

Enhancing Administrative Tax Record Imputations through Machine Learning: Utilizing Workgroups to Strengthen Production Processes

Kate Willyard, Mark Frame, Jadvir Gill, James HoShek, Amelia Ingram, Ming Ray Liao, Albert
Nedelman, Angelica Phillips, and Sam Shirazi

Social, Economic and Housing Statistics Division (SEHSD), U.S. Census Bureau

2024 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference

Thursday, October 24th at 10:30 AM

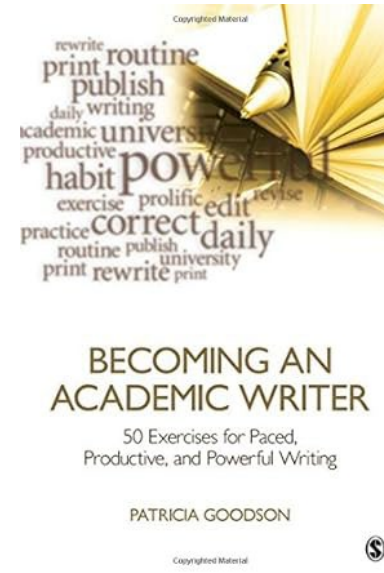
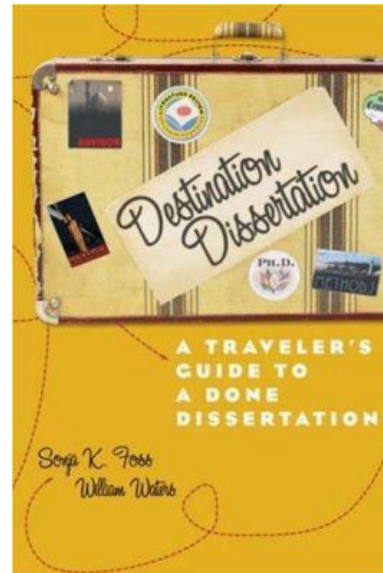
College Park Marriott Hotel and Conference Center, Room 0105, Hyattsville, MD 20783



Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion.
Any views expressed are those of the authors and not those of the U.S. Census Bureau.

The Machine Learning Workgroup

- Background
- Purpose
- Project charter
- Recruitment
- Sub team volunteers
- Assigned readings



Designing Research for Publication



Anne Sigismund Huff



PROJECT MANAGEMENT IN PRACTICE



MANTEL | MEREDITH | SHAFER | SUTTON | 4 E

Collaborative Efforts

- Regular workgroup meetings
 - Monthly team meetings to learn together
 - Sub team meetings to catch up and solve problems
- Meetings to determine research question
 - Met with senior researchers and leaders within the Social, Economic and Housing Statistics Division (SEHSD) to discuss needs
 - Met with senior researchers outside SEHSD in the Center for Statistical Research and Methodology (CSRM) to discuss viability

Finding the Best Research Question

- Evaluation of potential research questions
 - Timely
 - Achievable
 - Relevant
- Chosen research question: To what extent can machine learning methods enhance the geocoding of individual tax returns to Census blocks for the purpose of Small Area Income and Poverty Estimates (SAIPE)?

The Need to Impute Missing Geocodes

- SAIPE provides income and poverty estimates used in the distribution of federal funds to local jurisdictions
- The SAIPE school district model uses related school age children in poverty shares measured from individual tax returns
- Some individual tax returns cannot be geocoded
- Accurate geocoding reduces the potential bias of estimates

Literature Review – Traditional Methods

- Random allocation using:
 - land area
 - total housing units
 - total population
 - age/sex/race/ethnicity weighting factors
- Deterministic centroid allocation

References: Hibbert et al., 2009; Henry and Boscoe, 2008; Hurley et al., 2003; Lan Luo et al., 2010; Song Lin, 2016

Literature Review – Machine Learning Methods

- Combinations of different machine learning similarity measures or deep learning algorithms in text-based address matching
- Applications to other parts of the process, such as:
 - address parsing and address locating
 - test set selection
 - model selection
 - estimating match rate accuracy and other quality control tasks

References: Cruz et al., 2022; Lee et al., 2020

Literature Review – Challenges

- Traditional geocode imputation methods perform best when weighted by demographics, but less accurate in dense areas
- Machine learning methods so far are very computationally intensive for relatively modest gains

References: Henry and Boscoe, 2008; Hurley et al., 2003; Cruz et al., 2022

Utilization of Mailing Address Information to Geocode Individual Tax Exemptions

1040 U.S. Individual Income Tax Return 21

or Current Resident

Escondido, CA 92027-3640

Individual Tax Exemptions

Master Address File (MAF)

