Using Machine Learning and Open Source Tools to Improve the Commodity Flow Survey

Economic Reimbursable Surveys Division

James Hinckley – Business Development Staff

Christian Moscardi - Data Scientist

CIVIC DIGITAL FELLOWSHIP

United States Census Bureau

Flora Wang
Stanford University
Symbolic Systems

Raanan Gurewitsch
University of Pittsburgh
Information Science

Background

Commodity Flow Survey (CFS)

- Sponsored by BTS and Census
- Conducted every 5 years
- Respondents select and report data on a sample of shipments made in each calendar quarter

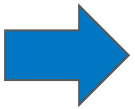
Problem: Significant number of establishments are out of scope (OOS)

Goal: Identify OOS establishments where there is no shipping activity

CIVIC DIGITAL FELLOWSHIP

Impact

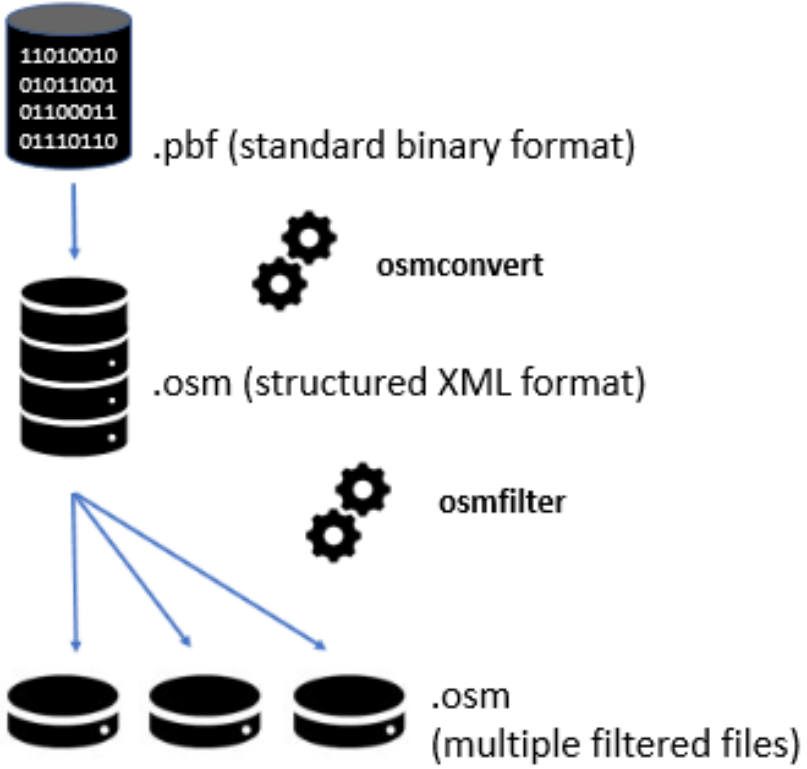
Higher rate of in-scope responses in future surveys



- More accurate shipping data
- Lower Census administrative costs
- Reduced burden on respondents

CIVIC DIGITAL FELLOWSHIP

OpenStreetMap Data Pipeline



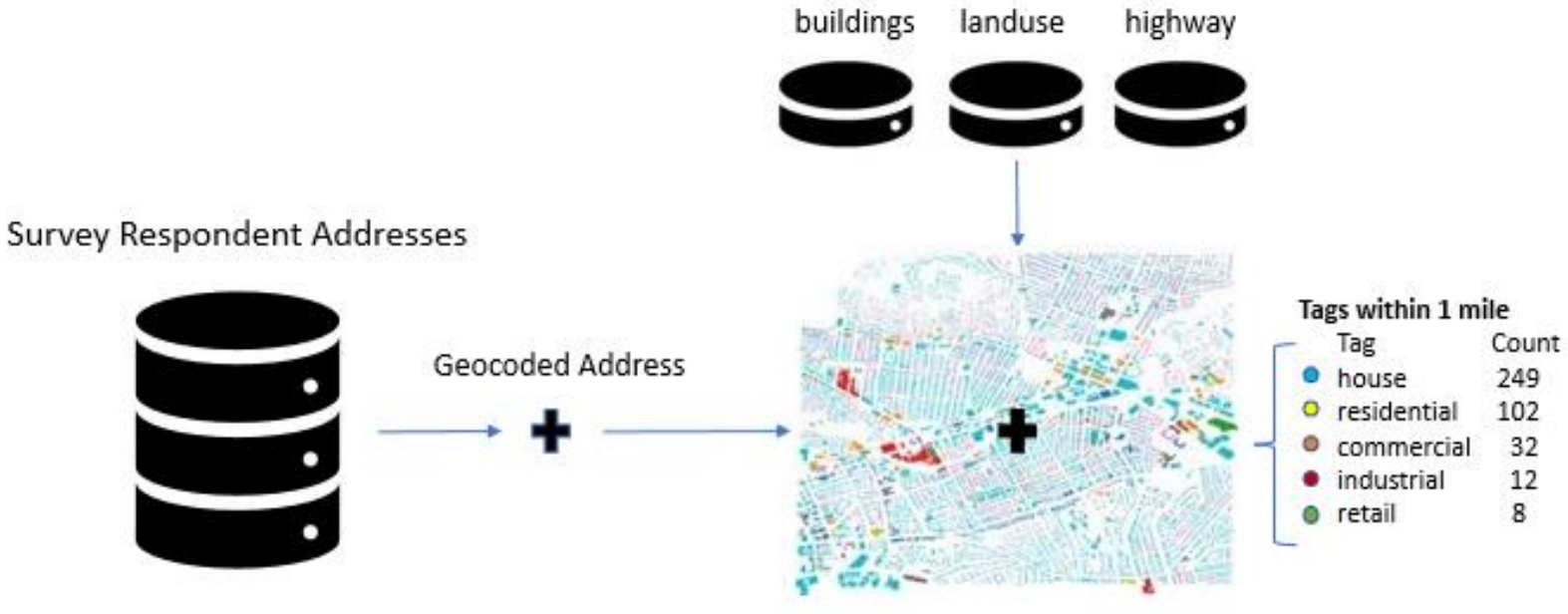
1. OpenStreetMap data for the United States is downloaded by region in compressed format

