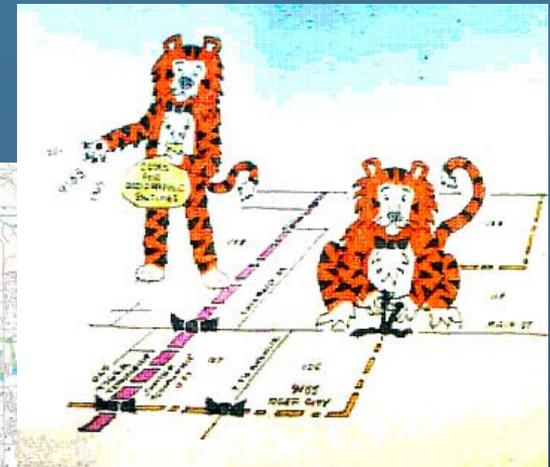
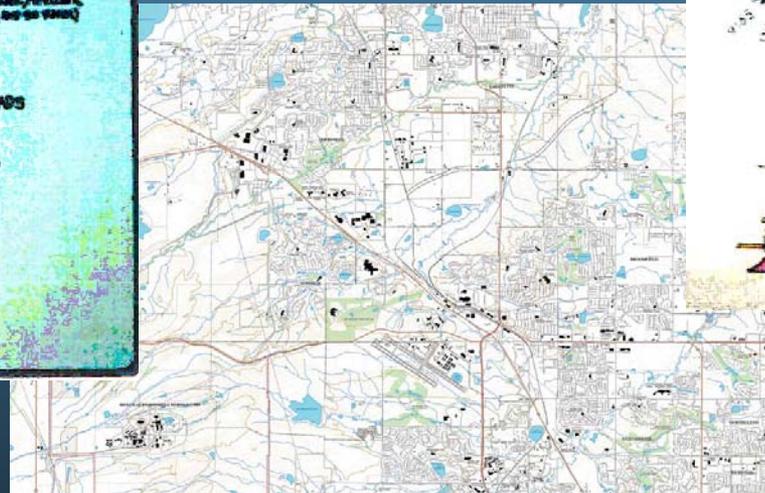
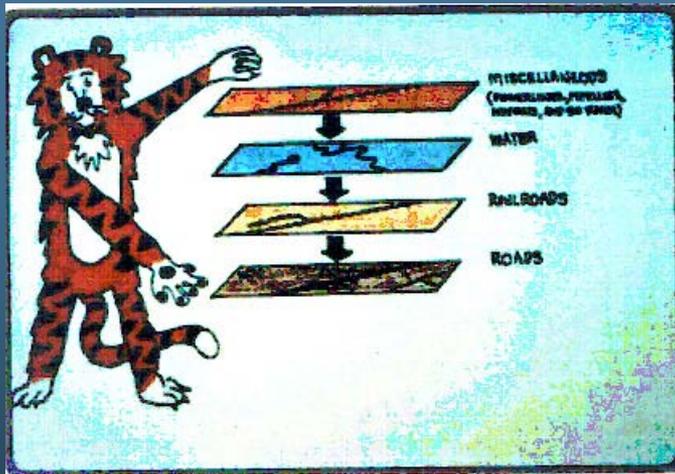


Quantifying the Quality of the MAF/TIGER Data Base: An Exploration

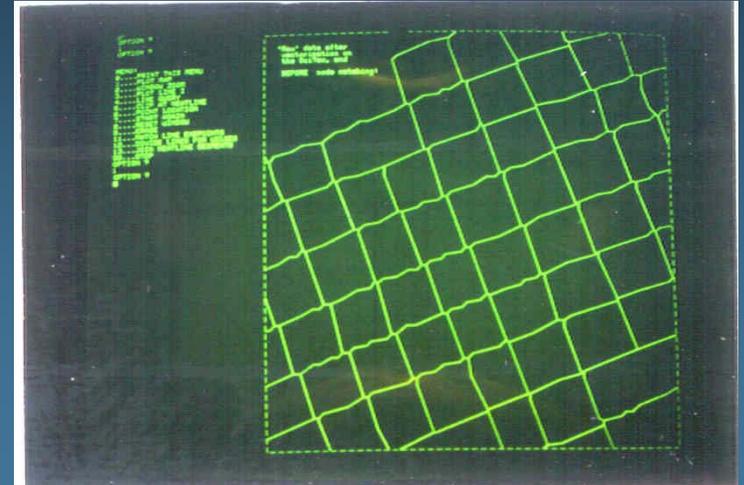
Stephen C. Guptill

June 9, 2011



GIS in 1970's & 1980's

- Data quality not a major concern
- Attention was on getting technology to work
- Accurate replication of analog maps = good data quality

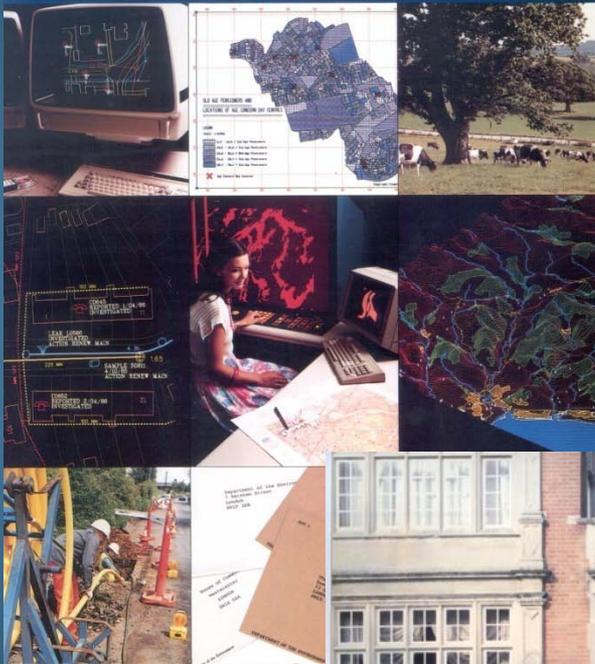


The Rise of GIS and Geospatial Data

Department of the Environment

Handling Geographic Information

Report of the Committee of Enquiry chaired by Lord Chorley



AUTO CARTO LONDON



edited by Michael Blakemore

OMB 3145-0058

P.T. 34.04
K.W. 0406000



Directorate for Biological, Behavioral,
and Social Sciences
Division of Social and Economic Science
Washington, D.C. 20550