Match

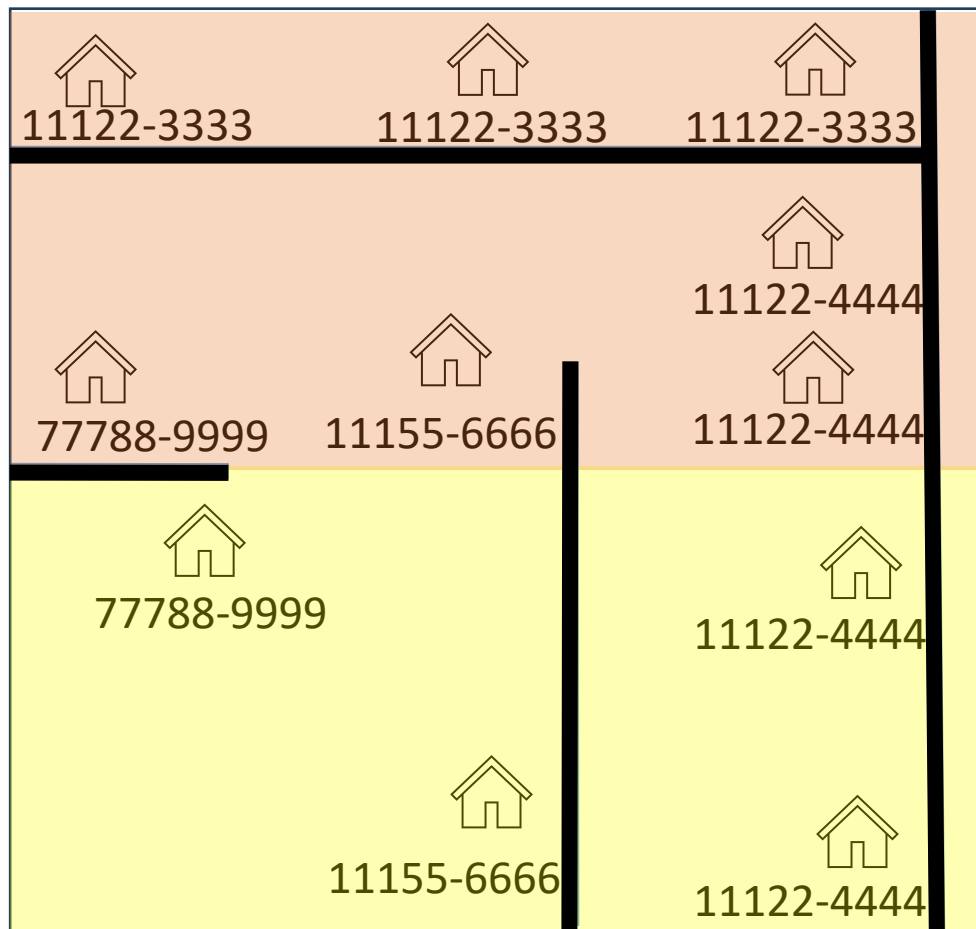
Use MAFID Geocode

No Match




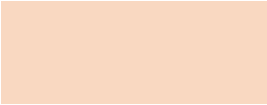
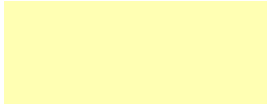
Randomly Allocate Geocode Based on Mailing Address Information

Introduction to Block-ZIP Pieces

Example City View of Roads, Mailable Addresses, ZIP Codes, and Two Census Blocks

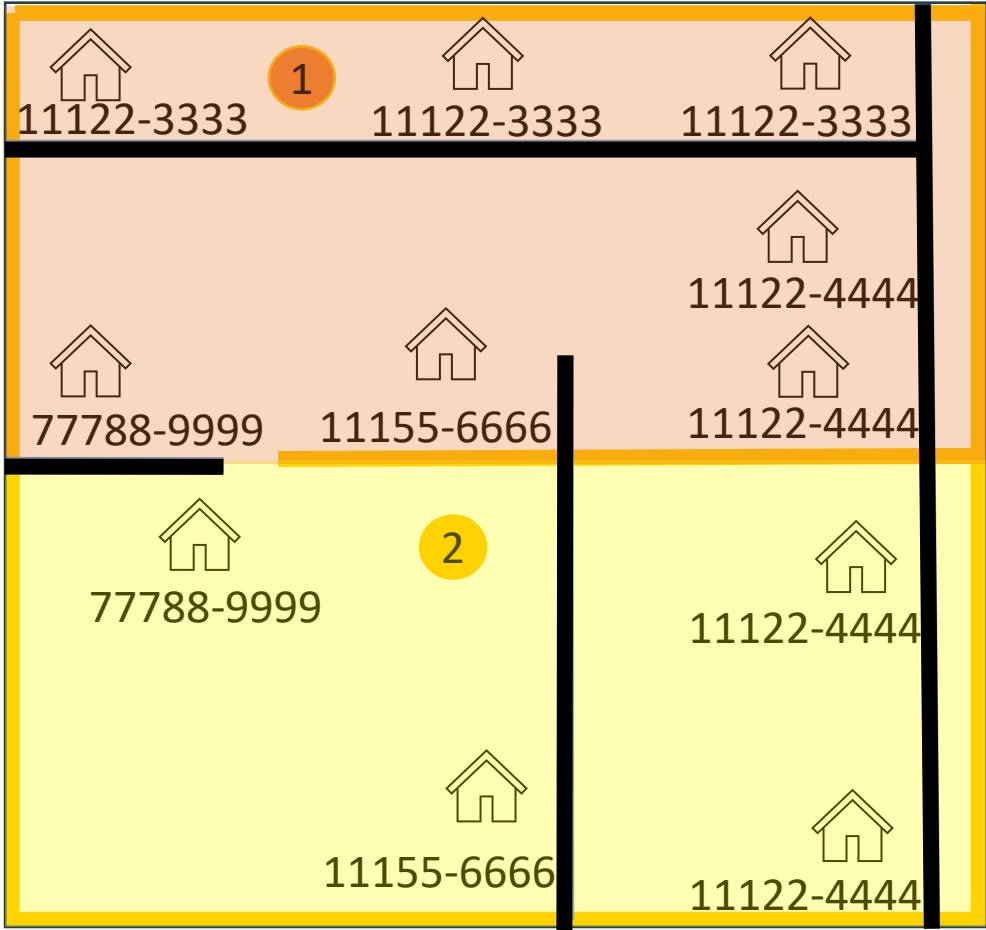


Legend





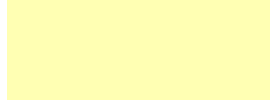

-  Road
-  Mailable Address
-  ZIP Code
-  Census Block 01-001-000001-0001
-  Census Block 01-001-000001-0002

