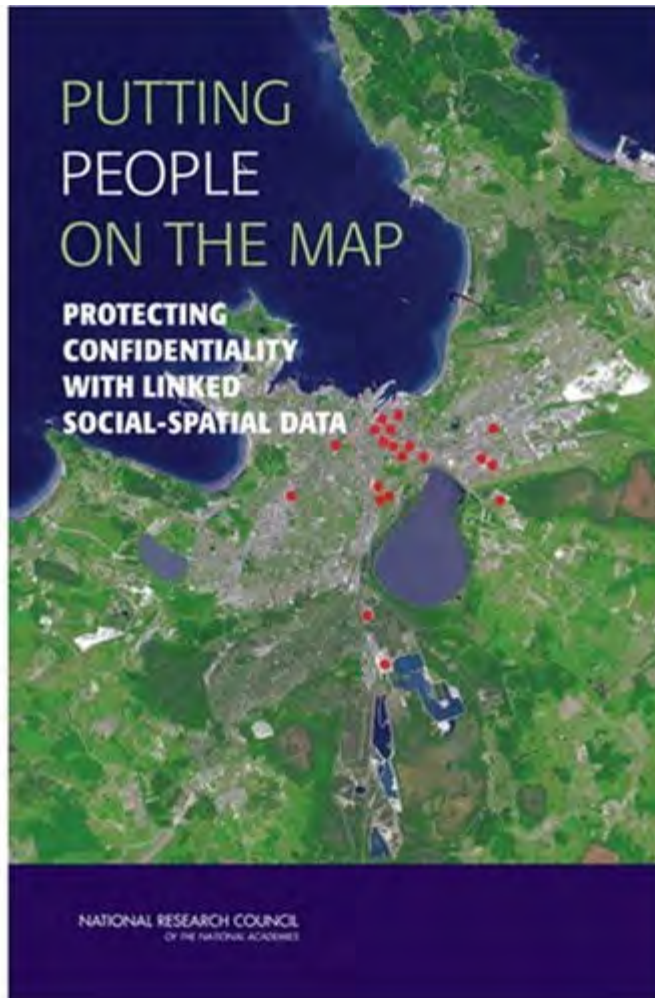
Panel of confidentiality issues arising from the integration of remotely sensed and self-identifying data

"No known technical strategy [….] for managing linked spatial-social data adequately resolves conflicts among the objectives of data linkage, open access, data quality, and confidentiality protection across datasets and across uses." (Conclusion 3)

National Research Council, 2007

Research Questions

What are the capabilities of reverse geocoding to identify individuals from published locations?

How does this vary with the methods employed for geocoding and reverse geocoding?

How does this vary with population density?

Experimental Design

- Actual building locations with known addresses
  - Travis County, TX (Austin)
  - Sample of 2,500 residential locations
  - Stratified across 5 population density classes
- Geocode using 5 different geocoders
- Reverse geocode using 3 different techniques
- Determine accuracy of reverse geocoding

