ESTIMATING EMPLOYMENT WITH NIGHTTIME LIGHTS AND TRANSPORTATION DATA USING MACHINE LEARNING

MARK FOLDEN
SENIOR PREDICTIVE ANALYTICS SPECIALIST
NCTCOG

MAY 13, 2025

I LOVE IT WHEN A PLAN COMES TOGETHER

PROBLEM TO SOLVE

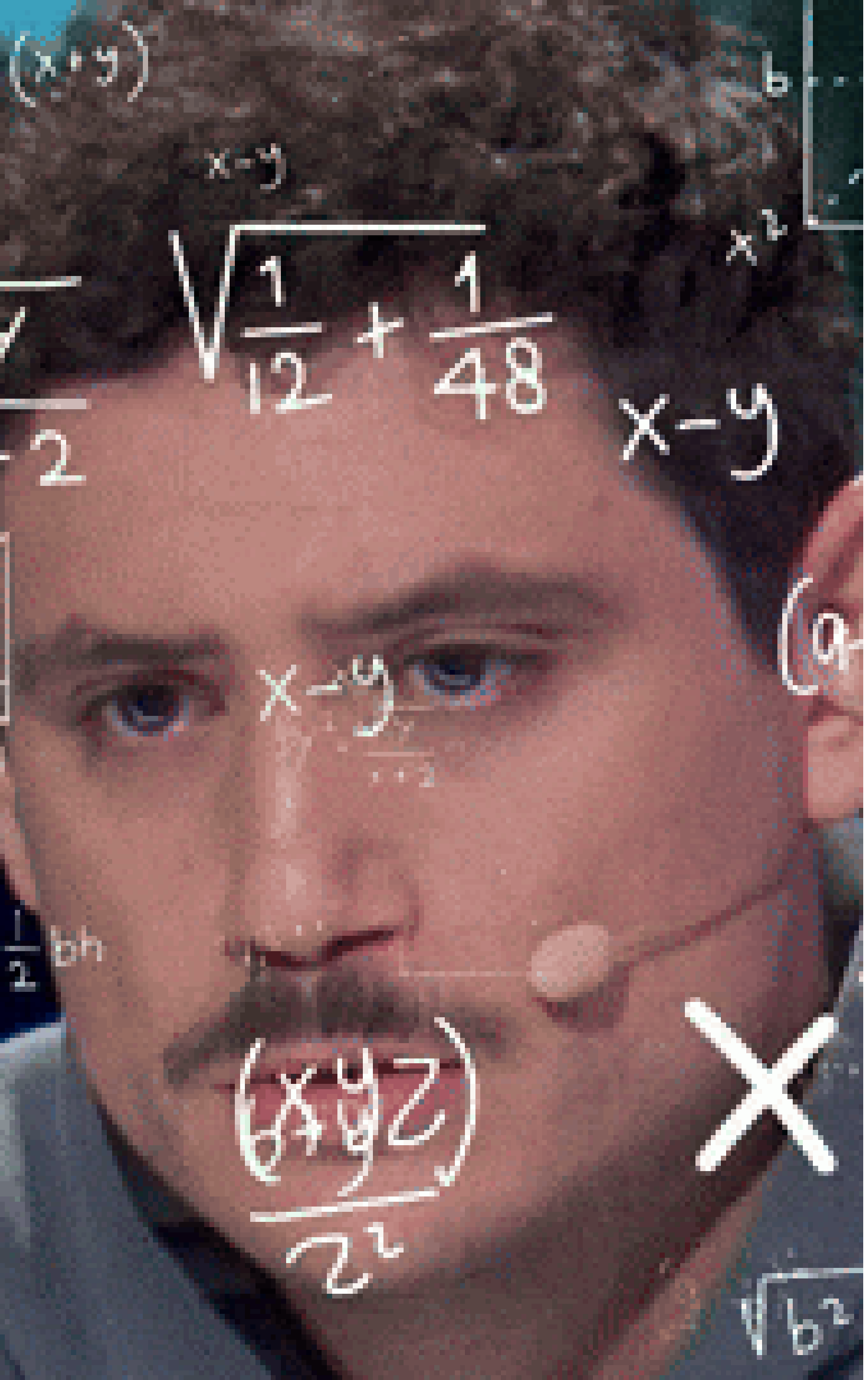
DATA & GIS PROCESSES

DATA SCIENCE PROCESS

TAKEAWAYS

FUTURE RESEARCH

Image Credit: https://giphy.com/gifs/a-team-a-team-imagines-john-hannibal-smith-8dx7Q9AXiMM24



PROBLEM TO SOLVE

- Small Area Estimates
- Sub-county estimates of Households, Population, Employment
- LODES is only published source of employment estimates; use as a starting point, but need to overcome several limitations
- Need a way to allocate to 30x30 meter grid; outputs should sum to the inputs, but be independent of zone structure
- Serve as starting point for Forecast
- Apply to past data and improve temporal consistency for Forecast model validation
- Apply in the future as new data becomes available, increased efficiency in generating updated data

Image Credit: https://giphy.com/gifs/newtral-problems-maths-daro-eme-hache-xFg0oNtbCCc4Epblyy

DATA & GIS PROCESSES

Small Area Estimates

- NCTCOG
- Already on 30x30m, but needs cleaning up

Landuse

- NCTCOG

Nighttime Lights

- Earth Observation Center, Colorado School of Mines

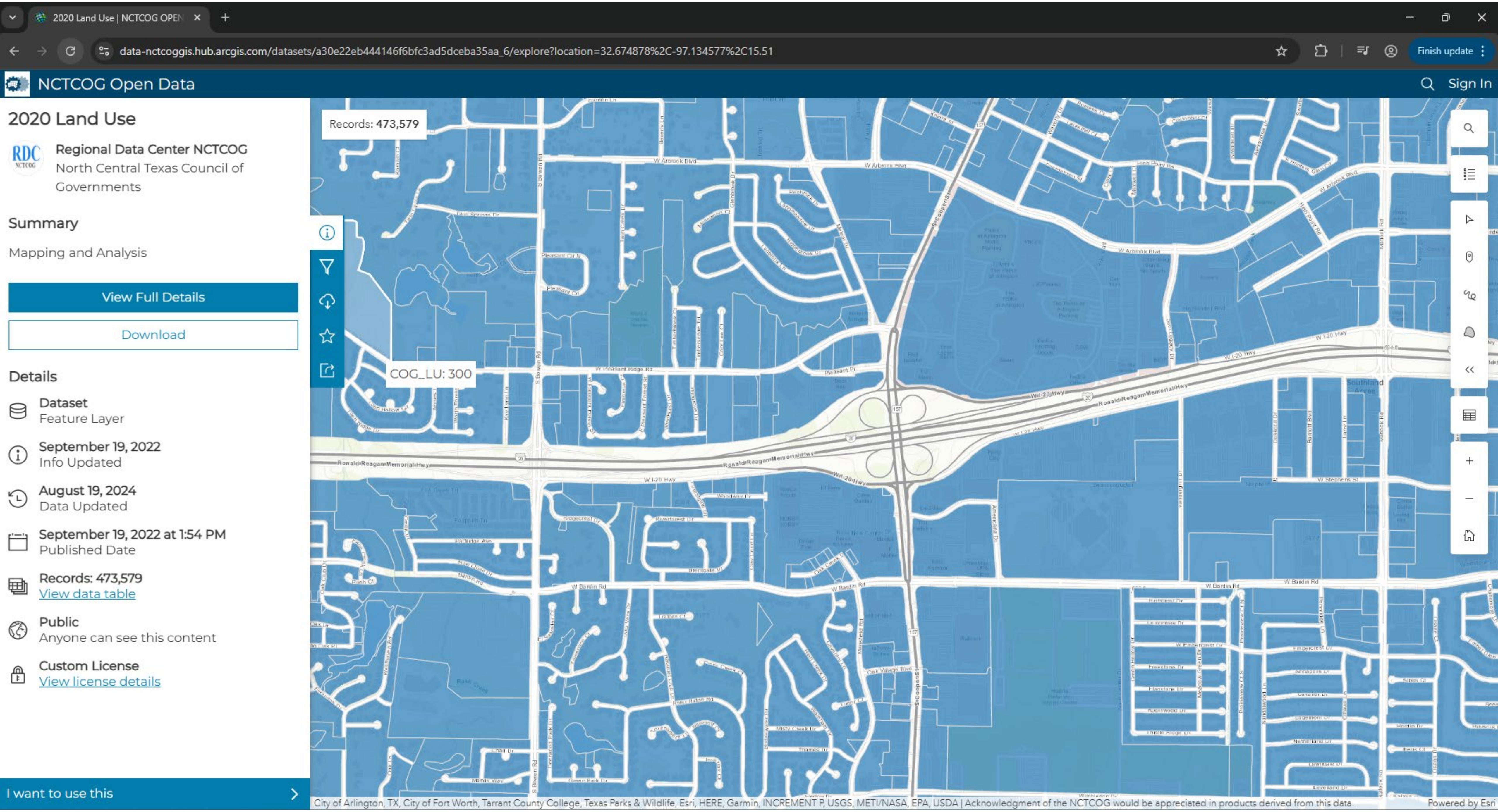
Roadway Network

- TxDOT

Traffic Counts

- TxDOT

https://giphy.com/gifs/nickelodeon-cook-bad-cooking-party-jBGhBiPXCFpxTB550m



2020 Land Use | NCTCOG OPEN
data-nctcoggis.hub.arcgis.com/datasets/a30e22eb444146f6bfc3ad5dceba35aa_6/explore?location=32.674878%2C-97.134577%2C15.51
Finish update
NCTCOG Open Data
Sign In
Records: 473,579
COG_LU: 300
W Arbrook Blvd
W Pleasant Ridge Rd
Ronald Reagan Memorial Hwy
W I-20 Hwy
W Bardin Rd
S Cooper St
S Bowen Rd
W Stephens St
Southland Acres
City of Arlington, TX, City of Fort Worth, Tarrant County College, Texas Parks & Wildlife, Esri, HERE, Garmin, INCREMENT P, USGS, METI/NASA, EPA, USDA | Acknowledgment of the NCTCOG would be appreciated in products derived from this data.
Powered by Esri