Example of Census Blocks

Example City View of Roads, Mailable Addresses, ZIP Codes, and Two Census Blocks

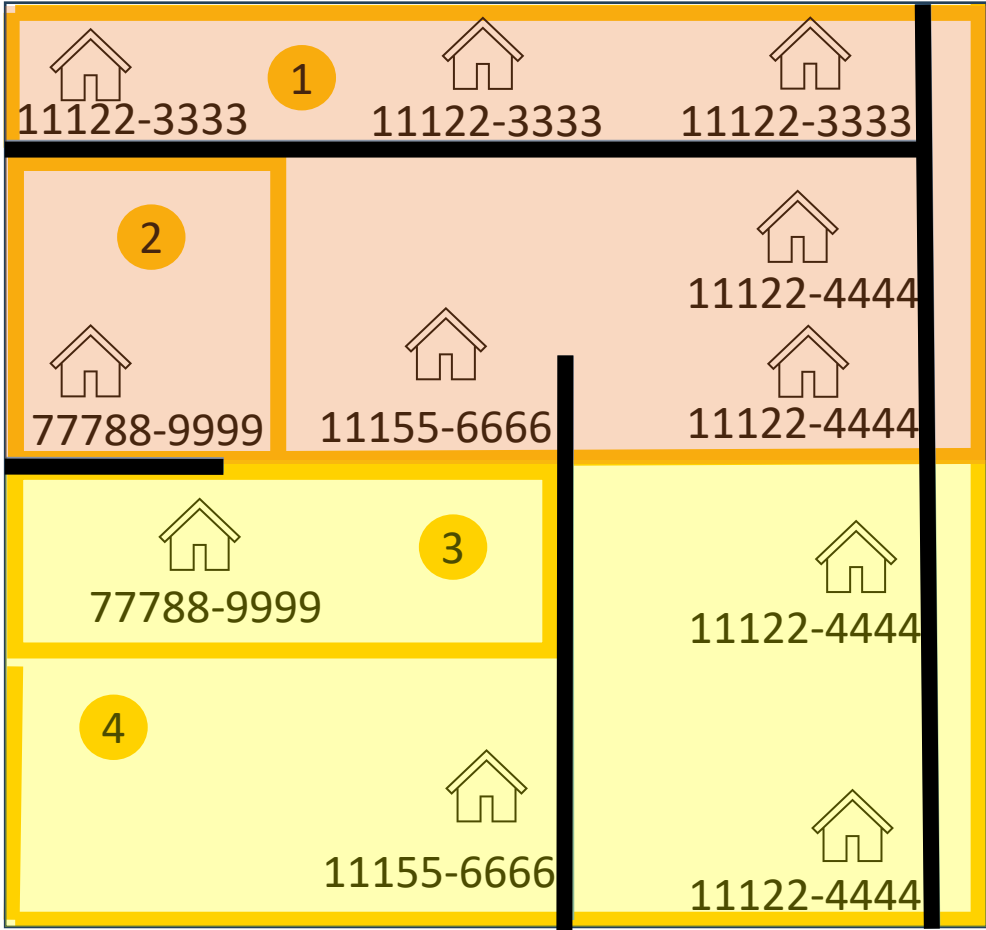


Legend





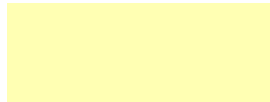

-  Road
-  Mailable Address
-  ZIP Code
-  Census Block 01-001-000001-0001
-  Census Block 01-001-000001-0002
-  Census Block

Example of Census Block-ZIP3 Pieces

Example City View of Roads, Mailable Addresses, ZIP Codes, Two Census Blocks, and Four Block-ZIP3 Pieces