Address Points
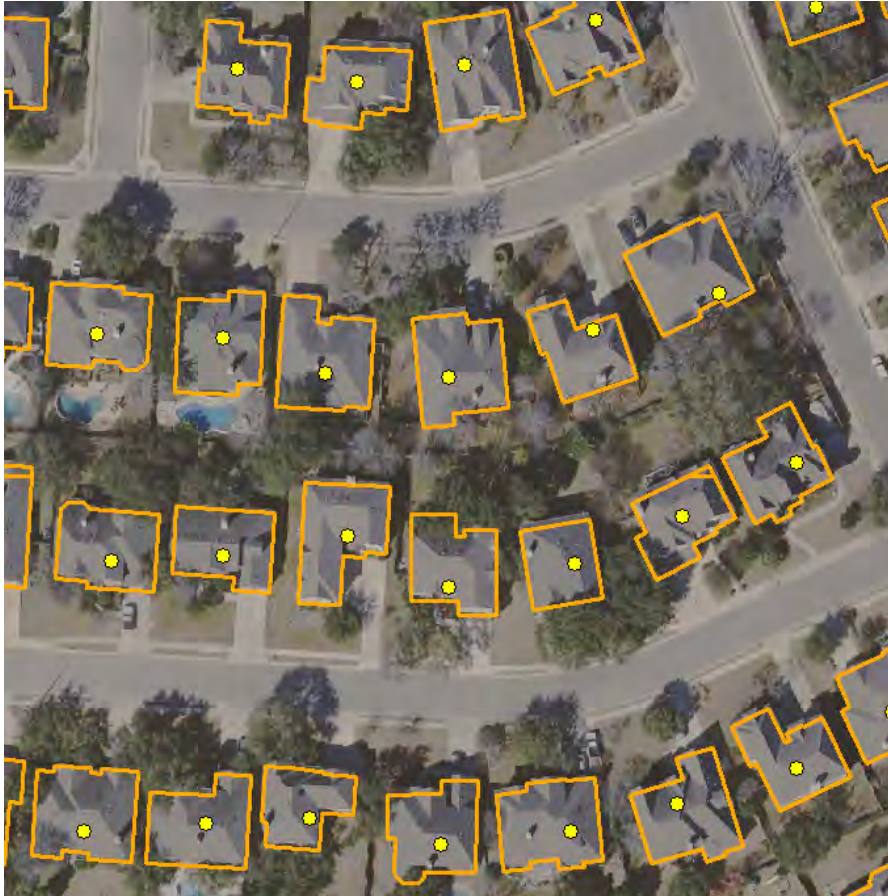
Geocoders

- TeleAtlas Address Points (commercial)
- TeleAtlas Streets (commercial)
- Google Maps (free API)
- StreetMap USA Pro 2007 in ArcGIS
- Geolytics 2007 (using TIGER 2007data)

Study Area – Travis County, TX

Population Density Zones
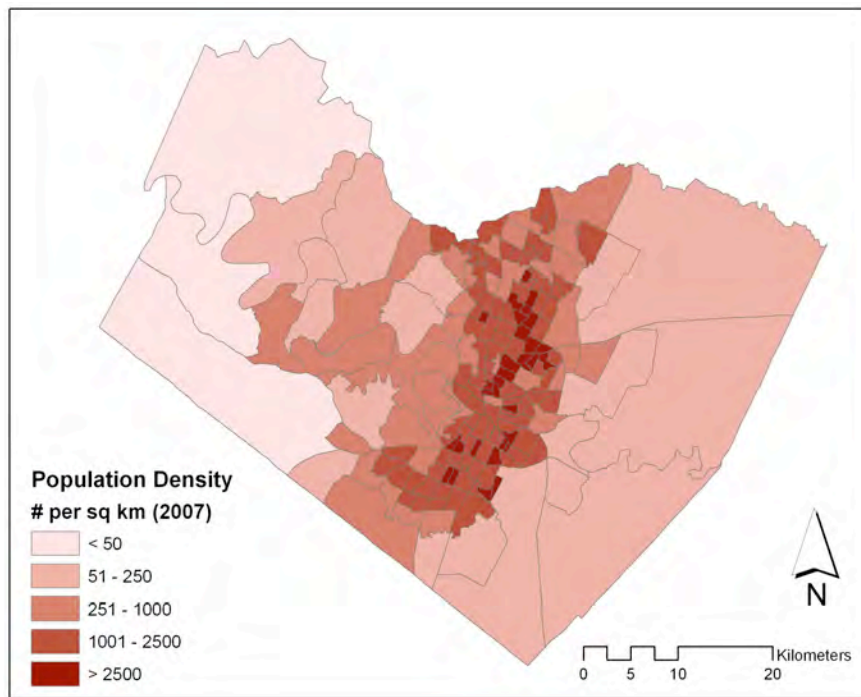
Sample of residential address points

Reverse Geocoding & Accuracy

- Original residential address points
  - Snap to nearest residential building
- TeleAtlas reverse street geocoding
  - Submit for commercial processing
- Google Maps reverse geocoding
  - Free API

- Accuracy of reverse matches
  1. Perfect match (street name and number)
  2. Close match (number within 10)
  3. Same street only

Results – Match Rates (%)

Geocoding Technique	Population density (people/km^2)					
	< 50	50 to 250	250 to 1000	1000 to 2500	> 2500	Total
TeleAtlas AP	43.6	70.2	91.8	92.8	93.6	78.4
TeleAtlas Street	93.2	92.8	98.0	99.8	96.8	96.1
Google Maps	92.2	95.6	98.2	99.0	96.2	96.2
StreetMap Pro	81.2	83.6	95.8	99.0	95.8	91.1
Geolytics	77.0	80.2	92.4	96.2	90.2	87.2
Combined	31.4	59.0	86.0	89.6	86.8	70.6

Results – Same Street Matches

	Reverse Geocoding Technique		
	Austin AP	Google Maps	TeleAtlas Street
Austin AP	100.0	96.7	90.1
TeleAtlas AP	99.5	92.0	89.6
Google Maps	99.2	32.0	90.5
TeleAtlas Street	88.2	92.5	99.5
StreetMap Pro	89.1	76.8	82.5
Geolytics	54.5	54.7	56.9

Geocoding Technique

Results – Close Reverse Matches

	Reverse Geocoding Technique		
	Austin AP	Google Maps	TeleAtlas Street
Austin AP	100.0	95.5	23.1
TeleAtlas AP	99.1	91.8	24.3
Google Maps	98.8	28.6	24.3
TeleAtlas Street	69.8	63.2	92.5
StreetMap Pro	60.8	42.9	44.0
Geolytics	33.8	27.4	29.3