2020 Land Use

Regional Data Center NCTCOG
North Central Texas Council of Governments

Summary

Mapping and Analysis

View Full Details

Download

Details

Dataset
Feature Layer

September 19, 2022
Info Updated

August 19, 2024
Data Updated

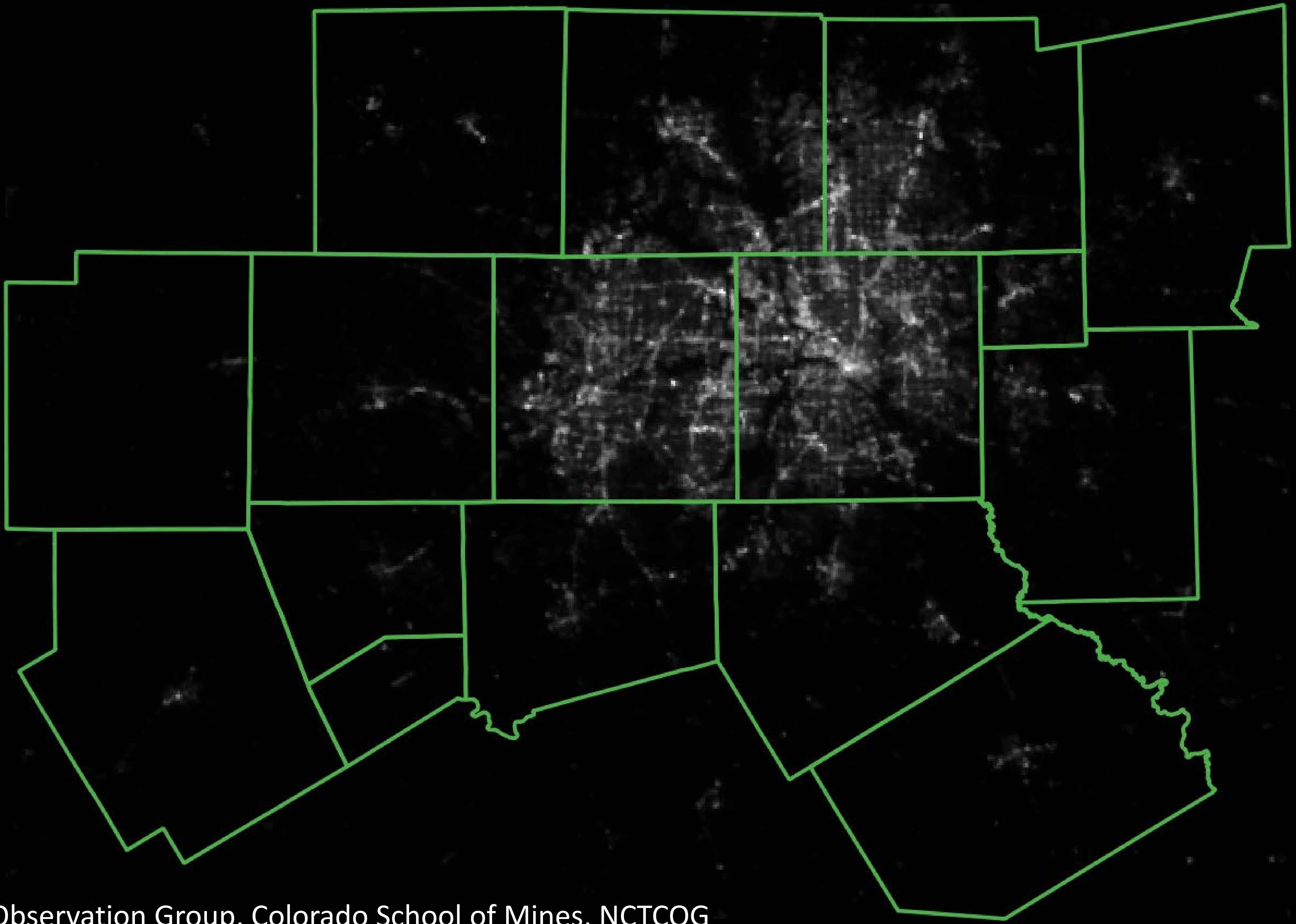
September 19, 2022 at 1:54 PM
Published Date

Records: 473,579
View data table

Public
Anyone can see this content

Custom License
View license details

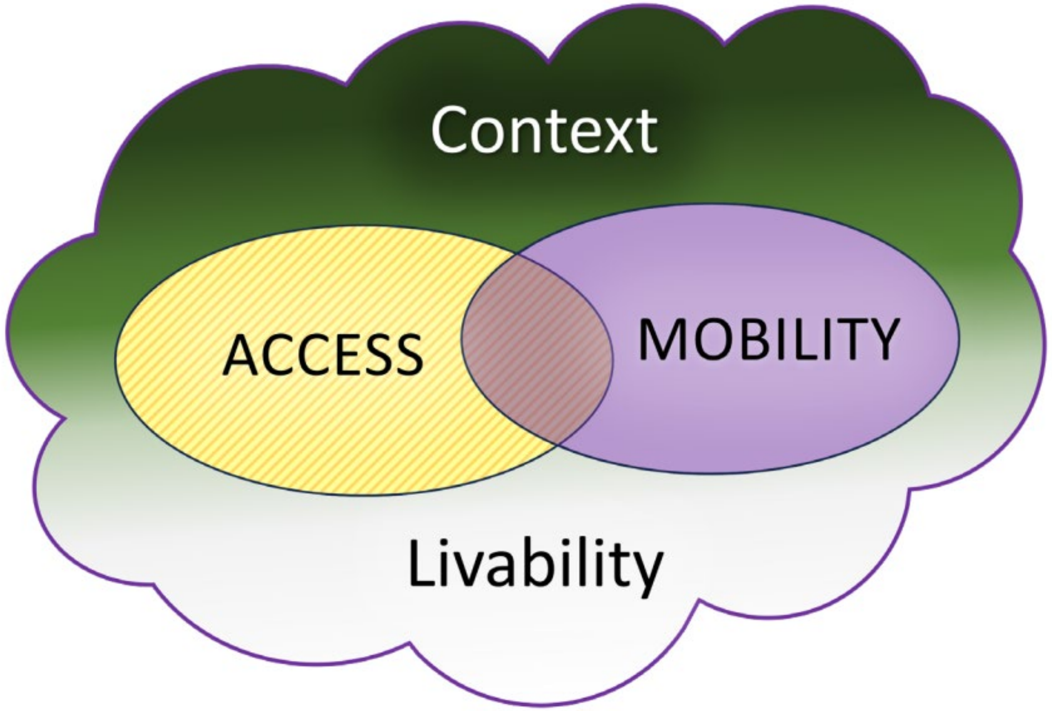
I want to use this



Source: Earth Observation Group, Colorado School of Mines, NCTCOG

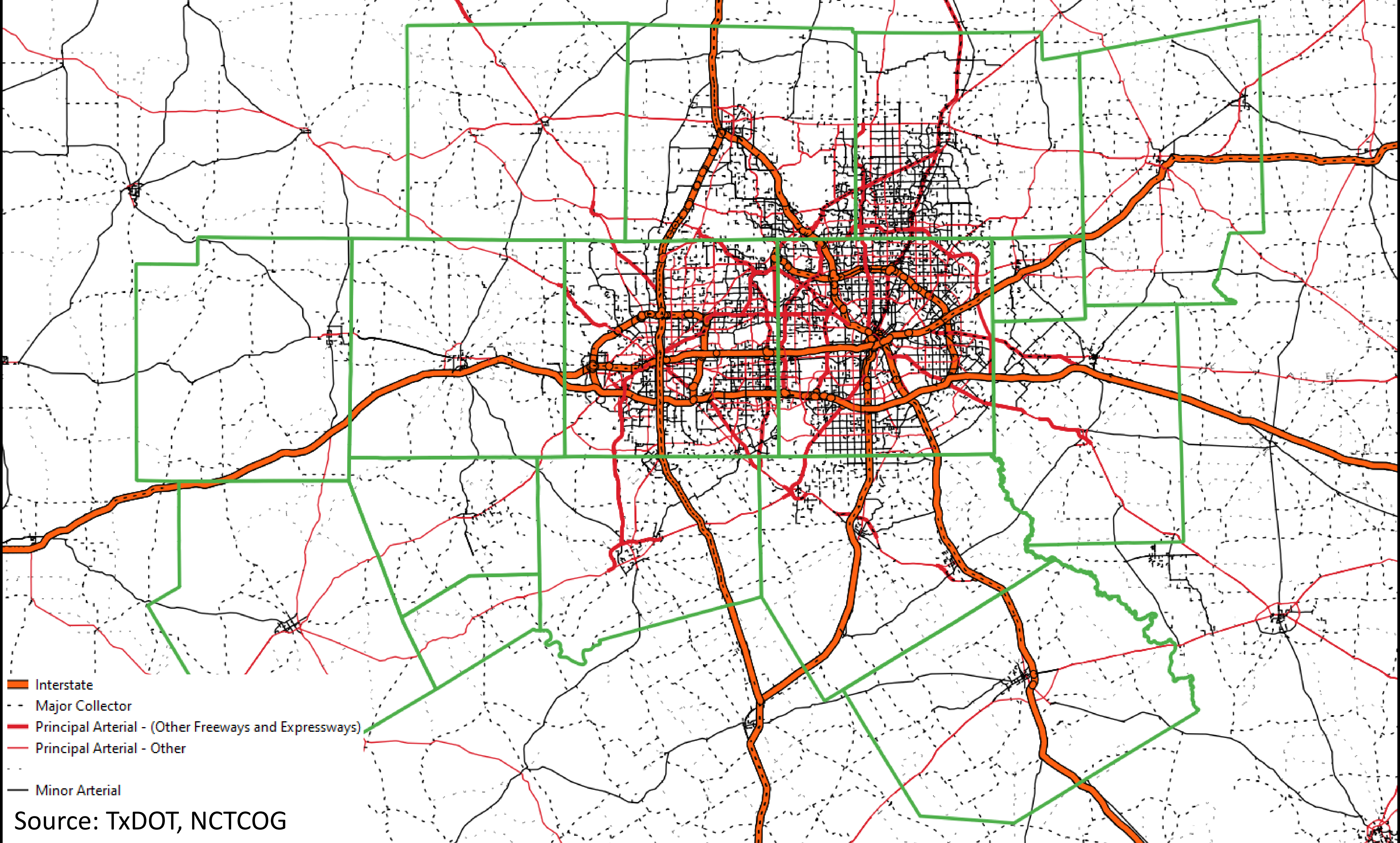
ROADWAY NETWORK: FUNCTIONAL CLASSIFICATION

- Transportation facilities can be classified on a continuum between interrelated goals
  - Access vs. Mobility
- For our purposes:
  - Mobility
    - Facility cannot have a driveway connected to it
    - Interstates
    - “Major Arterials – Other Freeways & Expressways”
  - Access
    - Facility can have a driveway connected directly to it

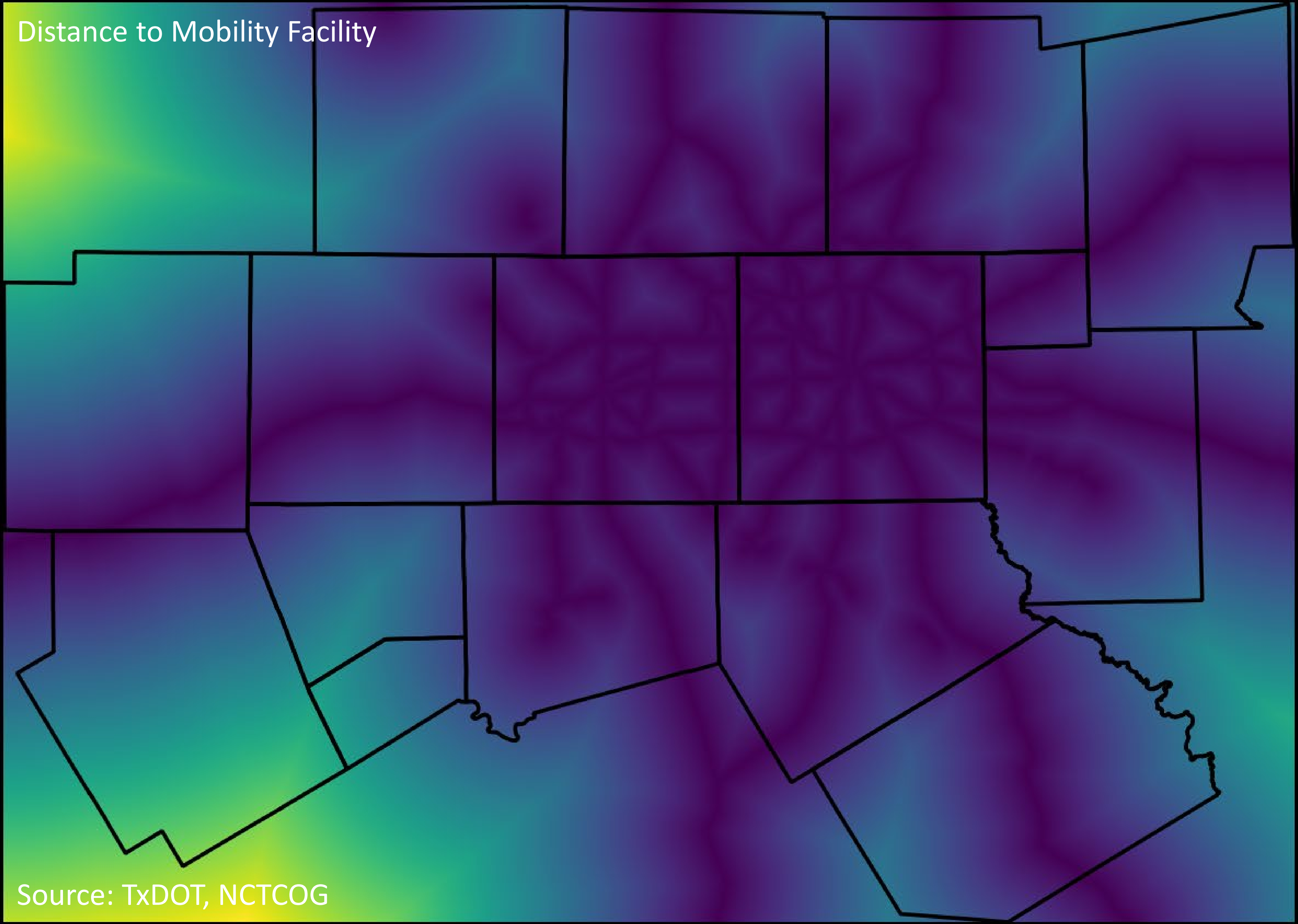


Source: FHWA

https://www.fhwa.dot.gov/planning/processes/statewide/related/hwy-functional-classification-2023.pdf

