ALL OVER THE MAP: COMPARING THE ACCURACY OF GEOCODING SOURCES

2019 IEPEC CONFERENCE PRESENTATION

Robert Saul, EMI Consulting
Brian Billing, AEP Ohio
August 21, 2019
AGENDA

1. “Why Does it Matter?”
2. Introduction
3. Methods
4. Results
5. Conclusion
“WHY DOES IT MATTER?”
BENEFITS OF SPATIAL ANALYSIS IN EE PROGRAMS
“WHY DOES IT MATTER?”
BENEFITS OF SPATIAL ANALYSIS IN EE PROGRAMS

MICRO-TARGETING

As energy efficiency programs become more successful, utilities are more frequently using spatial analysis to identify and influence hard-to-reach customers, such as:

• Low-income
• Environmentally disadvantaged
As energy efficiency programs become more successful, utilities are more frequently using spatial analysis to identify and influence hard-to-reach customers, such as:

- Low-income
- Environmentally disadvantaged

Many utilities are considering new program designs that hinge on specific geographic energy usage patterns, such as:

- Demand-response
- Electric vehicles
- Non-wires solutions
As energy efficiency programs become more successful, utilities are more frequently using spatial analysis to identify and influence hard-to-reach customers, such as:

- Low-income
- Environmentally disadvantaged

Many utilities are considering new program designs that hinge on specific geographic energy usage patterns, such as:

- Demand-response
- Electric vehicles
- Non-wires solutions

To better inform program design, utilities are combining program data with geographic data such as:

- Weather
- Traffic
- Demographics
“WHY DOES IT MATTER?”
LARGE SHIFTS IN GEOCODING SERVICES

1. OLD INFO IN STAND-ALONE SERVICES
2. CHANGE IN API REQUIREMENTS
3. CHANGES TO SERVICE CAPABILITIES
“WHY DOES IT MATTER?”
LARGE SHIFTS IN GEOCODING SERVICES

1. OLD INFO IN STAND-ALONE SERVICES

2. CHANGE IN API REQUIREMENTS

3. CHANGES TO SERVICE CAPABILITIES

“WHY DOES IT MATTER?”

LARGE SHIFTS IN GEOCODING SERVICES

1. OLD INFO IN STAND-ALONE SERVICES

2. CHANGE IN API REQUIREMENTS

3. CHANGES TO SERVICE CAPABILITIES

"Google has recently changed its API requirements...users are now required to provide an API key and enable billing."

https://www.littlemissdata.com/blog/ggmap-updated
“WHY DOES IT MATTER?”

LARGE SHIFTS IN GEOCODING SERVICES

1. OLD INFO IN STAND-ALONE SERVICES

2. CHANGE IN API REQUIREMENTS

3. CHANGES TO SERVICE CAPABILITIES

https://multiplottr.com/
“WHY DOES IT MATTER?”
WHAT ACCURACY IS NEEDED AND AT WHAT COST?
“WHY DOES IT MATTER?”
WHAT ACCURACY IS NEEDED AND AT WHAT COST?

https://community.tableau.com/thread/232540
AGENDA

1. “Why Does it Matter?”
2. Introduction
3. Methods
4. Results
5. Conclusion
## Review of Geocoding Services and Sites