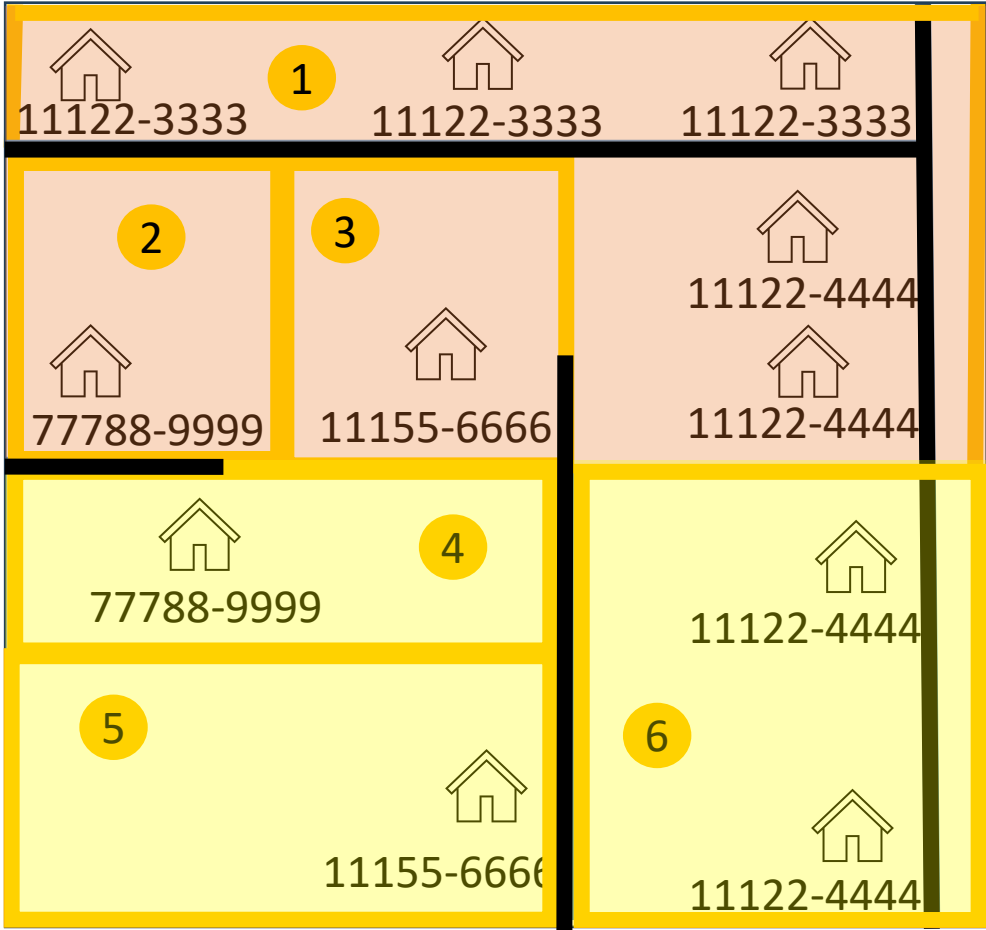


Legend





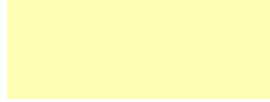

-  Road
-  Mailable Address
-  ZIP Code
-  Census Block 01-001-000001-0001
-  Census Block 01-001-000001-0002
-  Census Block ZIP3 Piece

Example of Census Block-ZIP5 Pieces

Example City View of Roads, Mailable Addresses, ZIP Codes, Two Census Blocks, and Six Block-ZIP5 Pieces

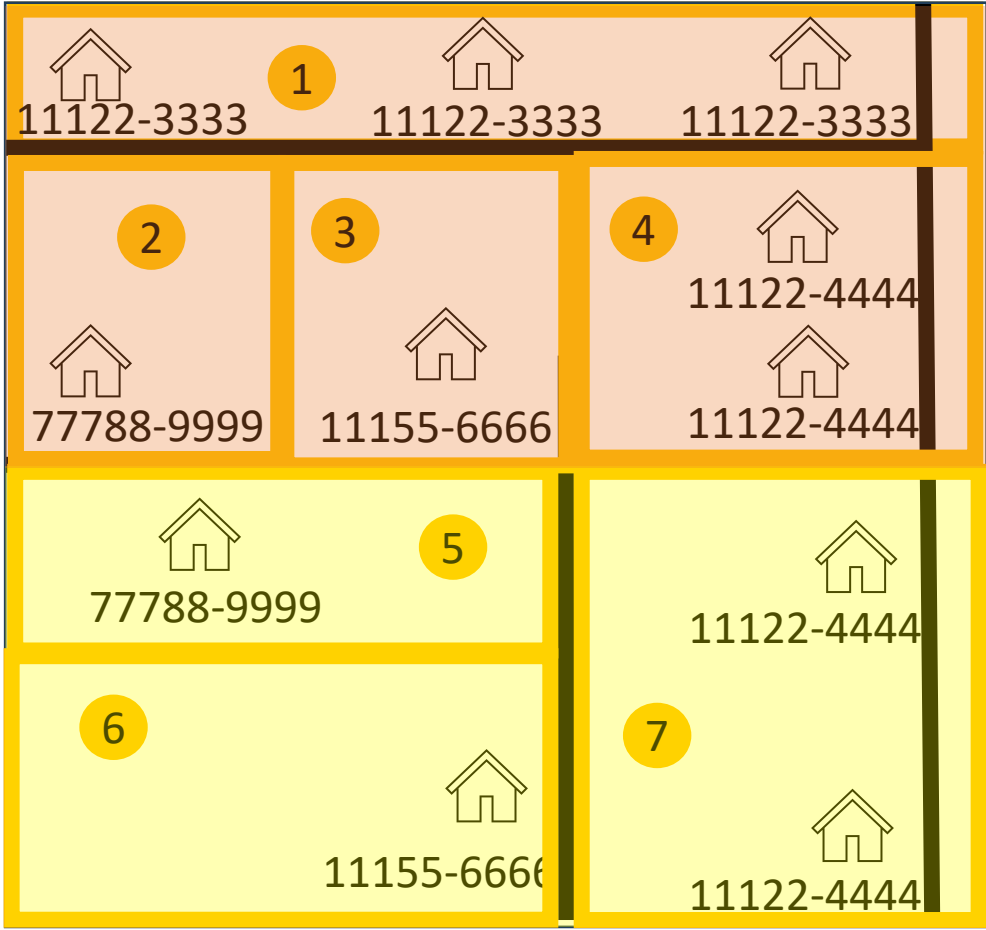


Legend





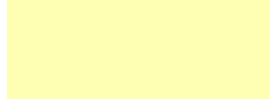

-  Road
-  Mailable Address
-  ZIP Code
-  Census Block 01-001-000001-0001
-  Census Block 01-001-000001-0002
-  Census Block ZIP5 Piece