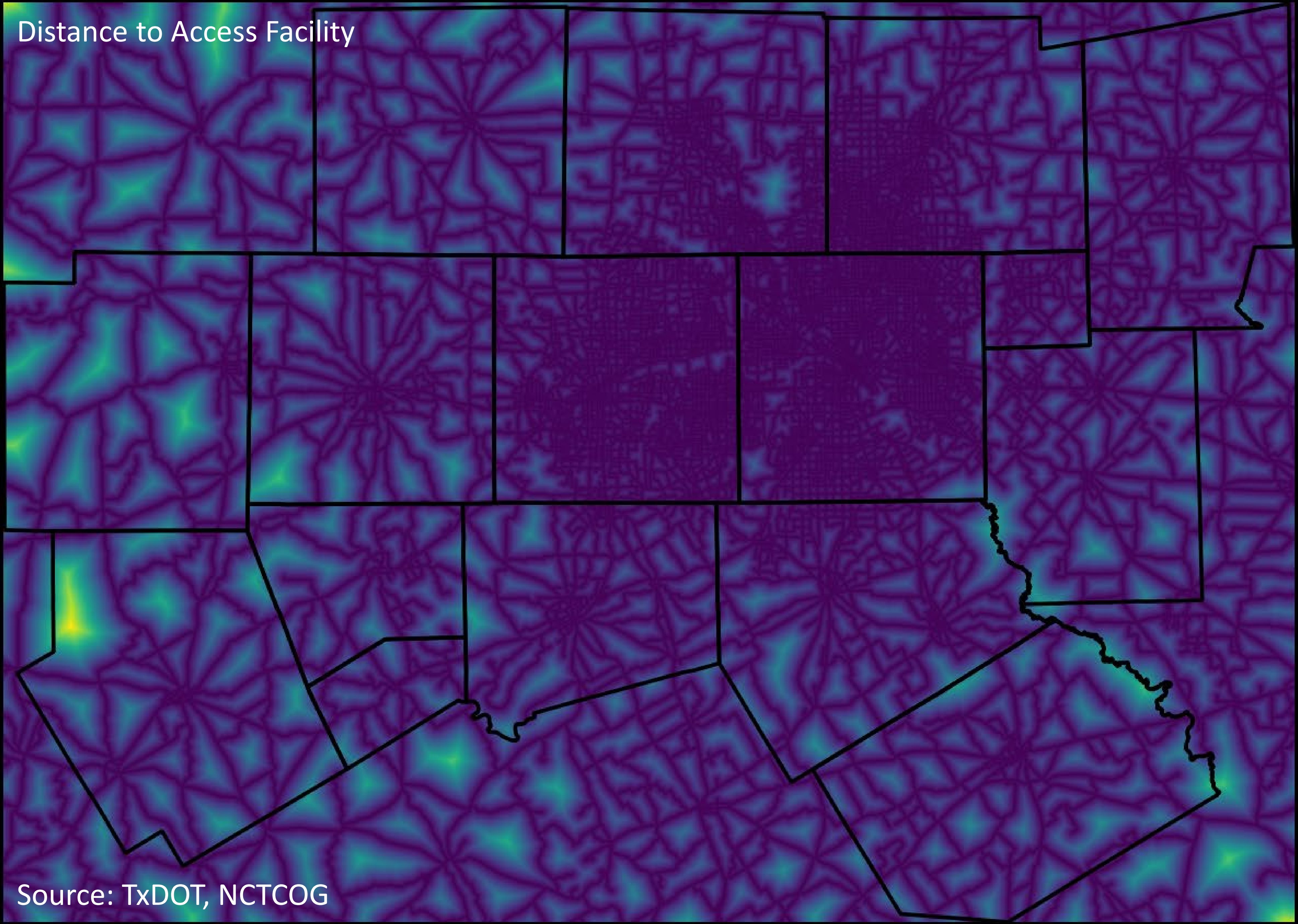

Interstate
Major Collector
Principal Arterial - (Other Freeways and Expressways)
Principal Arterial - Other
Minor Arterial