2. The data is converted to a larger but much more useful file format

3. The text files are then parsed through to find specific objects by their feature tags

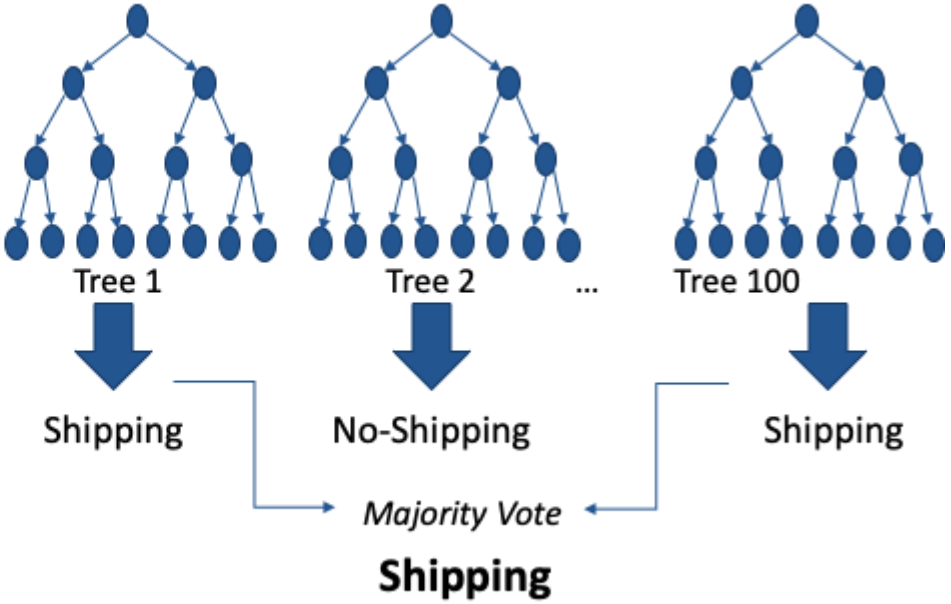
CIVIC DIGITAL FELLOWSHIP

OpenStreetMap Data Pipeline, contd.



Random Forest Classifier

Gather geographic data by location → **Ensemble Decision Tree-based ML model**

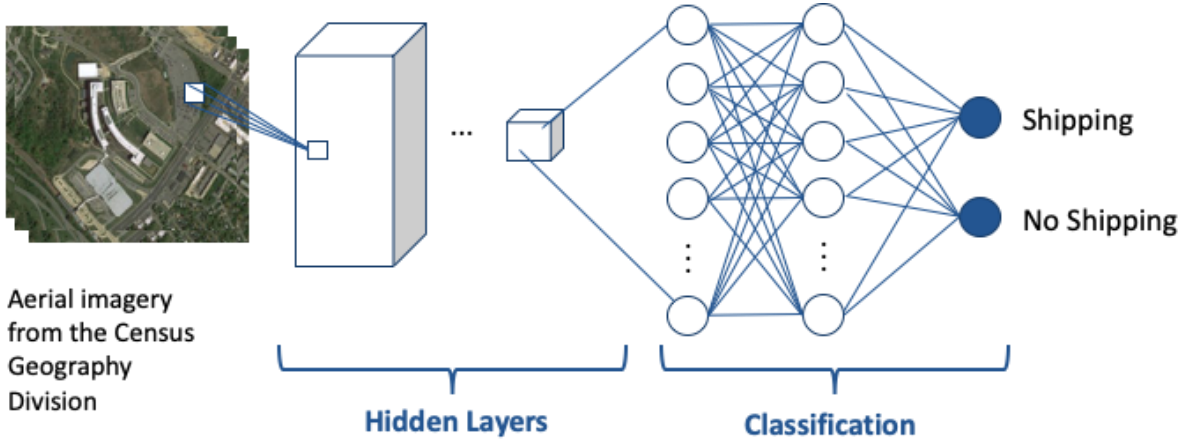


78% accuracy

CIVIC DIGITAL FELLOWSHIP

Convolutional Neural Network (CNN)

A deep learning approach to find patterns in images.



CNN assigns importance (learnable weights and biases) to aspects/objects in the image to be able to differentiate between shipping/no shipping classifications.

CIVIC DIGITAL FELLOWSHIP

CNN Results

85% validation accuracy

Model predictions on public data

(green = correct; red = incorrect)

CIVIC DIGITAL FELLOWSHIP

Future Impact

- 10% of respondents reported no shipping activity
- Model can lead to a **70% reduction** in error rate (to 3%)
- **$450,000** savings in respondent burden (using the OMB estimated 2.5 hours per questionnaire) for each CFS

CIVIC DIGITAL FELLOWSHIP

Acknowledgements

Special thanks to

- **James Hinckley**; Business Development Staff
- **Christian Moscardi**; Data Scientist
- **Keith Finlay, Carla Medalia, Barbara Wongus**; ERD
- **Ben Schultz, Berin Linfors**; CFS
- **Julie Parker, Mehdi Hashemipour**; BTS
- **Chris Kuang, Rachel Dodell**; Civic Digital Fellowship

CIVIC DIGITAL FELLOWSHIP