Geocoding Technique

Results – Perfect Reverse Matches

	Reverse Geocoding Technique		
	Austin AP	Google Maps	TeleAtlas Street
Austin AP	100.0	94.2	9.1
TeleAtlas AP	97.9	91.8	9.0
Google Maps	97.2	27.8	9.1
TeleAtlas Street	18.7	9.4	55.0
StreetMap Pro	17.0	7.3	8.2
Geolytics	7.1	2.3	2.6

Geocoding Technique

Effect of Population Density – Percent Perfect Matches

Geocoding	Reverse	Population density (people/km^2)				
		< 50	50 to 250	250 to 1000	1000 to 2500	> 2500
StreetMap Pro	Austin AP	22.9	16.9	15.3	16.1	17.3
Geolytics	Austin AP	14.0	7.1	7.0	5.4	6.5
Austin AP	TeleAtlas Street	2.5	6.8	11.6	11.4	8.3
TeleAtlas Street	TeleAtlas Street	51.6	63.7	56.7	53.8	49.8
Google Maps	TeleAtlas Street	4.5	6.4	12.3	11.2	7.4
TeleAtlas Street	Google Maps	9.6	9.8	8.8	9.6	9.4
Geolytics	Google Maps	1.9	2.0	2.6	2.5	2.3

Results Summary

- Accuracy of reverse geocoding varies greatly
- Building level (reverse) geocoding is typically most accurate
- Street geocoding is quite noisy
    - Easy to get the right street
    - Very few perfect matches
- Accuracy is substantially improved if knowledge of the original geocoding technique is available!
- No clear pattern with population density

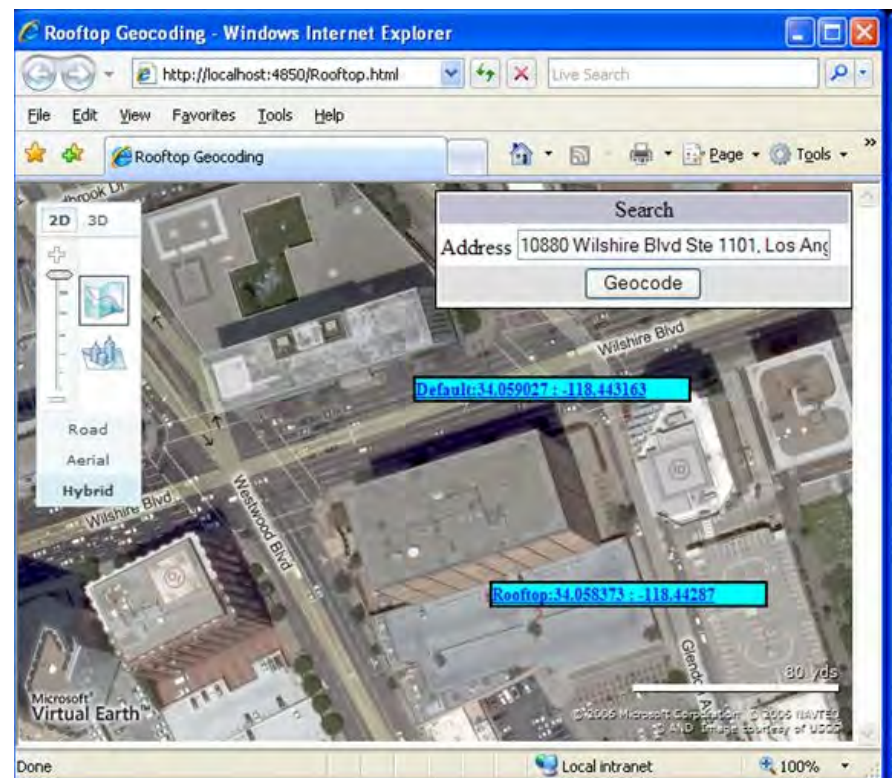# Rooftop Geocoding in Google Maps and Virtual Earth