Example of Census Block-ZIP9 Pieces

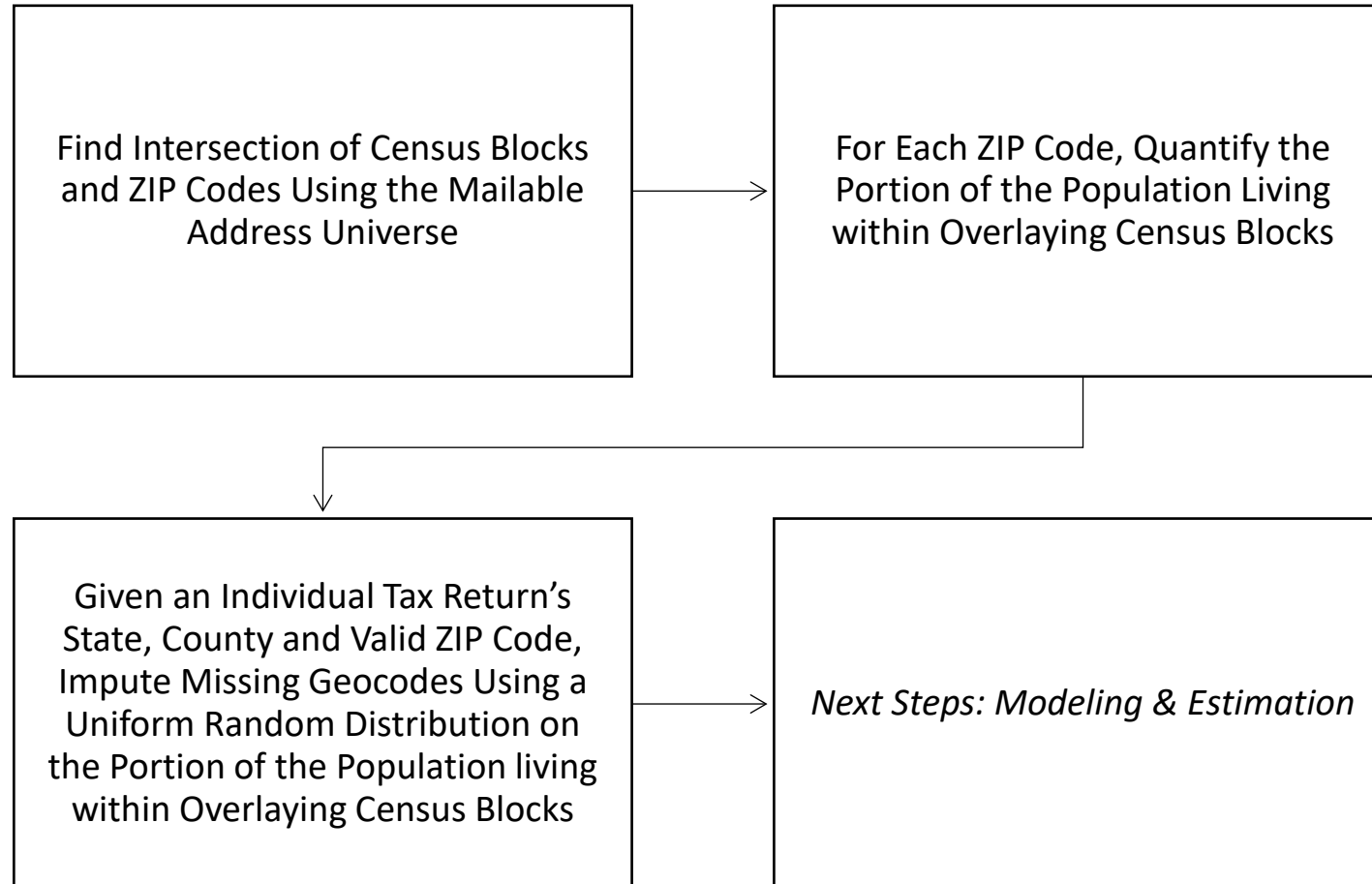
Example City View of Roads, Mailable Addresses, ZIP Codes, Two Census Blocks, and Seven Block-ZIP9 Pieces



Legend

-  Road
-  Mailable Address
-  ZIP Code
-  Census Block 01-001-000001-0001
-  Census Block 01-001-000001-0002
-  Census Block ZIP9 Piece

Current Process to Impute Missing Geocodes



Considering Another Framework to Allocate Individual Tax Exemptions without Geocodes Utilizing Mailing Address Information

Current Production Framework

- Based on Decennial Census shares

Machine Learning Framework

- Based on modeled estimates of the number of individual tax exemptions without geocodes in the SAIPE universe

Data Utilized in Machine Learning Framework

Target Variable

- Number individual tax exemptions without geocodes in the SAIPE universe
 - SAIPE universe
 - Tax exemptions with geocodes in the SAIPE universe
 - Non-filers in the SAIPE universe