Solicitation

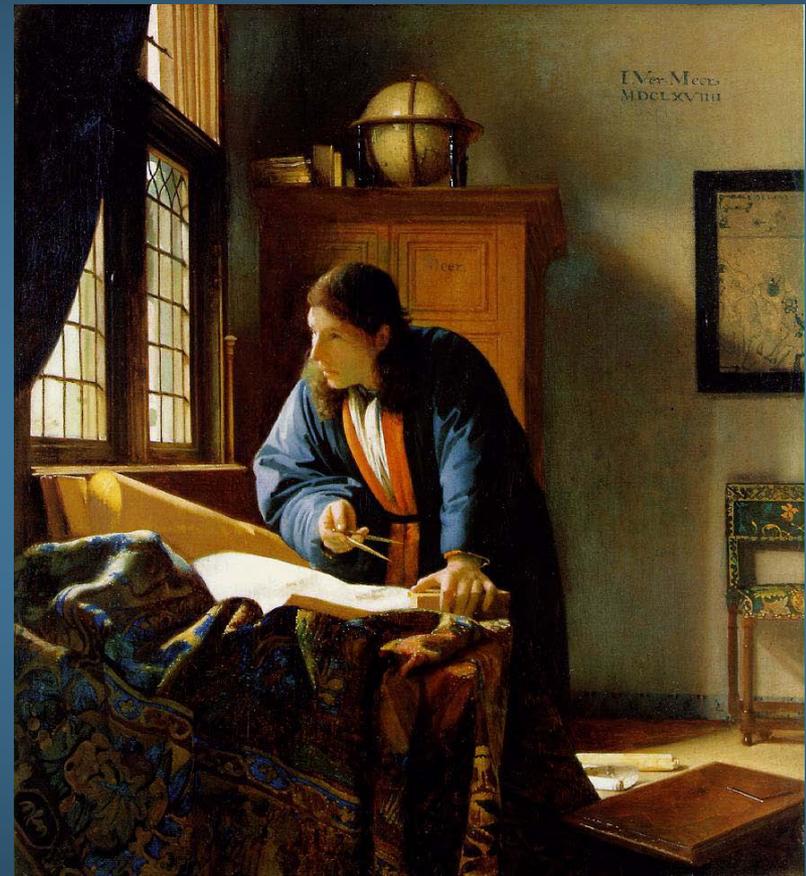
**NATIONAL CENTER
FOR GEOGRAPHIC
INFORMATION AND
ANALYSIS**

Deadline for Receipt of Proposals:
January 29, 1988



Evaluating Spatial Data Quality?

- As technology became more widespread, data bases were being shared beyond their original creators.
- Users, rather than spending resources to recreate data, might use an external data base.
- But the data need to be good enough to meet the user's "Fitness for Use" criteria.
- Data quality information typically provided as part of "Lineage" metadata
- How can you characterize and quantify measures of "Fitness for Use"?



The Geographer – Johannes Vermeer, 1668

Elements of Spatial Data Quality

- **Completeness** – presence and absence of features, their attributes and relationships;
 - commission – excess data present in a dataset,
 - omission – data absent from a dataset.
- **Logical Consistency** – degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical);
 - conceptual consistency – adherence to rules of the conceptual schema,
 - domain consistency – adherence of values to the value domains,
 - format consistency – degree to which data is stored in accordance with the physical structure of the dataset,
 - topological consistency – correctness of the explicitly encoded topological characteristics of a dataset.
- **Positional Accuracy** – accuracy of the position of features;
 - absolute or external accuracy – closeness of reported coordinate values to values accepted as or being true,
 - relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true,
 - gridded data position accuracy – closeness of gridded data position values to values accepted as or being true.
- **Temporal Accuracy** – accuracy of the temporal attributes and temporal relationships of features;
 - accuracy of a time measurement – correctness of the temporal references of an item (reporting of error in time measurement),
 - temporal consistency – correctness of ordered events or sequences, if reported,
 - temporal validity – validity of data with respect to time.
- **Thematic Accuracy** – accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships;
 - classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset),
 - non-quantitative attribute correctness – correctness of non-quantitative attributes (e.g. correctness of attribute values such as “road name” or “pavement type”),
 - quantitative attribute accuracy – accuracy of quantitative attributes.

SDQ Evaluation – Not Commonplace

- Some methods not well established
- Can be resource intensive
- Practiced (in some form) by National Mapping Agencies (UK OS, IGN), commercial firms (NAVTEQ, Tele-Atlas, GeoEye, Digital Globe)
- So what are others doing?

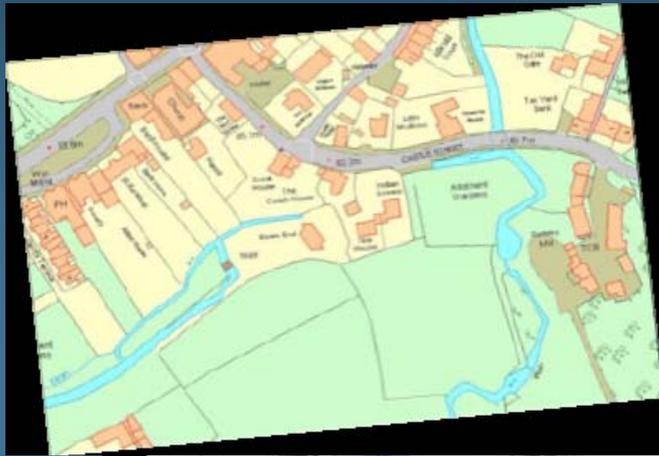
Observations of Ed Parsons, Google

When Good Enough, is Good Enough

6th International Symposium on Spatial Data Quality, St Johns, CA,
July 5-8, 2009

SDQ Evaluations

- UK Ordnance Survey case study (2008?)