<table>
<thead>
<tr>
<th>Geocoding Service</th>
<th>Free or Paid Service</th>
<th>Has API?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteryx</td>
<td>Free</td>
<td></td>
</tr>
<tr>
<td>ArcGIS</td>
<td>Paid</td>
<td>✓</td>
</tr>
<tr>
<td>BatchGeo</td>
<td>Free and Paid</td>
<td></td>
</tr>
<tr>
<td>GeoLytics</td>
<td>Paid</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>Free and Paid</td>
<td>✓</td>
</tr>
<tr>
<td>Google Sheets</td>
<td>Free and Paid</td>
<td></td>
</tr>
<tr>
<td>HERE</td>
<td>Free and Paid</td>
<td>✓</td>
</tr>
<tr>
<td>MapLarge</td>
<td>Paid</td>
<td>✓</td>
</tr>
<tr>
<td>OpenStreetMap</td>
<td>Free</td>
<td>✓</td>
</tr>
<tr>
<td>Texas A&amp;M Geoservices Geocoding</td>
<td>Free and Paid</td>
<td>✓</td>
</tr>
<tr>
<td>TomTom</td>
<td>Free and Paid</td>
<td>✓</td>
</tr>
<tr>
<td>United States Census Bureau Geocoder API</td>
<td>Free</td>
<td>✓</td>
</tr>
</tbody>
</table>
AEP Ohio: Electric investor-owned utility serving nearly 1.5 million customers.
INTRODUCTION

CLIENT AND PROGRAM

CLIENT

AEP Ohio: Electric investor-owned utility serving nearly 1.5 million customers.

GEOGRAPHY

Majority of AEP Ohio customers located in and around Columbus, OH.
INTRODUCTION

CLIENT AND PROGRAM

CLIENT

AEP Ohio: Electric investor-owned utility serving nearly 1.5 million customers.

GEOGRAPHY

Majority of AEP Ohio customers located in and around Columbus, OH.

DATA

Examined the geolocational data of 151 AEP Ohio residential survey respondents who completed a Qualtrics web survey in January 2019.

Survey respondents part of residential rebate program evaluation.
AGENDA

1. “Why Does it Matter?”
2. Introduction
3. Methods
4. Results
5. Conclusion
METHODS

DATASETS

RAW DATASET

Raw dataset of survey respondent service addresses as received from AEP Ohio.
METHODS
DATASETS

RAW DATASET

Raw dataset of survey respondent service addresses as received from AEP Ohio.

CLEAN DATASET

Cleaned dataset of survey respondent service addresses using cleaning best-practices recommended by the Harvard School of Public Health (HSPH 2017).
RAW DATASET

Raw dataset of survey respondent service addresses as received from AEP Ohio.

CLEAN DATASET

Cleaned dataset of survey respondent service addresses using cleaning best-practices recommended by the Harvard School of Public Health (HSPH 2017).

HYPOTHESIS

Differences in the geocoded locations of raw data and the cleaned data would give an indication of how effectively each service handles data quality issues.
METHODS

GEOCODING SOURCES TESTED

1. USED GOOGLE API AS “BASELINE”

2. PERFORMED SPOT-CHECK ON 20 RECORDS

3. COMPARED BASELINE TO 6 OTHER SERVICES
METHODS
GEOCODING SOURCES TESTED

1. USED GOOGLE API AS “BASELINE”

2. PERFORMED SPOT-CHECK ON 20 RECORDS

3. COMPARED BASELINE TO 6 OTHER SERVICES
METHODS

GEOCODING SOURCES TESTED

1. USED GOOGLE API AS “BASELINE”

2. PERFORMED SPOT-CHECK ON 20 RECORDS

3. COMPARED BASELINE TO 6 OTHER SERVICES
METHODS
GEOCODING SOURCES TESTED

1. USED GOOGLE API AS “BASELINE”

2. PERFORMED SPOT-CHECK ON 20 RECORDS

3. COMPARED BASELINE TO 6 OTHER SERVICES

<table>
<thead>
<tr>
<th>Geocoding Service</th>
<th>Geocoding Type</th>
<th>Geocoded as Baseline</th>
<th>Geocodes Performed on Both Raw and Clean Address Data</th>
<th>Geocodes Based on IP Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 Garmin</td>
<td>Map layer used as address locator tool in ArcMap</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ArcGIS Online</td>
<td>Online application</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Census API</td>
<td>API</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google API</td>
<td>API</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Sheets</td>
<td>Script within online application</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSM API</td>
<td>API</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualtrics</td>
<td>Locational data provided with survey responses</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AGENDA

1. “Why Does it Matter?”
2. Introduction
3. Methods
4. Results
5. Conclusion
RESULTS
SUMMARY STATISTICS FROM COMPARISON TO BASELINE

RAW VS CLEAN

In this particular experiment, found little difference in how the services geocoded the raw addresses and the cleaned addresses.