Commercial Address Points - TeleAtlas

51 million address points in the US

no coverage
moderate coverage
good coverage

Source: TeleAtlas, 2009
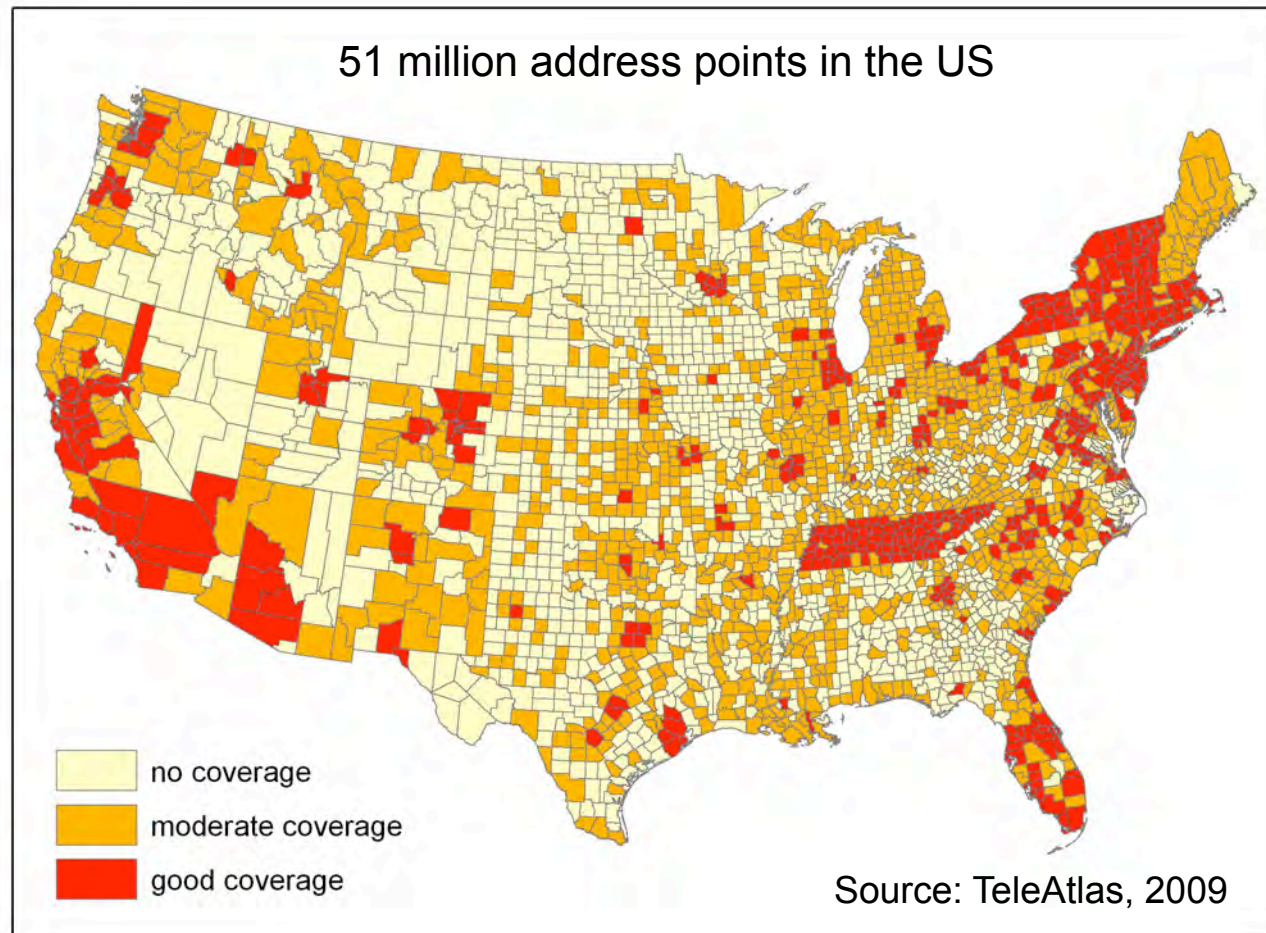
Licensed to Google, Virtual Earth, ArcGIS Business Analyst, Pitney Bowes / Group 1 / MapInfo

Reverse Geocoding

Reverse Geocoding Example using the GReverseGeocder v1.0.7
The example below uses the getPlacemarkProperty function to get the "PostalCode"
Copyright 2008, Nico Goeminne
nicogoeminne at gmail.com

Latitude (WGS84)

Longitude (WGS84)

Get Address

Reverse geocoding now supported in Google Maps and Microsoft Virtual Earth

Also supported in latest version of ArcGIS – requires some customization or ArcWeb services

Numerous free easy-to-use online utilities

Conclusions

- Accuracy of reverse geocoding
  - Varies greatly with geocoder / reverse geocoder combination
  - Between 2% and 98% perfect reverse matches
- Knowledge of original geocoding method is critical
  - noisy results from street geocoding can be reverse coded
- Trends:
  - Address points are the new standard in geocoding
  - Reverse geocoding is relatively easy

- Techniques to protect privacy may need to assume a worst-case scenario: very high resolution address data

Future Research

- Replicate in other study areas
- Examine urban/rural gradients more closely
- Experiment with different masking techniques
- Develop a framework for spatial κ-anonymity

Acknowledgements

- National Science Foundation
- UNM Research Allocation Committee
- American Civil Liberties Union