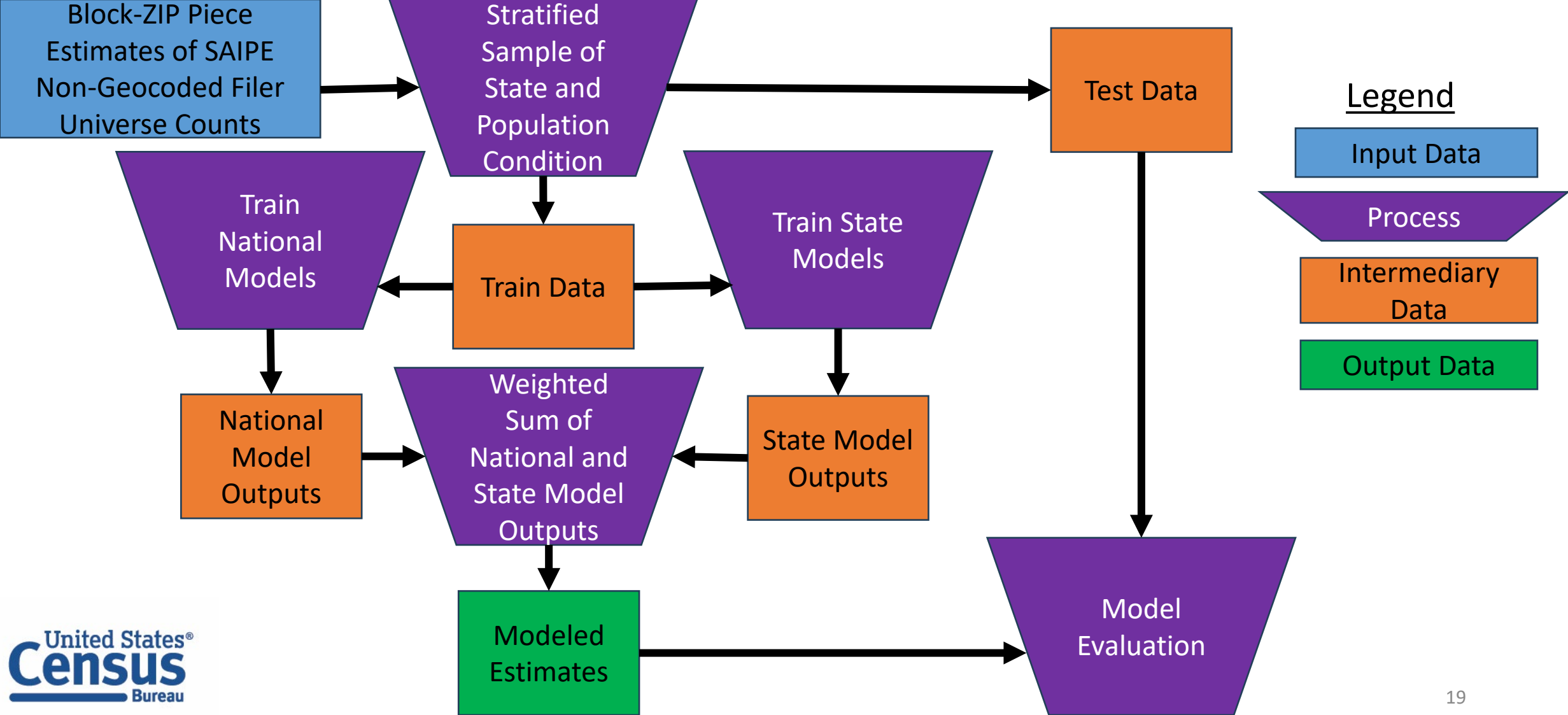
Feature Variables

- Mailable address counts
- Rural/urban housing units
- Decennial Census residential population counts by race/gender/age/group quarters type/household type

Machine Learning Method Overview

- Unit of Analysis
 - Census block
 - Census block-ZIP3 piece
 - Census block-ZIP5 piece
 - Census block-ZIP9 piece
- Data Sources
 - 2020 Decennial Census Microdata Detailed File
 - 2021 Income Year Individual 1040 Tax Records
 - 2022 Population Estimates Program County Population Estimates
- Models Considered
 - Lasso regression
 - Ridge regression
 - Logistic regression with elastic net
 - Decision tree
 - Random forest
 - k-Nearest neighbors
 - Naive bayes
 - AdaBoost trees
 - Gradient boosting trees
 - Support vector machine
 - Model averages

Machine Learning Modeling Process



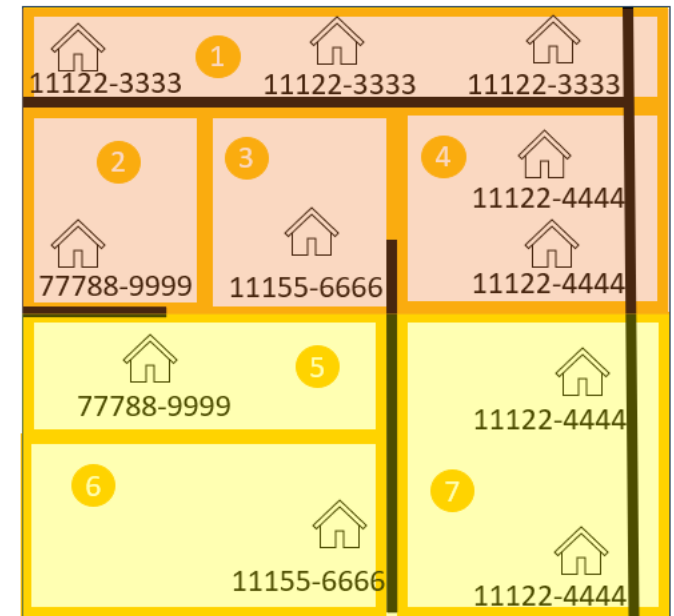
Machine Learning Model Evaluation Methods

- Matthews correlation: Correlation of Y-predict and Y-true
- Geometric mean: A geometric mean of recall of the two classes
- Area under the receiver operator curve
- Area under the precision recall curve

Utilization of Machine Learning Results (Part 1 of 5)