- Feature based lifecycle management
- Well defined object ontology
- Quality explicitly stated..

Quality Measures..

Completeness	??
Logical Consistency	??
Positional Accuracy	0.4 - 4.0m RMSE
Temporal Accuracy	??
Thematic Accuracy	??



Data Quality = Match to Capture Specification

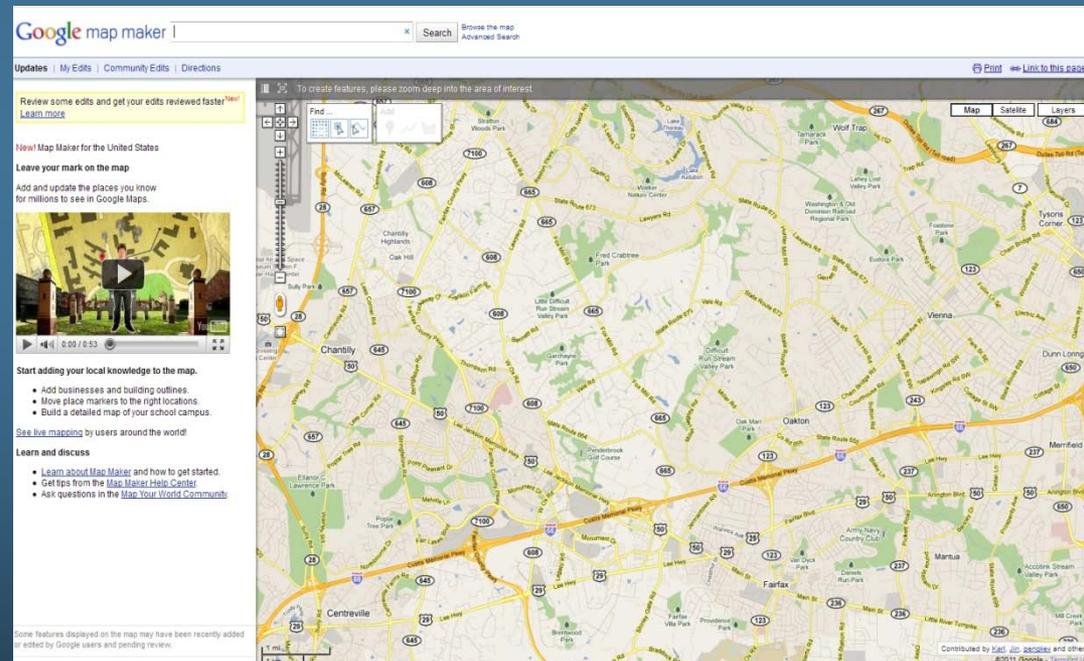
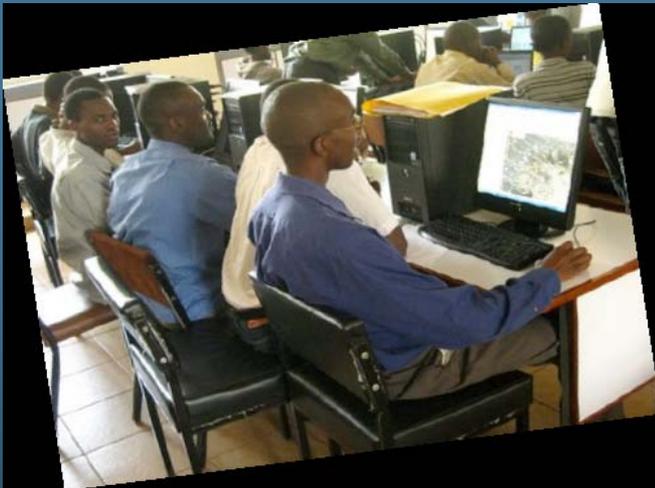
Google Map Philosophy

Data Quality = Fitness for Purpose – i.e. good enough to use

Data Quality = Uncertainty ?

Geoweb is self healing

Citizen Cartographers



From Parsons (2009)



By Rob
43 secs ago
Pending

Westpark Way Zeeland, Michigan, United States.
Road (Added)



javascript:void(0); Users



See Mike Dobson's analysis of Google Map Maker at: <http://blog.telemapics.com>



Spatial Data Quality – MAF/TIGER Context

- **Completeness** – account for all the feature instances
- **Logical Consistency** – road and boundary networks are self-consistent; the relationships between MAF Units, roads, MAF Structure Points (MSPs), and boundaries are consistent among entities
- **Positional Accuracy** – spatial feature instances are accurately located
- **Temporal Accuracy** – feature instances are up-to-date and valid
- **Thematic Accuracy** – feature classes, addresses, and attributes are correct



The Experiment

- Evaluate quality of TIGER roads, MAF units, and MSPs.
- Compare Census content to that of a [higher accuracy] local source, independently compiled, deemed to be complete, spatially accurate, and current.
- Given time/resource constraints, concentrated on **completeness** but also looked at some **logical consistency, positional accuracy, and thematic accuracy** elements as well.
- Test site – Loudoun County Virginia

Loudoun County

http://logis.loudoun.gov/weblogis/

Loudoun County, Virginia
WebLogis - Online Mapping System

Map Search Tools Results

1:400000 0 333333ft

Frederick County, MD

Jefferson County, WV

Clarke County, VA

Fauquier County, VA

Prince William County, VA

Fairfax County, VA

Prince George's County, MD

Calvert County, MD

Contents

- LandRecords
 - Tile Boundaries
 - ZIP Codes
 - Address Points
 - Road Centerline
 - Parcel Boundaries: Labels
 - Parcel Boundaries
 - Subdivisions: Labels
 - Subdivisions
- Districts
- EnvironmentalHealth
- Environmental
- Planning
- PublicSafety
- Schools
- Utilities
- Zoning
- BaseLayers