Source: TxDOT, NCTCOG



Distance to Mobility Facility

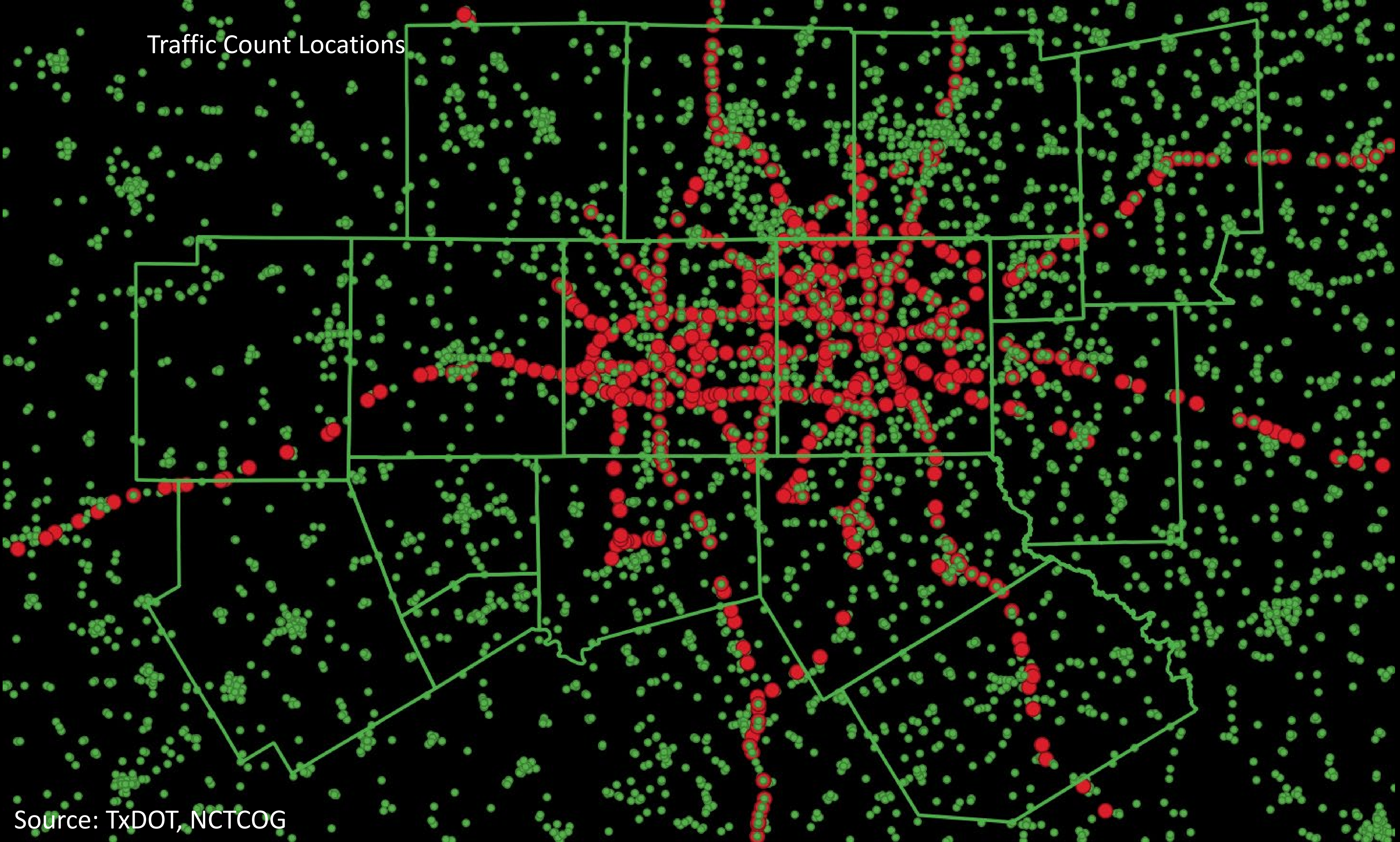
Source: TxDOT, NCTCOG



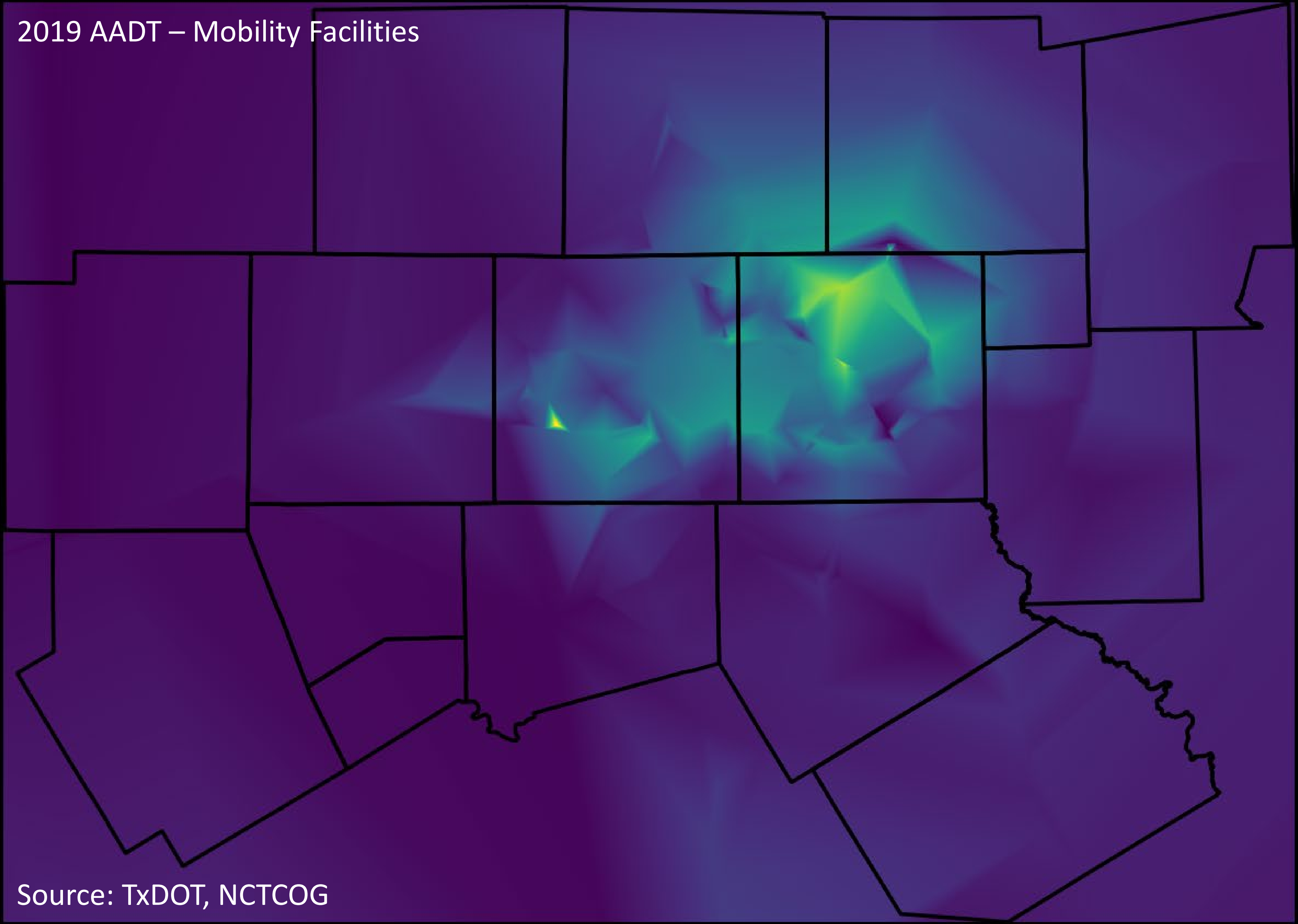
Distance to Access Facility

Source: TxDOT, NCTCOG

Traffic Count Locations



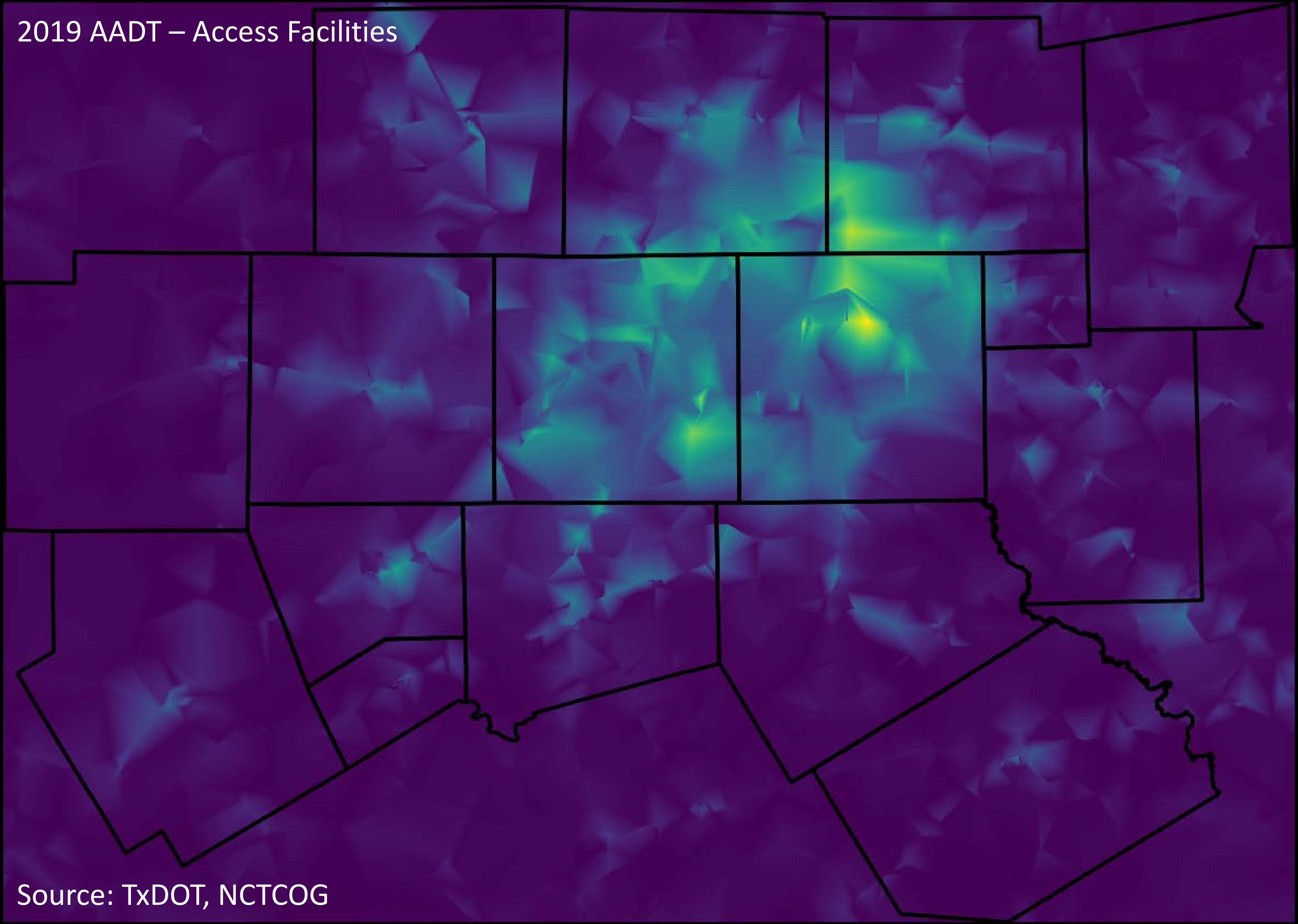
Source: TxDOT, NCTCOG



2019 AADT – Mobility Facilities

Source: TxDOT, NCTCOG

2019 AADT – Access Facilities



Source: TxDOT, NCTCOG

DATA SCIENCE PROCESS

CLEAN DATA

CHECK FOR NORMALITY

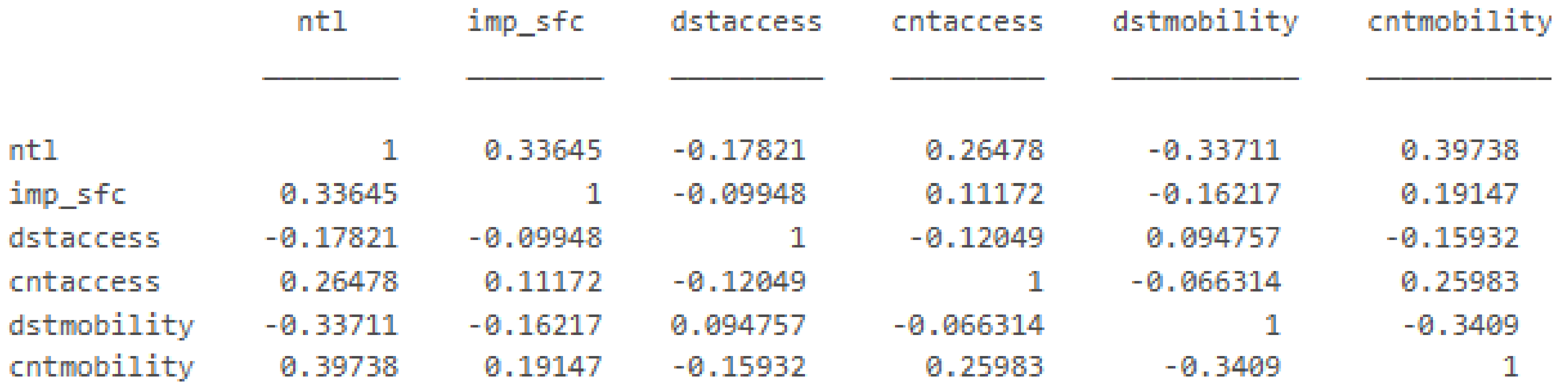
RESCALE (IF NECESSARY)

CHECK FOR CORRELATION

TRAIN ALGORITHM

VALIDATE & TEST

CORRELATION



```
                  ntl      imp_sfc    dstaccess    cntaccess    dstmobility    cntmobility
               ________    ________    _________    _________    ___________    ___________

ntl                   1     0.33645     -0.17821      0.26478       -0.33711        0.39738
imp_sfc         0.33645           1     -0.09948      0.11172       -0.16217        0.19147
dstaccess      -0.17821    -0.09948            1     -0.12049       0.094757       -0.15932
cntaccess       0.26478     0.11172     -0.12049            1      -0.066314        0.25983
dstmobility    -0.33711    -0.16217     0.094757    -0.066314              1        -0.3409
cntmobility     0.39738     0.19147     -0.15932      0.25983        -0.3409              1
```

Image Credit: https://www.kapwing.com/explore/3-spiderman-pointing-meme-template

ALGORITHMS

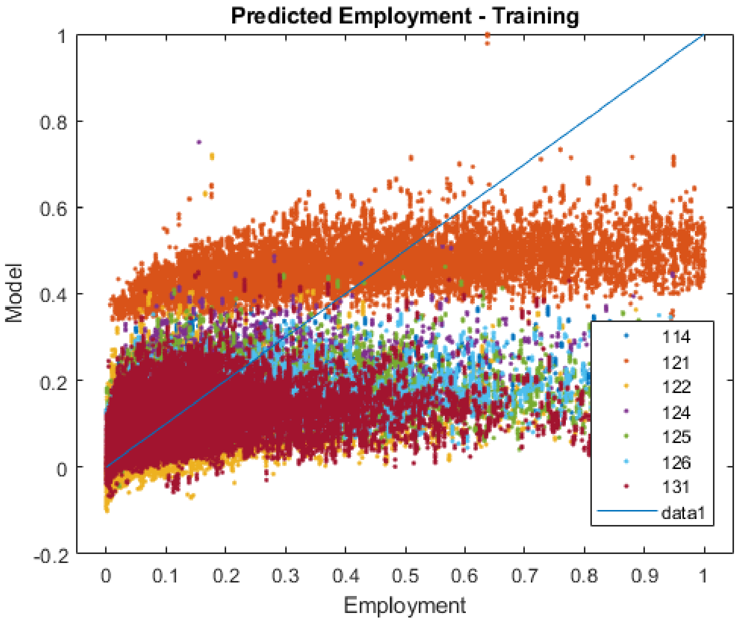
MATLAB 2022a with Statistics and Machine Learning Toolbox

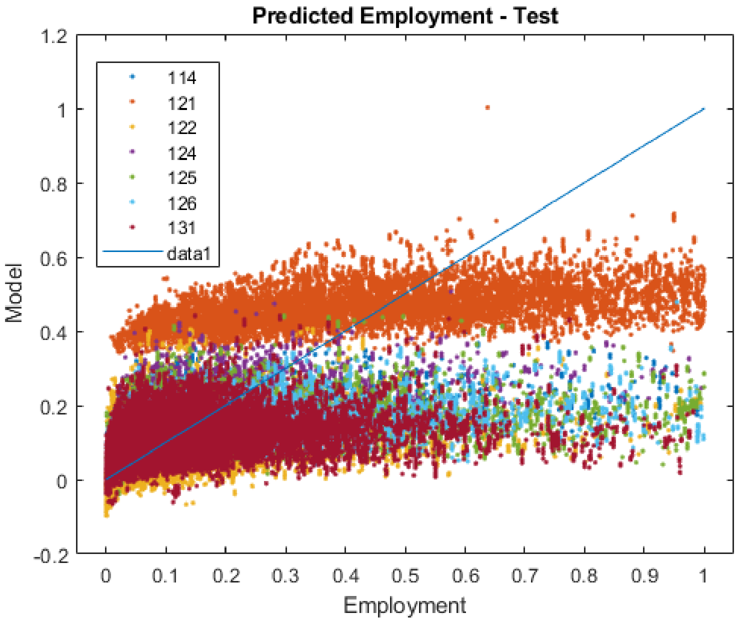
- Multivariate Linear Regression (Plain Vanilla)
- Support Vector Machine
- LS Boosted Ensemble
- Treebagger
- Bagged Ensemble
  - https://www.ibm.com/think/topics/bagging
  - https://www.mathworks.com/help/stats/ensemble-algorithms.html