Hypothetical Example

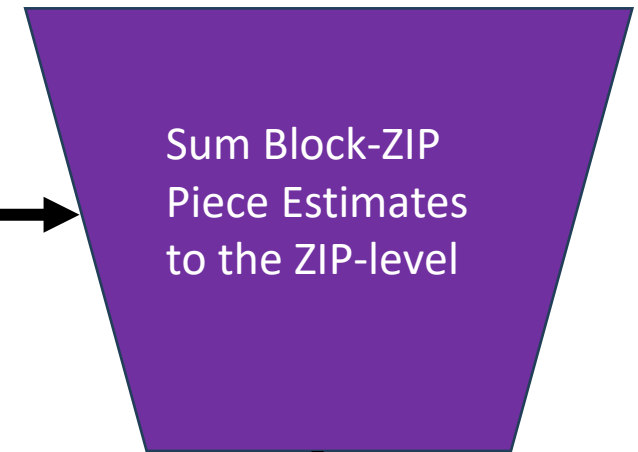
id	block	zip9	modeled estimate of SAIGE universe non-geocoded exemption counts
1	01-001-000001-0001	11122-3333	2
2	01-001-000001-0001	77788-9999	2
3	01-001-000001-0001	11155-6666	1
4	01-001-000001-0001	11122-4444	1
5	01-001-000001-0002	77788-9999	6
6	01-001-000001-0002	11155-6666	0
7	01-001-000001-0002	11122-4444	1



Utilization of Machine Learning Results (Part 2 of 5)

Hypothetical Example

id	block	zip9	modeled estimate of SAIPE universe non-geocoded exemption counts
1	01-001-000001-0001	11122-3333	2
2	01-001-000001-0001	77788-9999	2
3	01-001-000001-0001	11155-6666	1
4	01-001-000001-0001	11122-4444	1
5	01-001-000001-0002	77788-9999	6
6	01-001-000001-0002	11155-6666	0
7	01-001-000001-0002	11122-4444	1



zip9	modeled estimate
11122-3333	2
11122-4444	2
11155-6666	1
77788-9999	8

Utilization of Machine Learning Results (Part 3 of 5)

Hypothetical Example

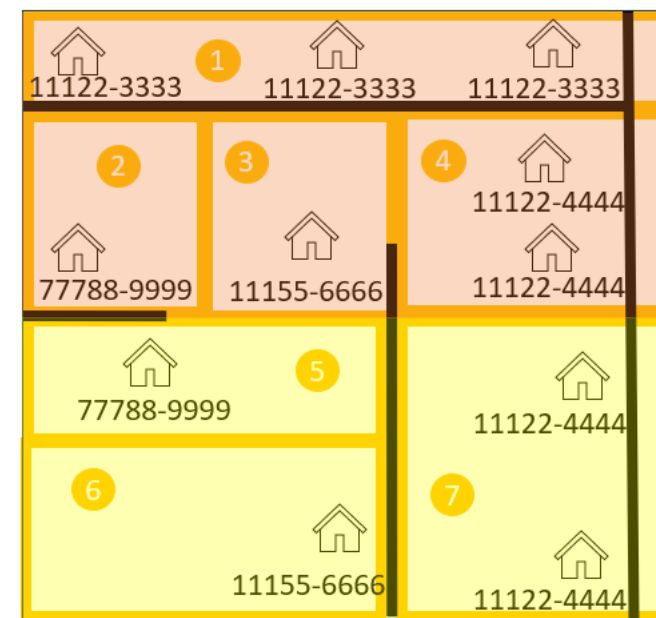
id	block	zip9	modeled estimate	zip9 sum	zip9 share
1	01-001-000001-0001	11122-3333	2	2	100%
2	01-001-000001-0001	77788-9999	2	8	25%
3	01-001-000001-0001	11155-6666	1	1	100%
4	01-001-000001-0001	11122-4444	1	2	50%
5	01-001-000001-0002	77788-9999	6	8	75%
6	01-001-000001-0002	11155-6666	0	1	0%
7	01-001-000001-0002	11122-4444	1	2	50%



Utilization of Machine Learning Results (Part 4 of 5)

Hypothetical Example

id	block	zip9	modeled estimate	zip9 sum	zip9 share
1	01-001-000001-0001	11122-3333	2	2	100%
2	01-001-000001-0001	77788-9999	2	8	25%
3	01-001-000001-0001	11155-6666	1	1	100%
4	01-001-000001-0001	11122-4444	1	2	50%
5	01-001-000001-0002	77788-9999	6	8	75%
6	01-001-000001-0002	11155-6666	0	1	0%
7	01-001-000001-0002	11122-4444	1	2	50%



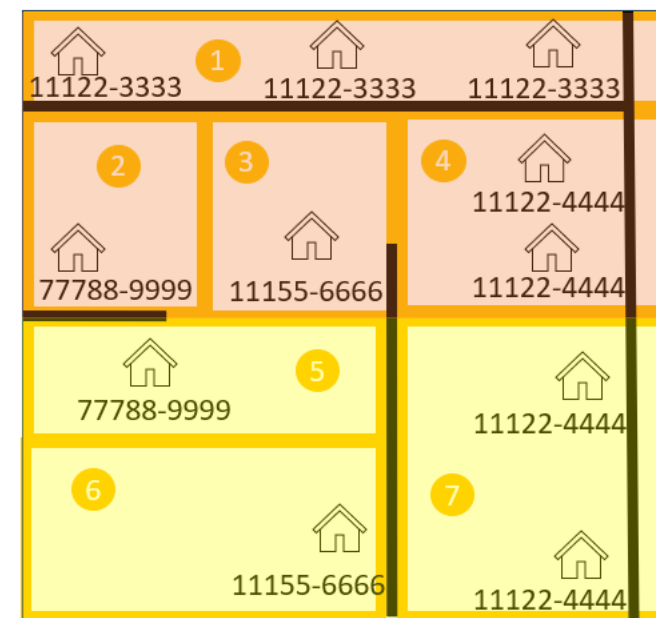
Non-geocoded individual tax record with ZIP code **11155-6666**

will have a geocode imputation of 01-001-000001-0001 100% of the time.