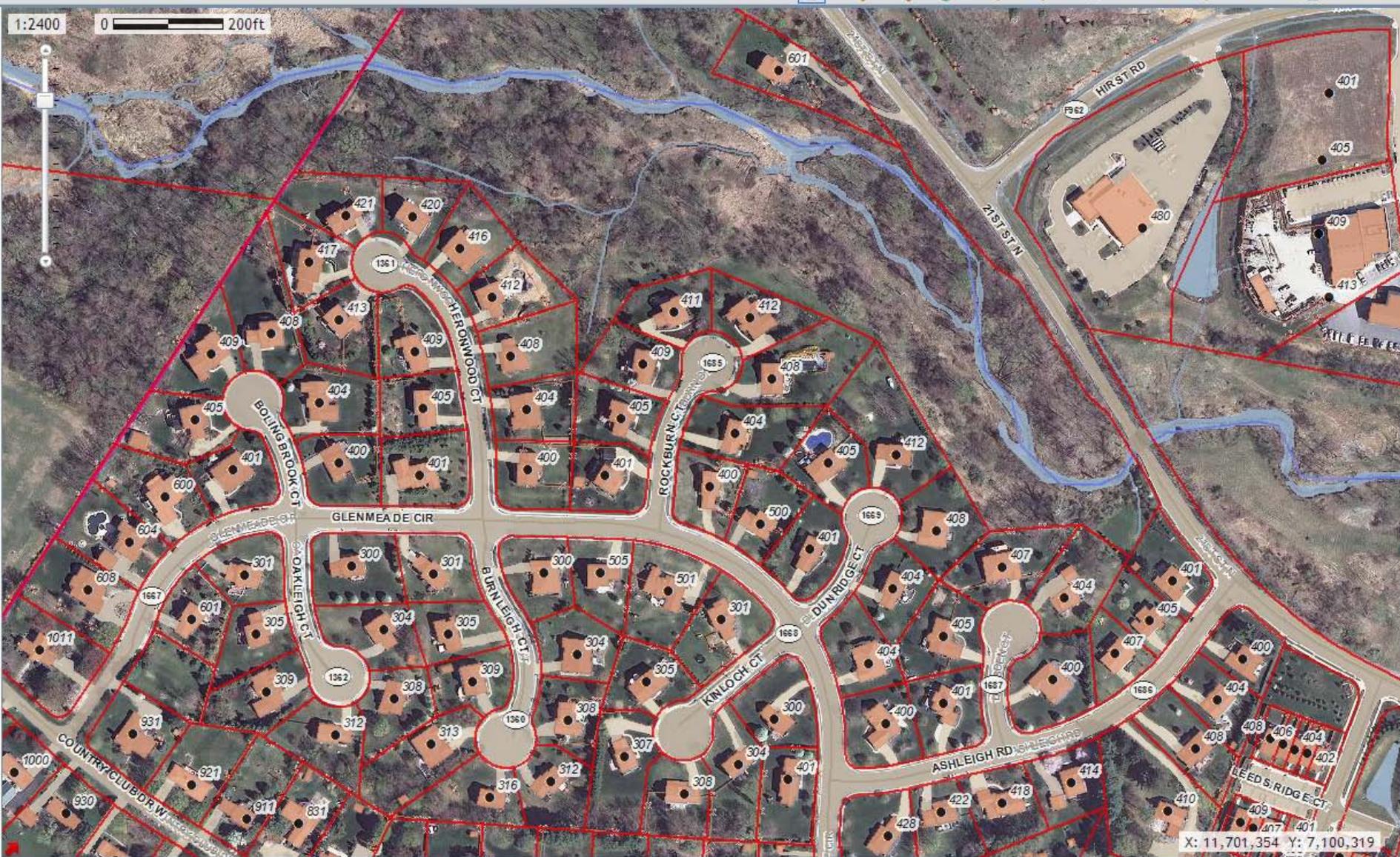
Change Transparency:

see metadata

Legend

- 2010 Population – 312,311
- Area – 521 square miles
- 4th fastest growing county (2000-10) in US
- 2nd richest county (\$112,021 median annual household income) in US