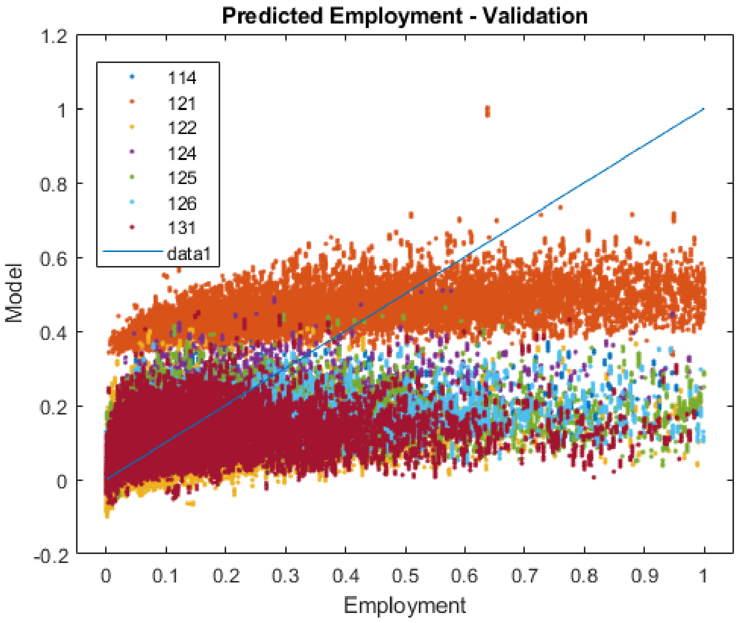
n = 1,143,918

- Training
  - 30% : 342,967
- Test
  - 20%: 228,935
- Validation
  - 50%: 572,016