Likely speaks to good data practices on the part of AEP Ohio rather than the capabilities of the geocoding services.
**RESULTS**

**SUMMARY STATISTICS FROM COMPARISON TO BASELINE**

In this particular experiment, found little difference in how the services geocoded the raw addresses and the cleaned addresses. Likely speaks to good data practices on the part of AEP Ohio rather than the capabilities of the geocoding services.

- Qualtrics was the outlier.

<table>
<thead>
<tr>
<th>GEOCODING SERVICE</th>
<th>UNMATCHED GEOCODING RECORDS</th>
<th>AVERAGE DISTANCE FROM BASELINE COORDINATE (IN MILES)</th>
<th>MEDIAN DISTANCE FROM BASELINE COORDINATE (IN MILES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Sheets</td>
<td>0</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Census API</td>
<td>10</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>2013 Garmin</td>
<td>0</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>OSM API</td>
<td>61</td>
<td>0.32</td>
<td>0.11</td>
</tr>
<tr>
<td>Qualtrics</td>
<td>0</td>
<td>95.30</td>
<td>6.15</td>
</tr>
</tbody>
</table>
RESULTS

SUMMARY STATISTICS FROM COMPARISON TO BASELINE

RAW VS CLEAN

In this particular experiment, found little difference in how the services geocoded the raw addresses and the cleaned addresses. Likely speaks to good data practices on the part of AEP Ohio rather than the capabilities of the geocoding services.

BASELINE VS OTHER SERVICES

- Qualtrics was the outlier.
- OSM and the Census did not find matches for all addresses.

<table>
<thead>
<tr>
<th>Geocoding Service</th>
<th>Unmatched Geocoding Records</th>
<th>Average Distance from Baseline Coordinate (in Miles)</th>
<th>Median Distance from Baseline Coordinate (in Miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Sheets</td>
<td>0</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Census API</td>
<td>10</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>2013 Garmin</td>
<td>0</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>OSM API</td>
<td>61</td>
<td>0.32</td>
<td>0.11</td>
</tr>
<tr>
<td>Qualtrics</td>
<td>0</td>
<td>95.30</td>
<td>6.15</td>
</tr>
</tbody>
</table>
RESULTS
SUMMARY STATISTICS FROM COMPARISON TO BASELINE

In this particular experiment, found little difference in how the services geocoded the raw addresses and the cleaned addresses. Likely speaks to good data practices on the part of AEP Ohio rather than the capabilities of the geocoding services.

- Qualtrics was the outlier.
- OSM and the Census did not find matches for all addresses.
- Excluding Sheets, Census API had results closest to baseline.

<table>
<thead>
<tr>
<th>Geocoding Service</th>
<th>Unmatched Geocoding Records</th>
<th>Average Distance from Baseline Coordinate (in Miles)</th>
<th>Median Distance from Baseline Coordinate (in Miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Sheets</td>
<td>0</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Census API</td>
<td>10</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>2013 Garmin</td>
<td>0</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>OSM API</td>
<td>61</td>
<td>0.32</td>
<td>0.11</td>
</tr>
<tr>
<td>Qualtrics</td>
<td>0</td>
<td>95.30</td>
<td>6.15</td>
</tr>
</tbody>
</table>
RESULTS
FURTHER EXPLORING DIFFERENCES