Loudoun County Data Sets from Office of Mapping and Geographic Information



TIGER Roads and Loudoun Co. Roads

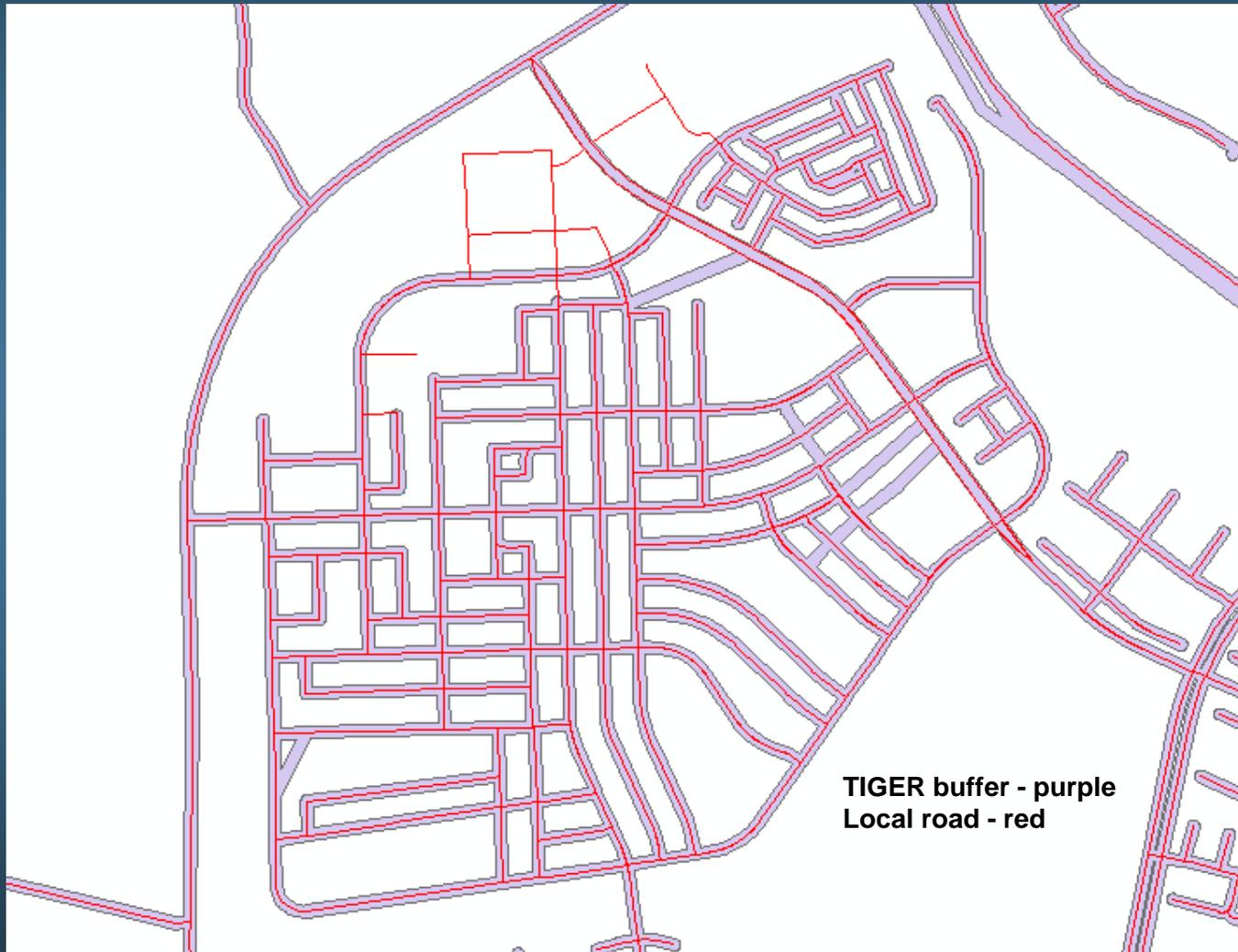
Summary Statistics :

	<i>TIGER</i>	<i>Percentage</i>	<i>Local</i>	<i>Percentage</i>
Total	22,470		19,926	
Named Roads	19,452	86.57%	19,479	97.76%
Unnamed Roads	3,018	13.43%	447	2.24%
Unnamed S1200, S1400	1,481	6.59%	-	0.00%
Private driveways, access roads* , trails, alley	1,903	8.47%	447	2.24%
Road edges outside county boundary	95	0.42%	201	1.01%

NB: Comparison was based on the number of edges.
 All the unnamed local roads are access ramps
 85 of the tiger roads edges outside the county boundary are named.
 S1100 – Primary roads, S1200 – Secondary roads, S1400 – Local roads

TIGER Roads and Loudoun Co. Roads

Local roads intersected with the TIGER (buffer width 7.6 meters)



TIGER Roads and Loudoun Co. Roads

Local roads intersection with TIGER buffer		
	Total	Percentage
Local within 7.6 m TIGER buffer	16,530	82.96%
Named	16,290	81.75%
Unnamed	240	1.20%
Partially within the 7.6m of TIGER buffer	1,529	7.67%
Named	1,442	7.24%
Unnamed	87	0.44%
Outside the TIGER 7.6m buffer	1,867	9.37%
Named	1,747	8.77%
Unnamed	120	0.60%

TIGER Roads and Loudoun Co. Roads

TIGER roads intersected with Local roads (buffer width 7.6 meters)



TIGER Roads and Loudoun Co. Roads

TIGER intersection with Local roads		
	Total	Percentage
TIGER within 7.6 m of local road buffer	18,183	80.92%
Named	17,858	79.47%
Unnamed	325	1.45%
Unnamed S1200, S1400	76	0.34%
Partially within the 7.6m of local buffer	1,167	5.19%
Named	1,030	4.58%
Unnamed	137	0.61%
Unnamed S1200, S1400	67	0.30%
Outside the 7.6m local road buffer	3,120	13.89%
Named	564	2.51%
Unnamed	2,556	13.14%
Unnamed S1200, S1400	1,338	5.95%

TIGER Roads and Loudoun Co. Roads

TIGER (blue) lacks road features which are captured in local roads (red)