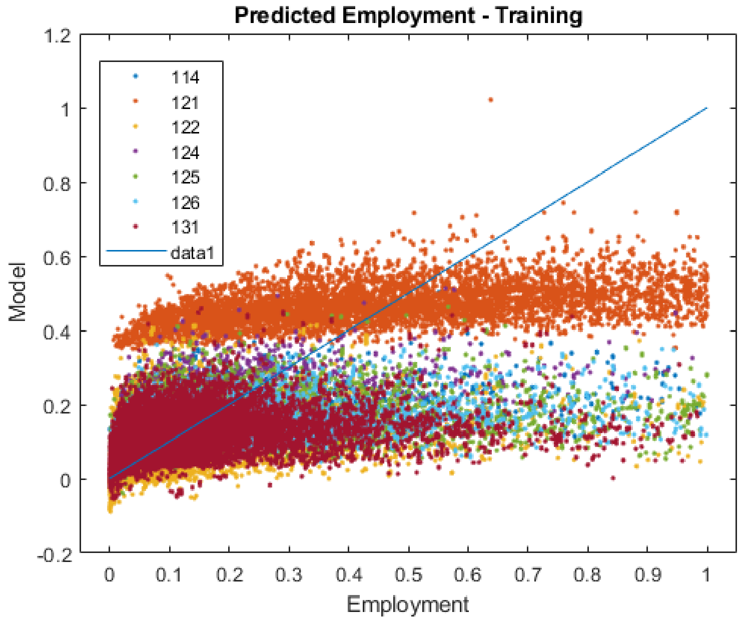
MULTIVARIATE LINEAR REGRESSION



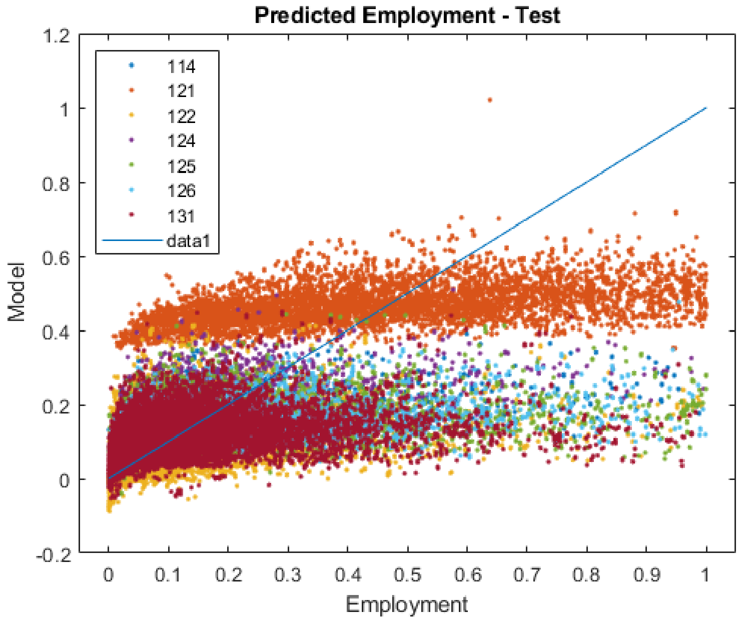




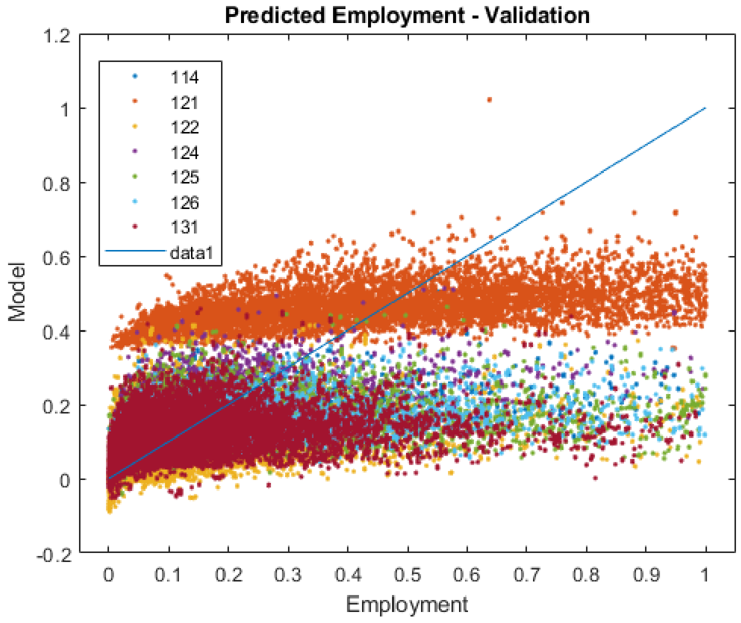
SUPPORT VECTOR MACHINE



Predicted Employment - Training

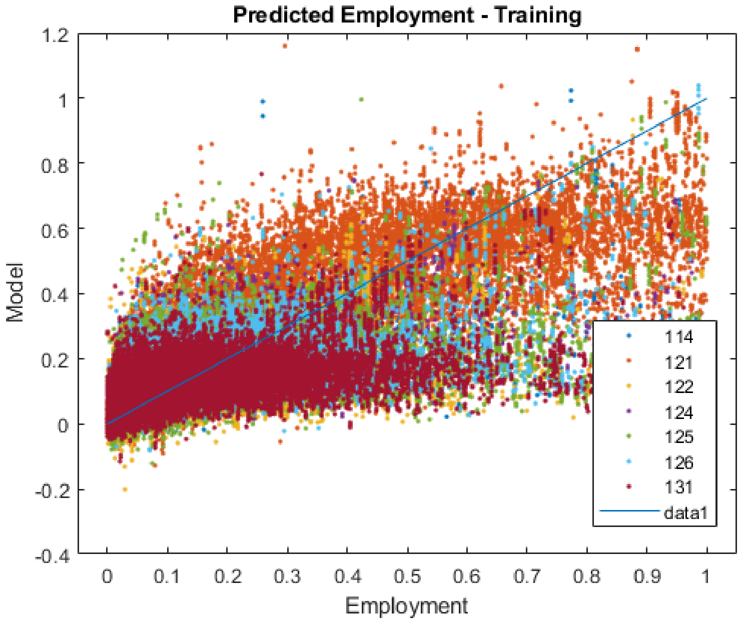


Predicted Employment - Test

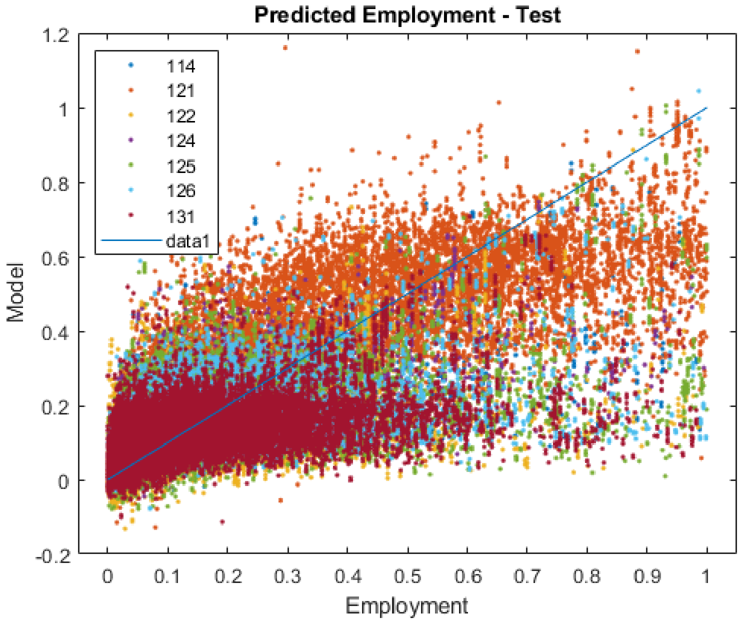


Predicted Employment - Validation

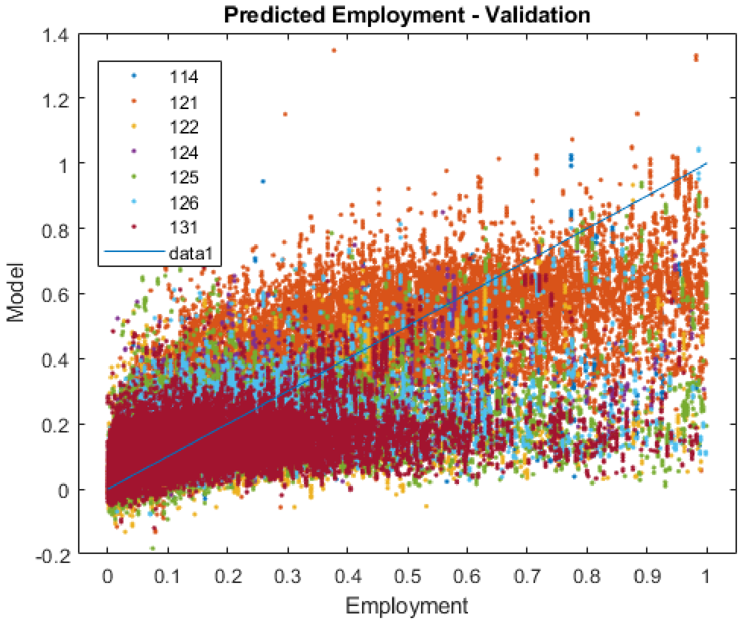
LS BOOSTED ENSEMBLE



Predicted Employment - Training

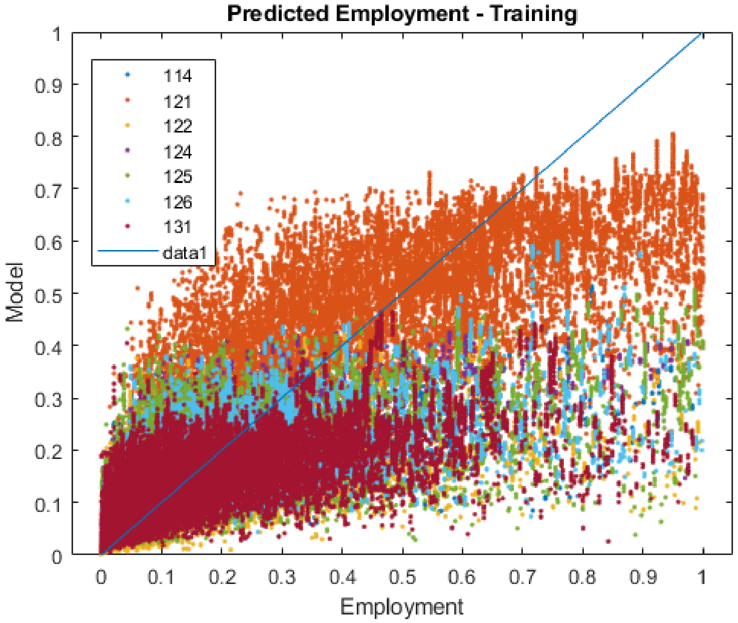


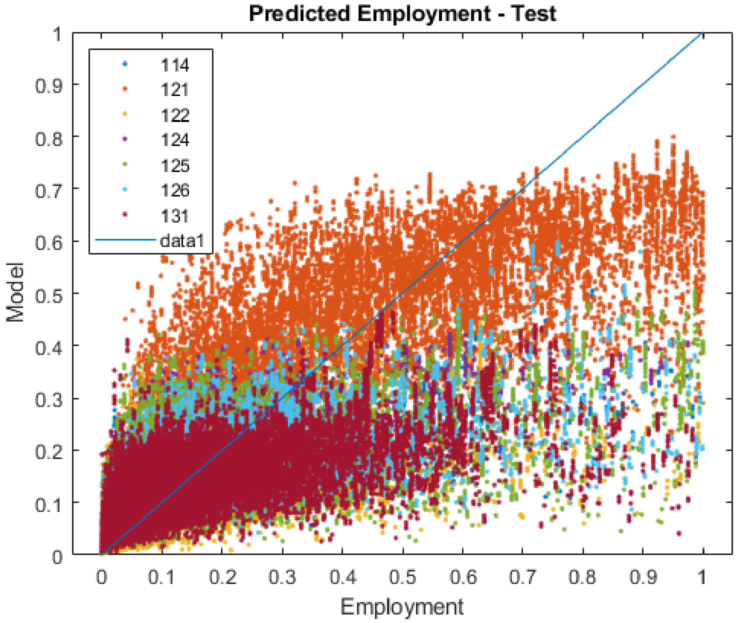
Predicted Employment - Test

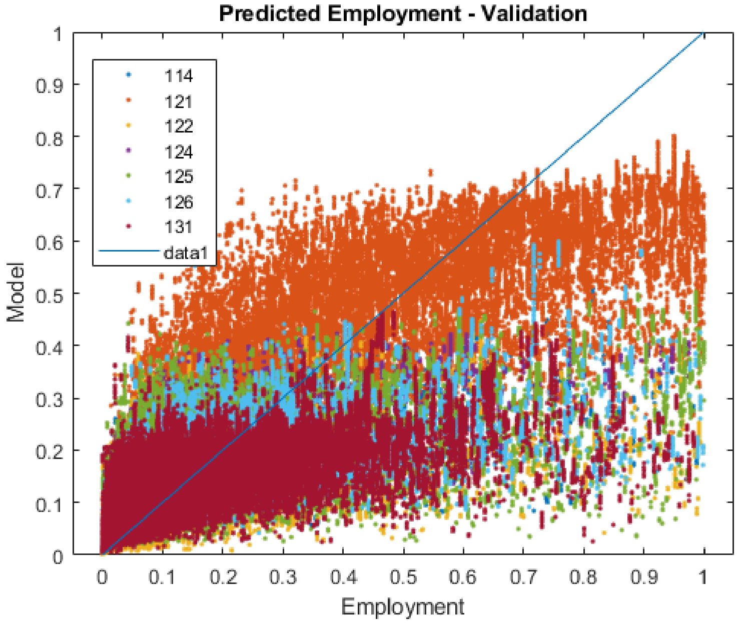


Predicted Employment - Validation

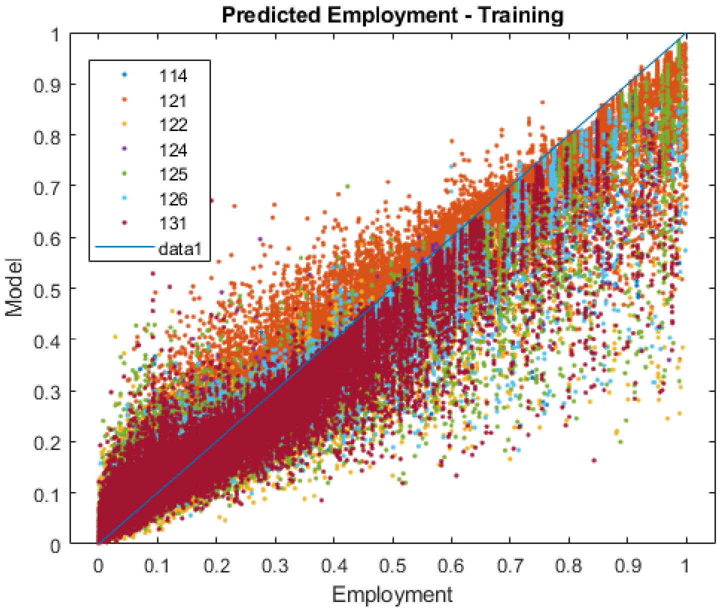
TREEBAGGER

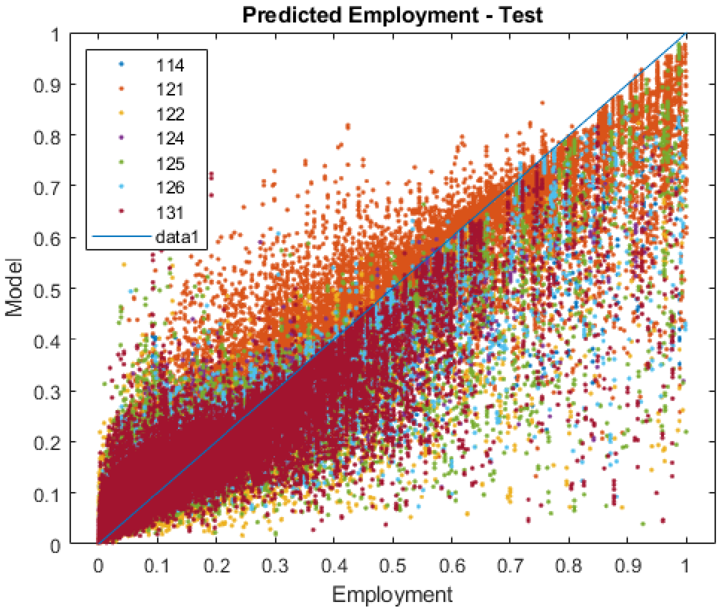

Predicted Employment - Training

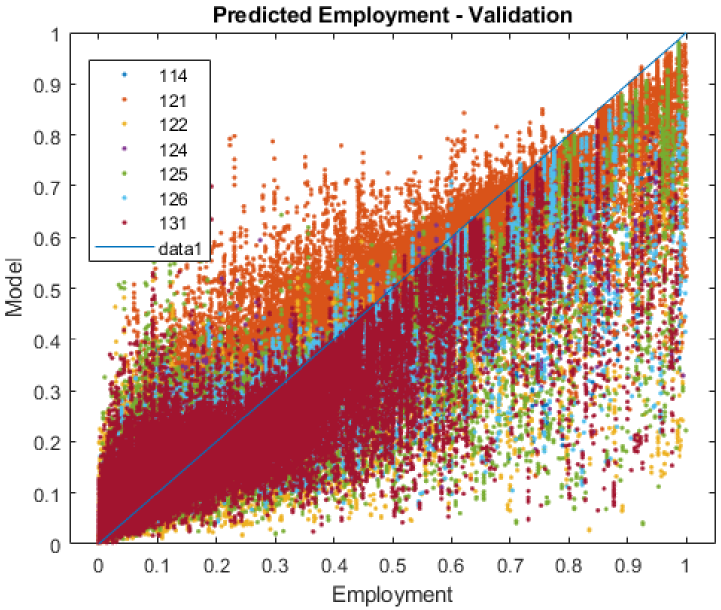

Predicted Employment - Test


Predicted Employment - Validation

BAGGED ENSEMBLE







SUMMARY

R-SQUARED

- R^2 closer to 1.0 the better
- Don’t want a large drop from Training to Test and Validation

Algorithm	Training	Test	Validation
Multivariate Linear Regression	0.65027	0.64930	0.65293
Support Vector Machine	0.64824	0.64711	0.65109
LS Boosted Ensemble	0.75222	0.74138	0.74352
Treebagger	0.80157	0.78811	0.78892
Bagged Ensemble	0.97062	0.93480	0.93508

TAKEAWAYS

Machine Learning techniques compared to traditional Multinomial Least Squares regression is a tradeoff between model performance and interpretability

You can offset some of that tradeoff with additional analysis and trying multiple methods

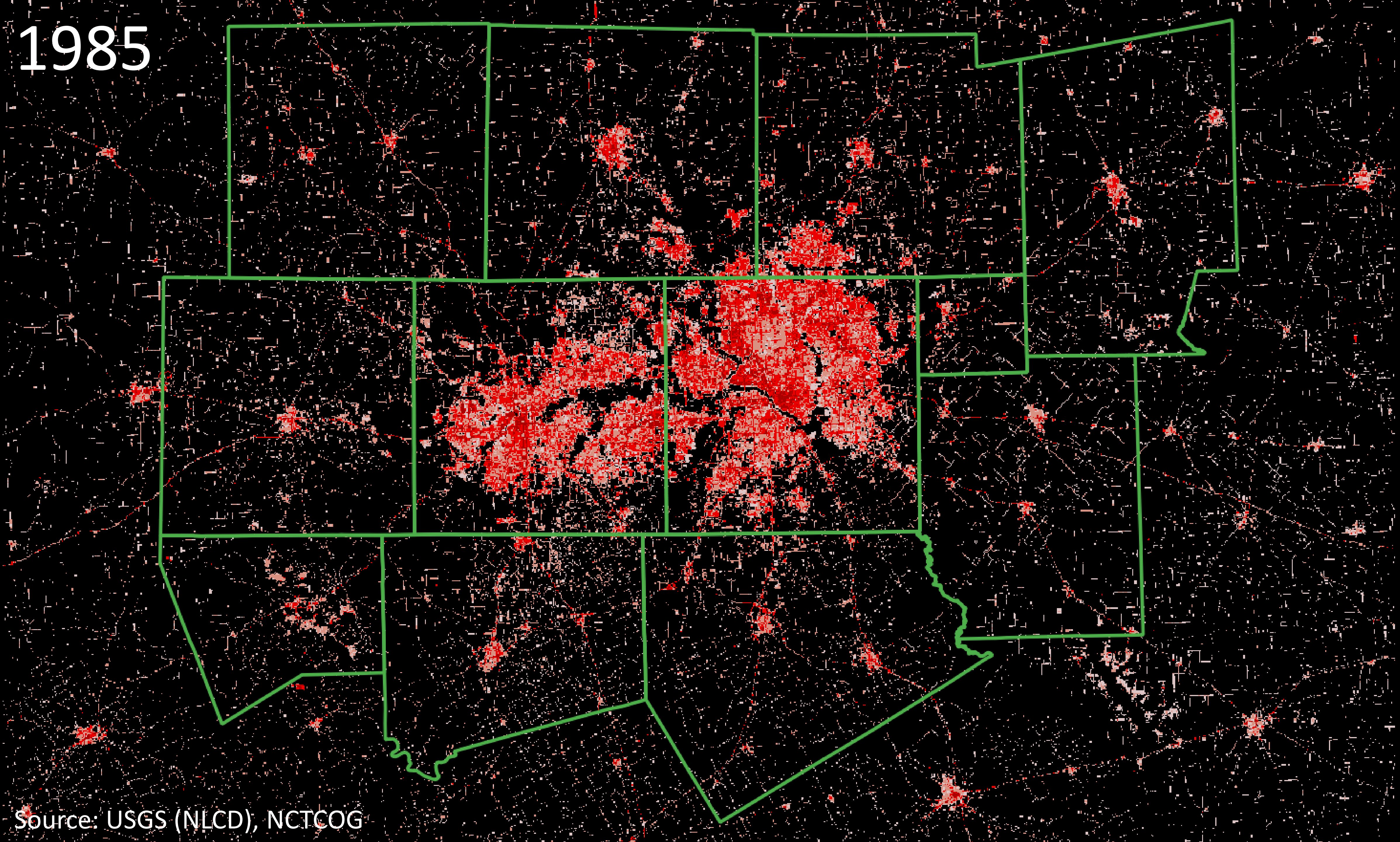
You might not know exactly how an algorithm works, but go ahead and try it and see if you get an improvement