1. QUALTRICS: SURVEY RESPONDENT LOCATION

2. GEOCODING SERVICES VISUAL COMPARISON

3. COMPARING BASELINE TO ARCGIS ONLINE
RESULTS
FURTHER EXPLORING DIFFERENCES

1. QUALTRICS: SURVEY RESPONDENT LOCATION

2. GEOCODING SERVICES VISUAL COMPARISON

3. COMPARING BASELINE TO ARCGIS ONLINE
RESULTS
FURTHER EXPLORING DIFFERENCES

1. QUALTRICS: SURVEY RESPONDENT LOCATION

2. GEOCODING SERVICES VISUAL COMPARISON

3. COMPARING BASELINE TO ARCGIS ONLINE
RESULTS
FURTHER EXPLORING DIFFERENCES

1. QUALTRICS: SURVEY RESPONDENT LOCATION

2. GEOCODING SERVICES VISUAL COMPARISON

3. COMPARING BASELINE TO ARCGIS ONLINE
AGENDA

1. “Why Does it Matter?”
2. Introduction
3. Methods
4. Results
5. Conclusion
Geocode a small amount of records (less than 100): may be easiest to use to Google Sheets.
**CONCLUSION**

**TIPS AND TRICKS**

**SMALL BATCH**

Geocode a small amount of records (less than 100): may be easiest to use to Google Sheets.

**LARGE BATCH**

Geocode a larger amount of records (less than 40k): using the Google API may be free to use.*

* Statement was true at the time the paper was written.
**CONCLUSION**

**TIPS AND TRICKS**

**SMALL BATCH**

Geocode a small amount of records (less than 100): may be easiest to use to Google Sheets.

**LARGE BATCH**

Geocode a larger amount of records (less than 40k): using the Google API may be free to use.*

**INCREASE ACCURACY**

Can perform a second set of geocoding on a subset of records using a different geocoding service.

Can check the subset of geocodes that are flagged as potentially less accurate based on the geocoders accuracy field.

* Statement was true at the time the paper was written.
## Conclusion

### Summary Table

<table>
<thead>
<tr>
<th>Geocoding Service</th>
<th>Recommended Use</th>
<th>Cost for Small Number of Geocoded Locations</th>
<th>Cost for Large Number of Geocoded Locations</th>
<th>Accuracy</th>
<th>Includes Accuracy Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google API</td>
<td>Accurate geocoding of a small number of records; have help from a proficient coder</td>
<td>Free (less than 40,000 calls a month)</td>
<td>High (between $4 and $5 per 1,000 calls)</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Google Sheets</td>
<td>Accurate geocoding of a small record batch.</td>
<td>Free (unclear when pay barrier starts)</td>
<td>Unknown (unclear when pay barrier starts)</td>
<td>High</td>
<td>No</td>
</tr>
<tr>
<td>Census API</td>
<td>Fairly accurate geocoding of large number of records; unmatched records won’t pose an issue; have help from a proficient coder; pair with another service for unmatched addresses</td>
<td>Free</td>
<td>Moderate – High</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2013 Garmin North America</td>
<td>Fairly accurate geocoding of large number of records; fast processing time; own Garmin map layers; have help from an ArcGIS user</td>
<td>Cost of acquiring Garmin map layer (cost unknown)</td>
<td>Moderate – High</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>OSM API</td>
<td>Less accurate geocoding of large number of records; unmatched records won’t pose an issue; have help from a proficient coder; pair with another service for unmatched addresses</td>
<td>Free</td>
<td>Low – Moderate</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Qualtrics</td>
<td>Quick check that survey respondents are answering in roughly the predicted pattern</td>
<td>Free (no additional cost with Qualtrics service)</td>
<td>Low</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>ArcGIS Online</td>
<td>Accurate geocoding; high price</td>
<td>Approximately $4 for 1,000 geocodes</td>
<td>High</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
CONCLUSION

WHAT ACCURACY IS NEEDED AND AT WHAT COST?
CONCLUSION

WHAT ACCURACY IS NEEDED AND AT WHAT COST?

https://oehha.ca.gov/calenviroscreen/maps-data
Questions?

Robert Saul
rsaul@emiconsulting.com
t. 206.388.0973