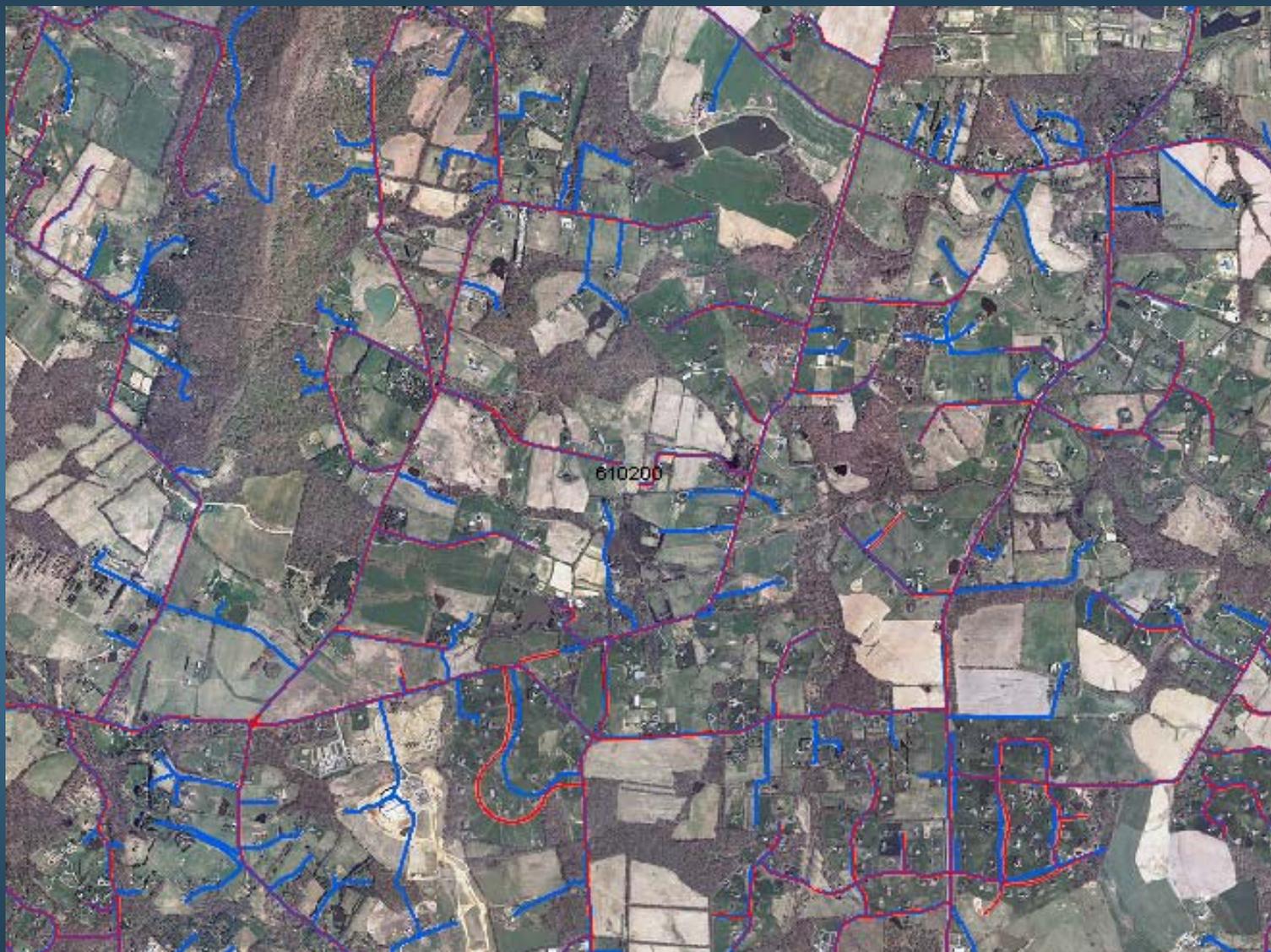
TIGER Roads and Loudoun Co. Roads

Local roads (red) match with imagery but TIGER (blue) is spatially inaccurate



TIGER Roads and Loudoun Co. Roads

Roads missing in the local roads file (shown in red) which are present in TIGER (shown in blue).
These roads appear to be mostly private driveways.



TIGER Roads and Loudoun Co. Roads

Local roads not in TIGER (red) and TIGER roads (black) are missing in imagery. They may be new since the imagery, under construction, or proposed roads.





MAF Addresses and Local Addresses

- Initial plan was to compare the congruence between the local address list (~124,000 records) and the MAF (~129,000 records).
- However two difficulties:
 - Lack of adequate time to process local address list through Geography Division address standardization and matching software
 - Local address list did not distinguish between residential and non-residential addresses
- **PLAN B**
 - Compare MAF and Local files based on base street name

MAF Addresses and Local Addresses

	MAF Location Address	MAF Mailing Address	Local Address List
# of base street names	5833	5633	5558

MAF Addresses and Local Addresses

Location Address Comparison

Street Names in MAF not Local	712 (12.2%)
Street Names in Local not MAF	436 (7.5%)

MAF Addresses and Local Addresses

Mailing Address Comparison

Street Names in MAF not Local	729 (12.9%)
Street Names in Local not MAF	329 (5.8%)

MAF Addresses and Local Addresses

Addresses Associated with Street Names

Commission/Omission Sets	Minimum number of addresses (MAF units) in set
Location Address streets not in Local	4502
Mailing Address streets not in Local	2563
Mailing and Location Address streets not in Local	2056
Local streets not in MAF Mailing Address	5338