Even two algorithms that work similarly can get you different results

Categorical variables can be VERY powerful

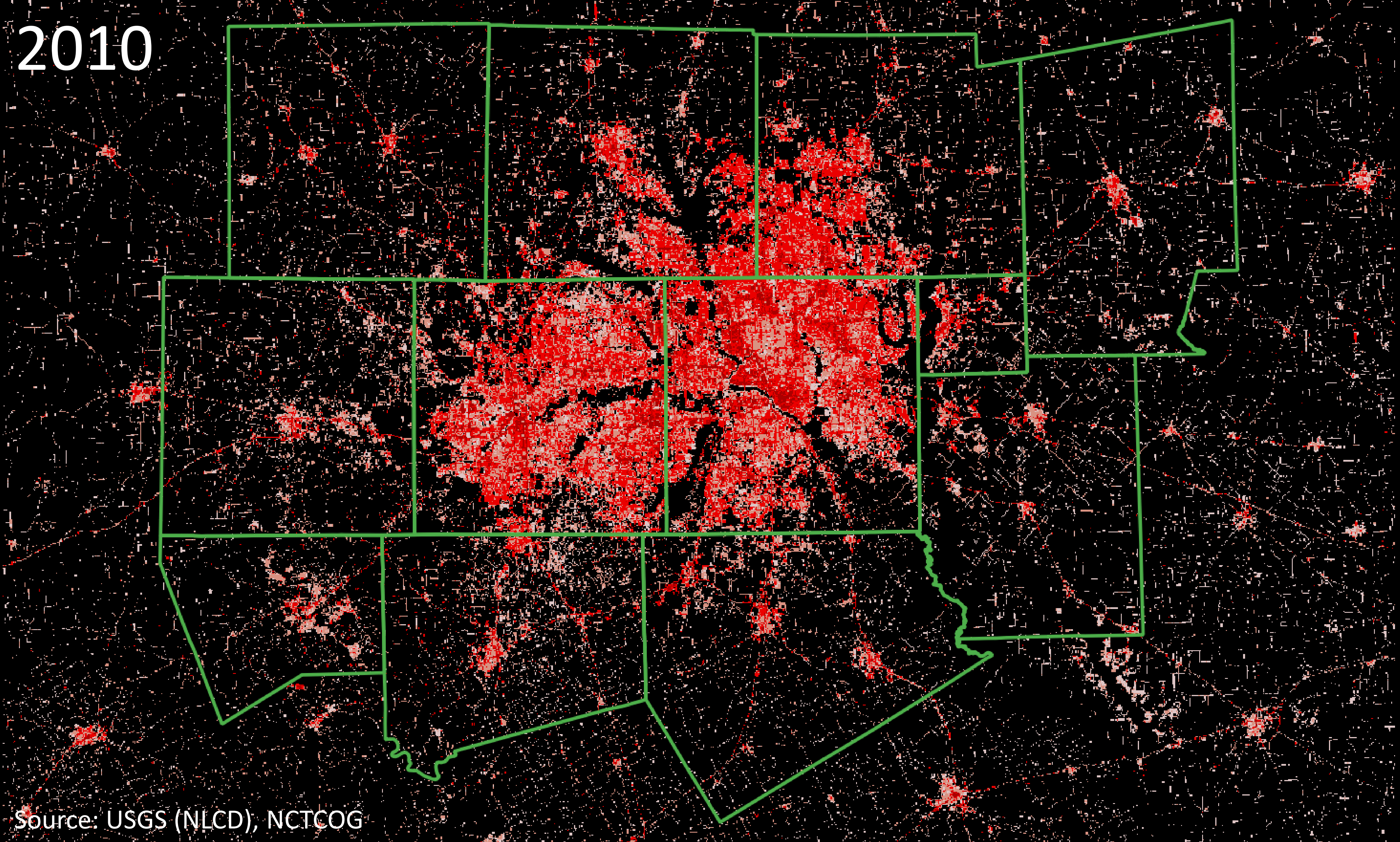
More data != better

Image: https://tenor.com/view/the-it-crowd-it-crowd-moss-moss-the-it-crowd-the-it-crowd-moss-gif-24976144

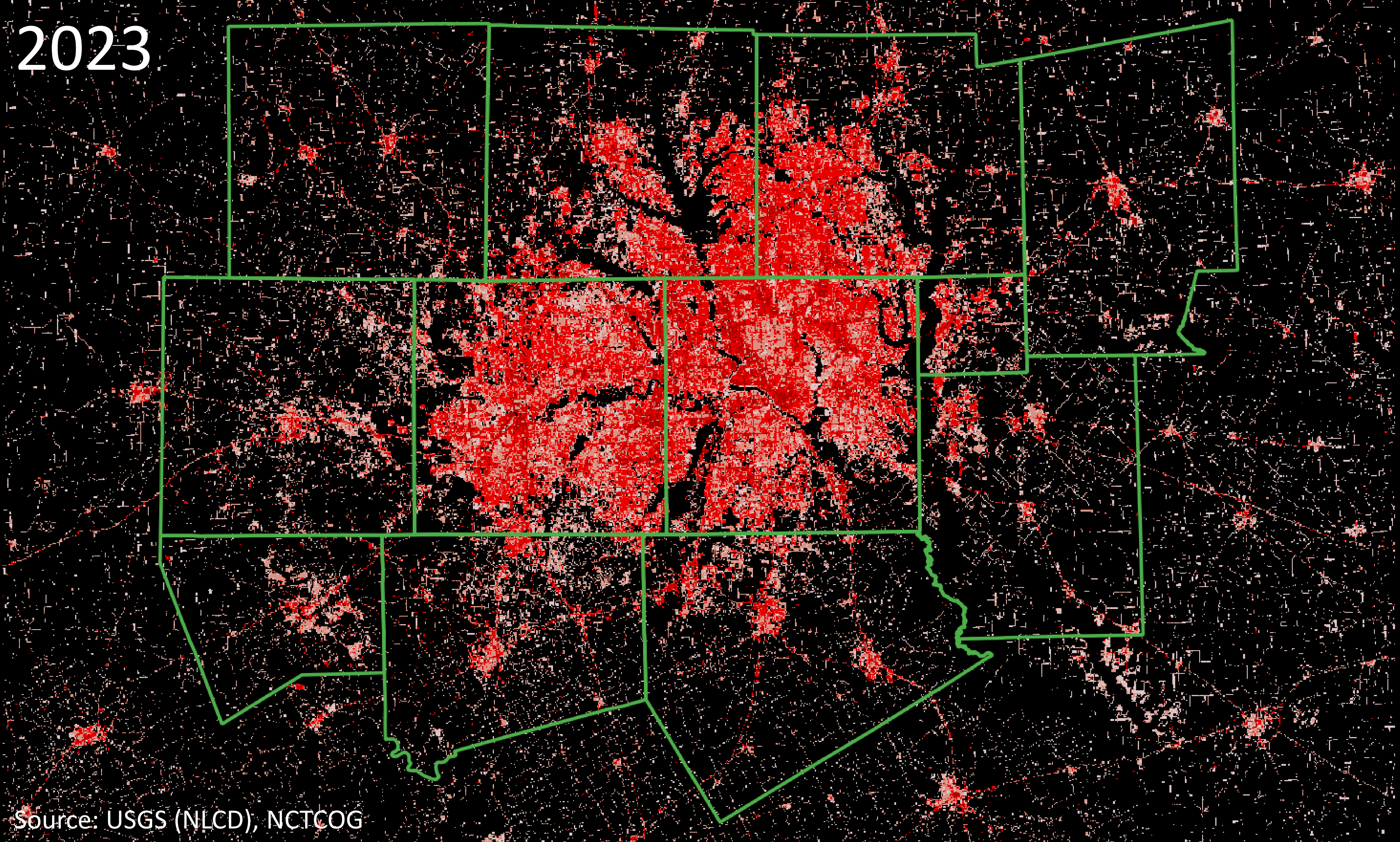


Source: USGS (NLCD), NCTCOG

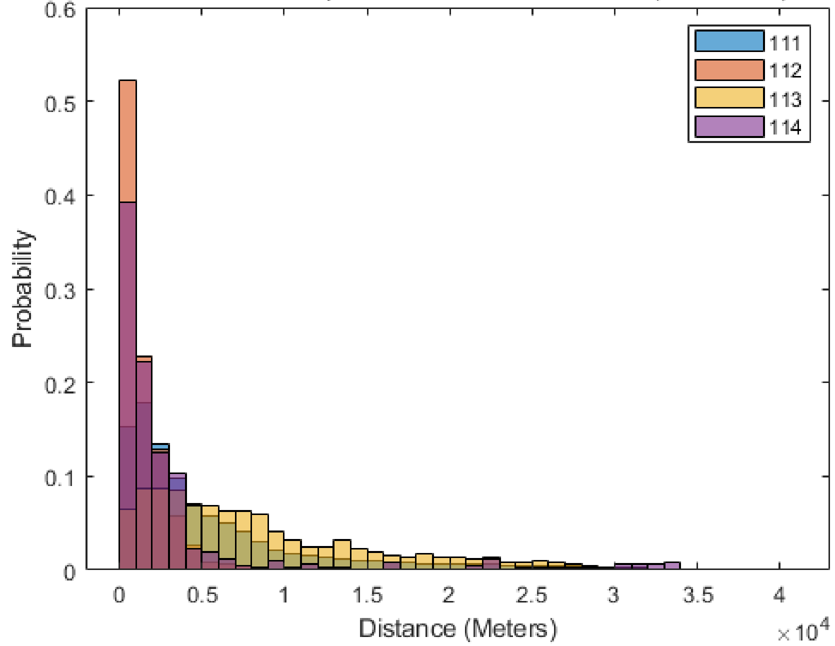
2010



Source: USGS (NLCD), NCTCOG

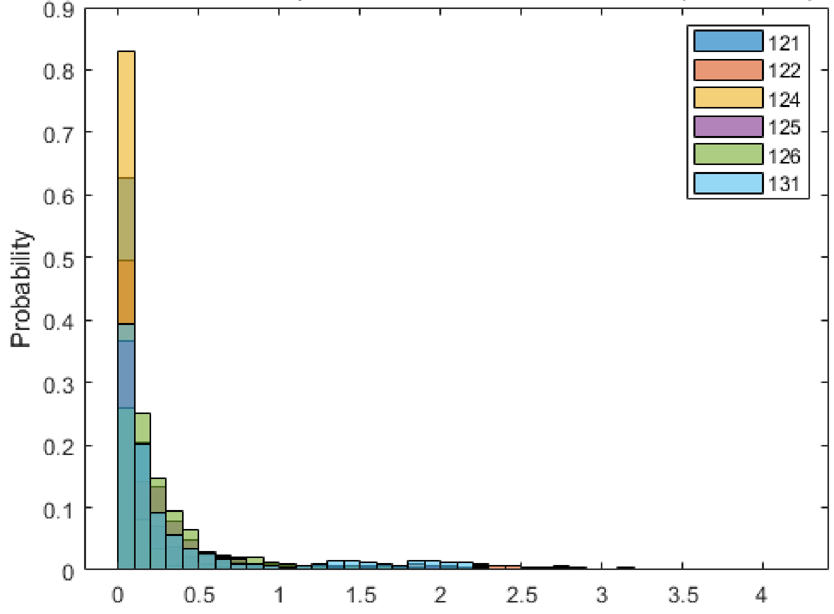


Source: USGS (NLCD), NCTCOG



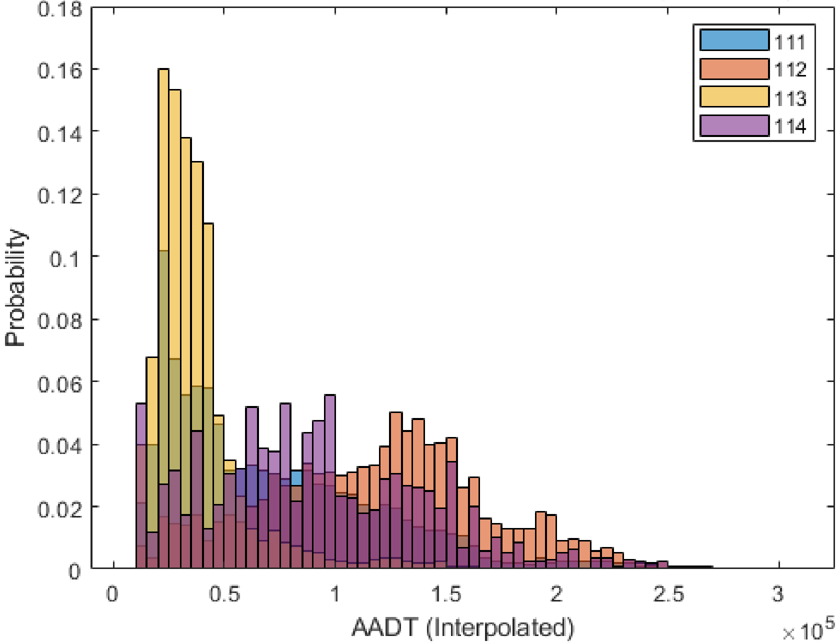
Distance to Mobility Classified Facility