Utilization of Machine Learning Results (Part 5 of 5)

Hypothetical Example

id	block	zip9	modeled estimate	zip9 sum	zip9 share
1	01-001-000001-0001	11122-3333	2	2	100%
2	01-001-000001-0001	77788-9999	2	8	25%
3	01-001-000001-0001	11155-6666	1	1	100%
4	01-001-000001-0001	11122-4444	1	2	50%
5	01-001-000001-0002	77788-9999	6	8	75%
6	01-001-000001-0002	11155-6666	0	1	0%
7	01-001-000001-0002	11122-4444	1	2	50%



Non-geocoded individual tax record with ZIP code 77788-9999

will have a geocode imputation of 01-001-000001-0001 25% of the time
and 01-001-000001-0002 75% of the time.

Methods to Compare Frameworks

Measuring Precision

- Use random distribution to impute missing geocodes 10 times
- Quantify the average value of the share of related children in poverty, its percent error, absolute deviation and average deviation

Measuring Accuracy

- Remove the geocode from a different 20% of the individual tax records that are geocoded five times, impute geocodes and determine if there is a difference between the real and imputed result for each geocoded individual tax record
- Quantify the portion of correctly geocoded individual tax records at the level of the Census block, tract and school district

Discussion of Results

- Best performing machine learning models
 - Random forest
 - k-Nearest neighbors
- Overall precision of traditional versus machine learning model
 - Machine learning model is guaranteed to be more precise
- Overall accuracy of traditional versus machine learning model
 - Machine learning model accuracy under development
 - Traditional model performed better than expected

Conclusion

- Reflections: Developing a machine learning workgroup is a great way to empower employees to utilize data science training to solve problems
- Next Steps: Utilize more known information on family structure
- Challenges: Model and input processing is computationally expensive
- Contact Information:
 - Kate Willyard, Small Area Estimates Branch, U.S. Census Bureau
 - Katherine.a.Willyard@census.gov

References

- Cruz, Paula, Leonardo Vanneschi, Marco Painho, and Paulo Rita. “Automatic Identification of Addresses: A Systematic Literature Review.” *ISPRS International Journal of Geo-Information*, Vol. 11, Issue 1, December 29, 2021, P11. <https://doi.org/10.3390/ijgi11010011>
- Hibbert, James, Angela Liese, Andrew Lawson, Dwayne Porter, Robin Puett, Debra Standiford, Lenna Liu and Dana Dabelea. “Evaluating Geographic Imputation Approaches for Zip Code Level Data: An Application to a Study of Pediatric Diabetes.” *International Journal of Health Geographics*, Vol. 8, Issue 54, October 8, 2009. <http://www.ij-healthgeographics.com/content/8/1/54>
- Henry, Kevin A., and Francis P. Boscoe. “Estimating the Accuracy of Geographical Imputation.” *International Journal of Health Geographics*, Vol. 7, Issue 3, January 23, 2008. <https://link.springer.com/article/10.1186/1476-072X-7-3>
- Hurley, Susan E., TM. Saunders, R. Nivas, A. Hertz, P. Reynolds. “Post Office Box Addresses: A Challenge for Geographic Information System-Based Studies.” *Epidemiology*, Vol. 14, Issue 4, July 2003, P386-391. DOI: 10.1097/01.EDE.0000073161.66729.89.
- Lan Luo, Sara McLafferty, Fahui Wang. “Analyzing Spatial Aggregation Error in Statistical Models of Late-Stage Cancer Risk: a Monte Carlo Simulation Approach.” *International Journal of Health Geographics*, Vol. 9, Issue 51, October 19, 2010. <https://link.springer.com/article/10.1186/1476-072X-9-51>
- Lee, Kangjae, Alexis Richard C. Claridades, and Jiyeong Lee. “Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques.” *Applied Sciences*, Vol. 10, August 13, 2020, P5628,. <https://doi.org/10.3390/app10165628>
- Song, Lin, Laina Mercer, Jon Wakefield, Amy Laurent, David Solet. “Using Small-Area Estimation to Calculate the Prevalence of Smoking by Subcounty Geographic Areas in King County, Washington, Behavioral Risk Factor Surveillance System, 2009-2013.” *Preventing Chronic Disease*, Vol. 13, Issue 5, May 5, 2016. <http://dx.doi.org/10.5888/pcd13.150536>
- U.S. Census Bureau. “Standard Hierarchy of Census Geographic Entities.” 2020. <https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf>

Thank You Machine Learning Workgroup and Supporters!

- Literature review sub team
 - Albert Nedelman
 - Amelia Ingram
 - Angelica Phillips
- Input data and current process evaluation sub team
 - Ming-Ray Liao
 - Sam Shirazi
- Machine learning sub team
 - Angelica Phillips
 - Jadvir Kaur Gill
 - James Ho Shek
 - Mark Frame
 - Ming-Ray Liao
- Senior SEHSD leaders
 - Alfred Gottschalck
 - Carolyn Gann
 - David Powers
 - David Waddington
 - James Mouser
 - Jasen Taciak
 - Sandy Dietrich
 - Wesley Basel
- Senior CSRM leaders
 - Jerry Maples
 - Ryan Janicki
 - Scott Holan
 - Soumendra Lahiri
 - William Bell