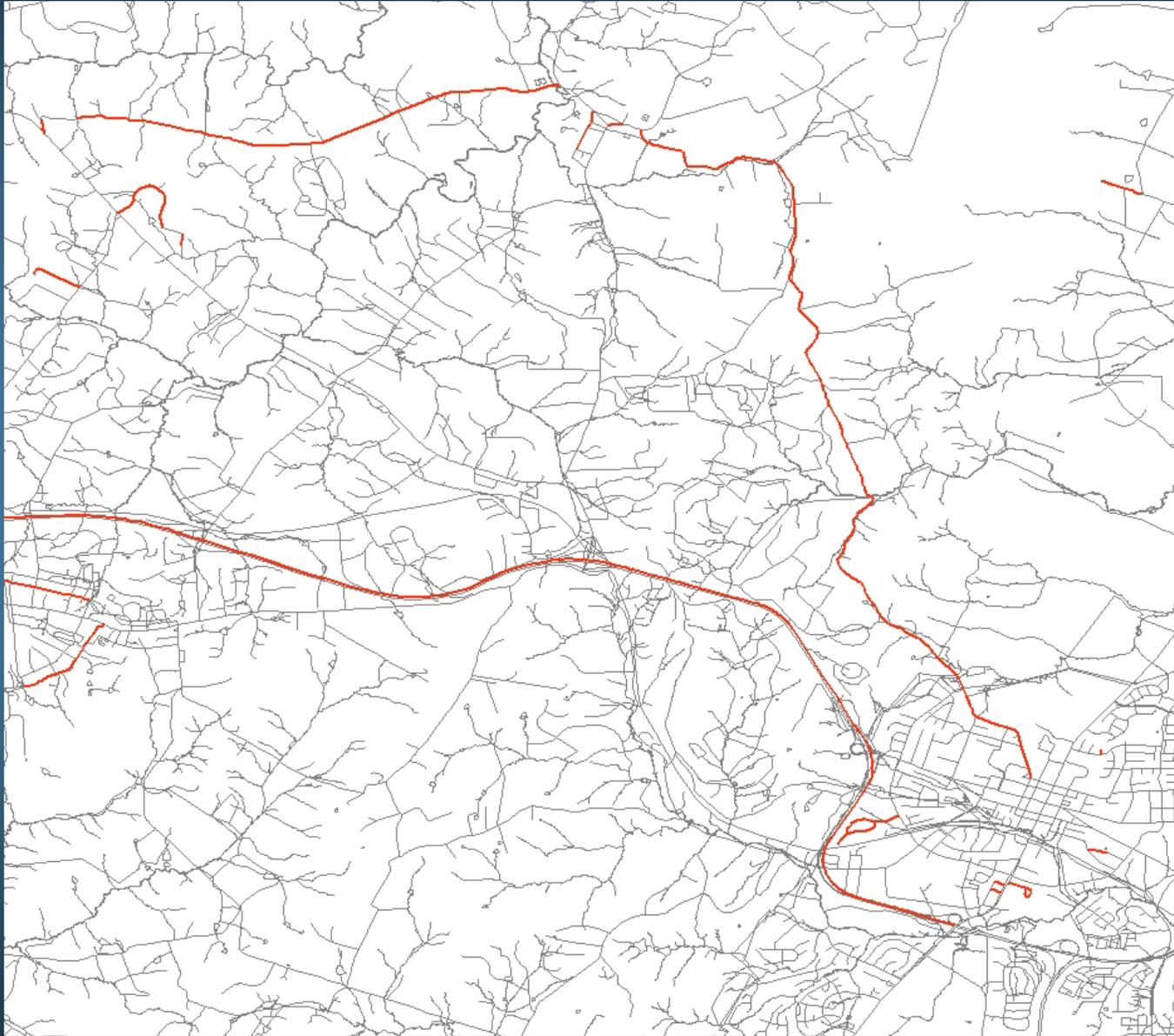


MAF Addresses and Local Addresses

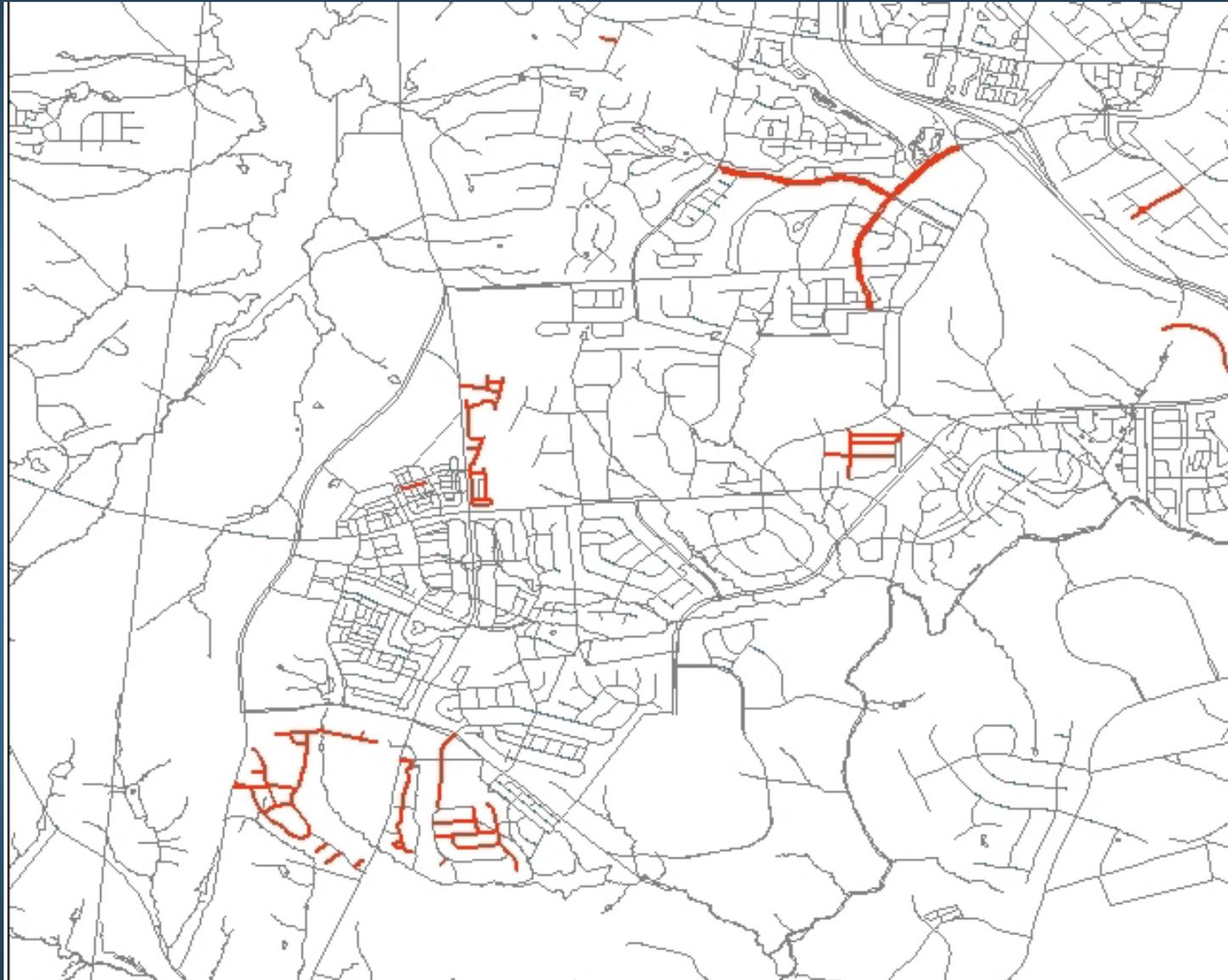
Cross-Check with Spatial Data

- For the 729 names in MAF, but NOT in local - - we matched those names to TIGER streets and the number of unique names matched is 81 (548 names did not match exactly).
- For the 329 names in local but NOT in MAF - - we matched those names to TIGER streets and the number of unique names matched is 111 (218 names did not match exactly).