New & Re-Development Residential Grid Cells (1985-2010)



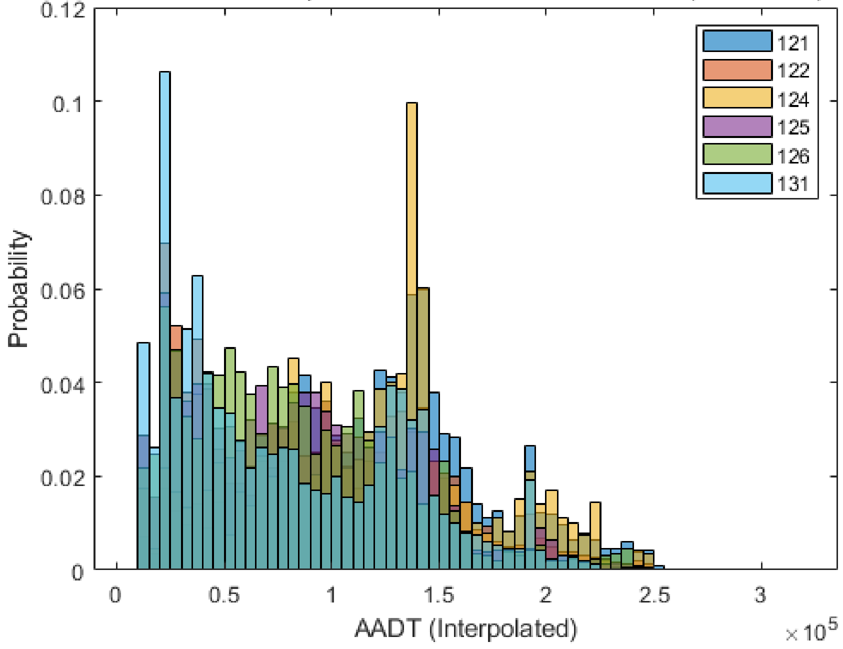
Distance to Mobility Classified Facility

New & Re-Development Non-Residential Grid Cells (1985-2010)



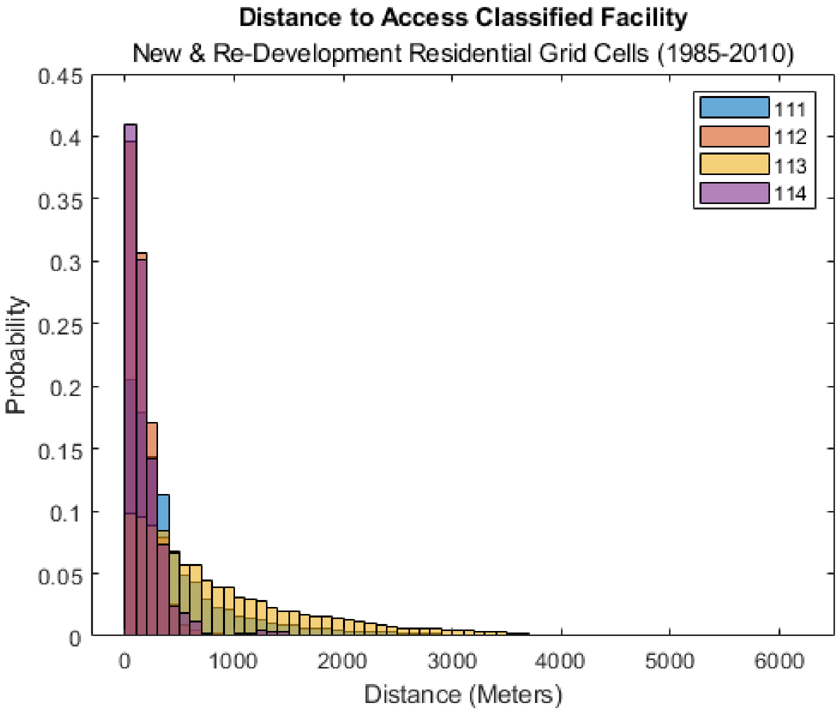
2010 AADT (Interpolated) for Mobility Classified Facility

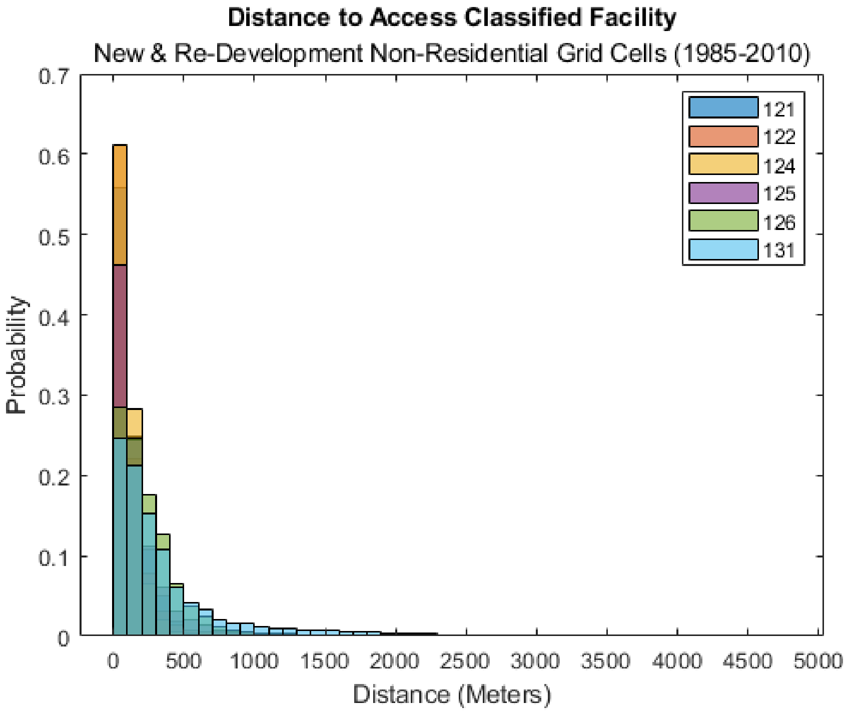
New & Re-Development Residential Grid Cells (1985-2010)

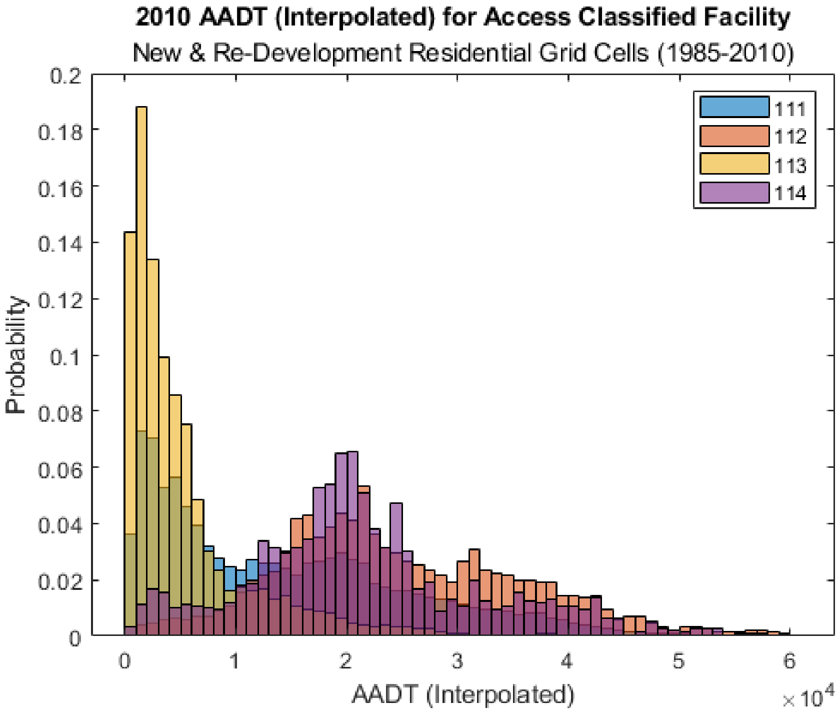


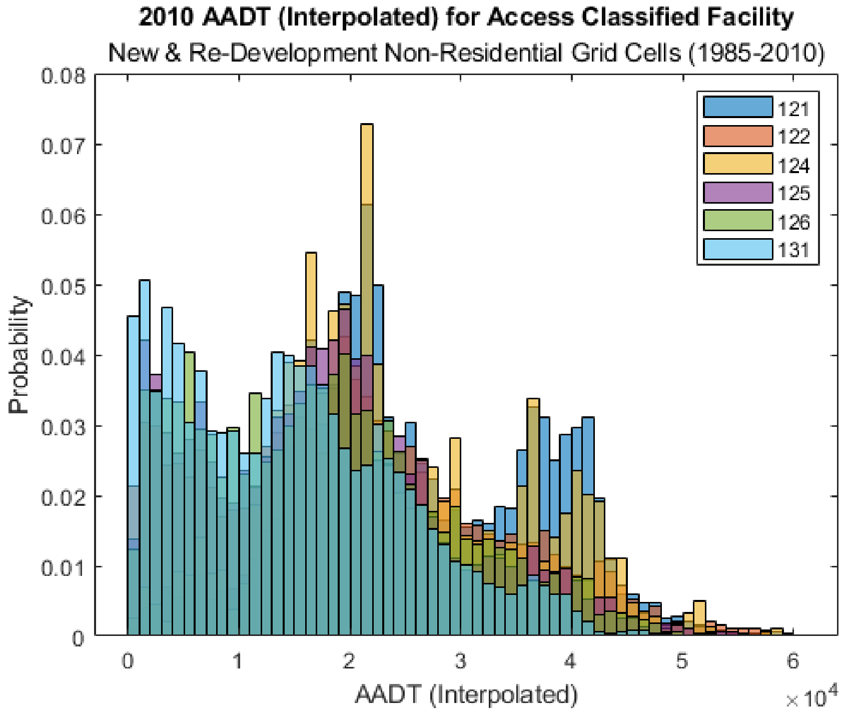
2010 AADT (Interpolated) for Mobility Classified Facility

New & Re-Development Non-Residential Grid Cells (1985-2010)


Distance to Access Classified Facility
New & Re-Development Residential Grid Cells (1985-2010)
Probability
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Distance (Meters)
0
1000
2000
3000
4000
5000
6000
111
112
113
114


Distance to Access Classified Facility
New & Re-Development Non-Residential Grid Cells (1985-2010)
Probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Distance (Meters)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
121
122
124
125
126
131


2010 AADT (Interpolated) for Access Classified Facility
New & Re-Development Residential Grid Cells (1985-2010)
Probability
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
AADT (Interpolated)
0
1
2
3
4
5
6
×10^4
111
112
113
114


2010 AADT (Interpolated) for Access Classified Facility
New & Re-Development Non-Residential Grid Cells (1985-2010)
Probability
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
AADT (Interpolated)
0
1
2
3
4
5
6
×10^4
121
122
124
125
126
131

QUESTIONS?