Example of Roads in MAF – In TIGER – But Not Local Address File



Example of Roads in Local Address File – In TIGER But Not in MAF



Roads Flagged because of Naming Variants



West Virginia Ave

Virginia Ave West

Saint Paul St.

St. Paul St.



MSPs and Parcels

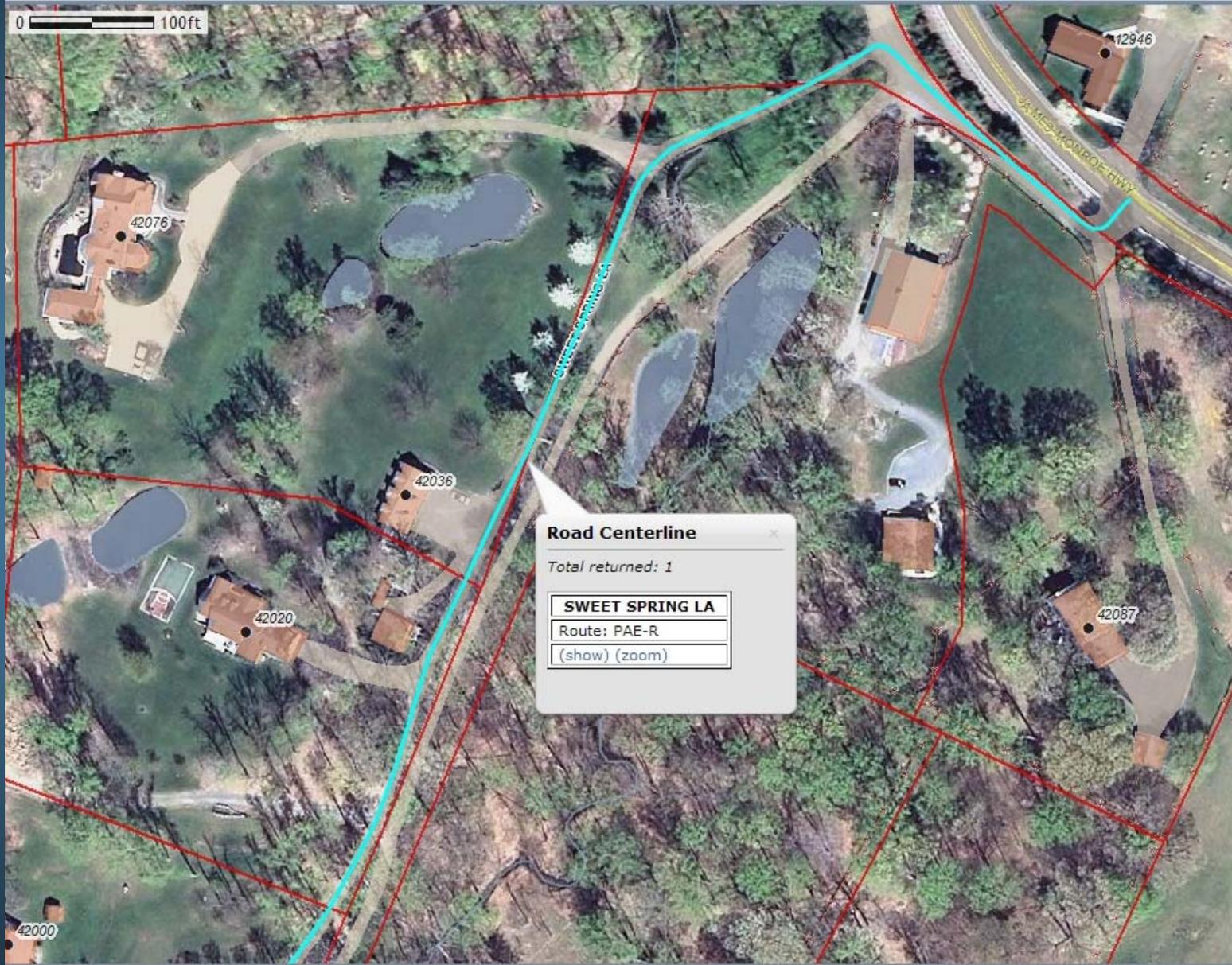
Summary Statistics

- Loudoun Data
 - 105,482 Address Points
 - 105,703 Parcels
 - 95,295 Parcels with Address Points
- MSPs
 - 92,646 MSPs
 - 904 MSPs have no house number
- Find MSPs that fall within a parcel
 - Result = 67,783 MSPs (73.2%)
- Find MSPs that also match parcel house number
 - Result = 48,954 MSPs (52.8%)

Evaluate This...



It's Complicated...



Conclusions

- This exploration has provided a preliminary estimate on aspects of the quality of the TIGER, MAF, and MSP datasets.
- More precise measures would require further processing of MAF/TIGER data, augmentation of the reference datasets, followed by more detailed testing.
- Quantification of the quality of MAF/TIGER requires evaluation of spatial and attribute data and an appropriate set of analytical tools.
- Lack of standardization/harmonization of address (street name) data makes evaluation process difficult.
- MAF/TIGER characteristics make it complicated to compile a reference dataset suitable for data quality comparisons.

Conclusions

- May need to revise/expand spatial data quality measures
 - e.g. completeness measure for spatial elements based on match/no match binomial data; possibly replace with ordinal scale - match/partial match/no match;
 - or perhaps base completeness of road segments on existence of topological edge; use positional accuracy to describe spatial correspondence of segments
- All elements that are measured on binomial scale (completeness, topological consistency, etc.) are given equal weight. Should missing a segment of an Interstate highway be treated the same as missing a driveway? Should errors be ranked on an ordinal scale...fatal/non-fatal/pass?



Special Thanks Go To:

- Frank Ussher
- Richard Watson
- Paul Namie
- Dante Terango
- Tom Fleming
- Helen Zassypkina



Questions, Comments?

sguptill@guptillgeoscience.com