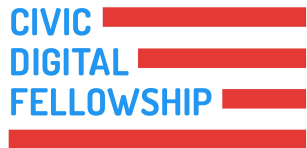


Using Machine Learning and Open Source Tools to Optimize the Commodity Flow Survey

Economic Reimbursable Surveys Division

James Hinckley - Branch Chief, Commodity Flow Survey

Christian Moscardi - Data Scientist



Flora Wang
Stanford University
Symbolic Systems

Raanan Gurewitsch
University of Pittsburgh
Information Science

Background

Commodity Flow Survey (CFS)

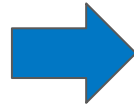
- Sponsored by BTS and Census
- Conducted every 5 years
- Respondents provide sampling of shipment from each quarter

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

Goal: Identify OOS establishments where there is no shipping activity

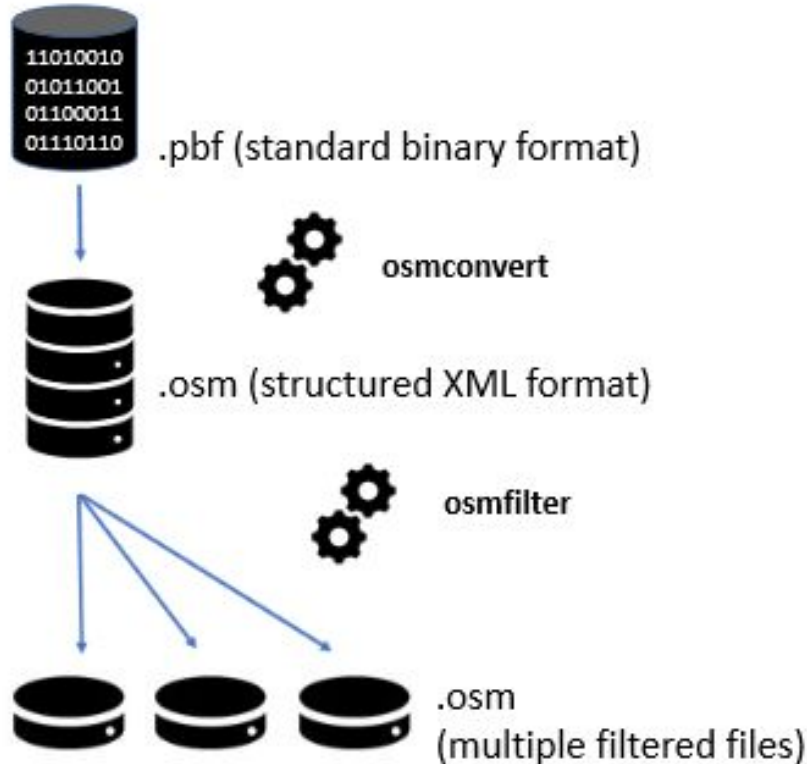
Impact

Higher rate of
in-scope
responses in
future surveys



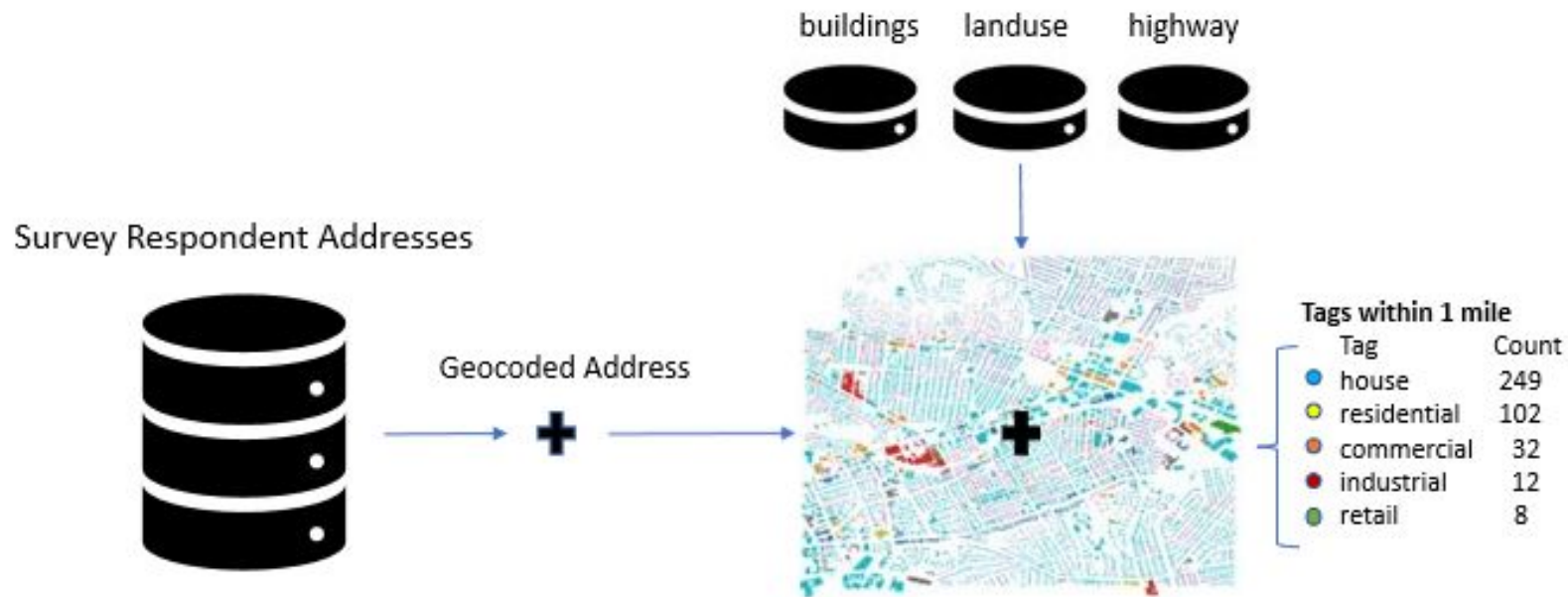
- More accurate shipping data
- Lower Census administrative costs
- Reduced burden on respondents

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

OpenStreetMap Data Pipeline, contd.

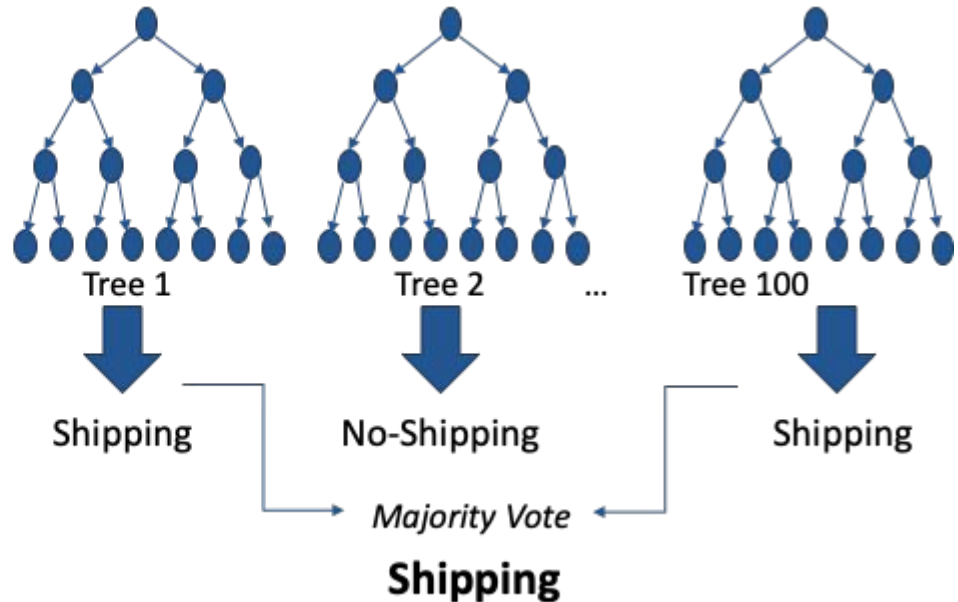


Random Forest Classifier

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

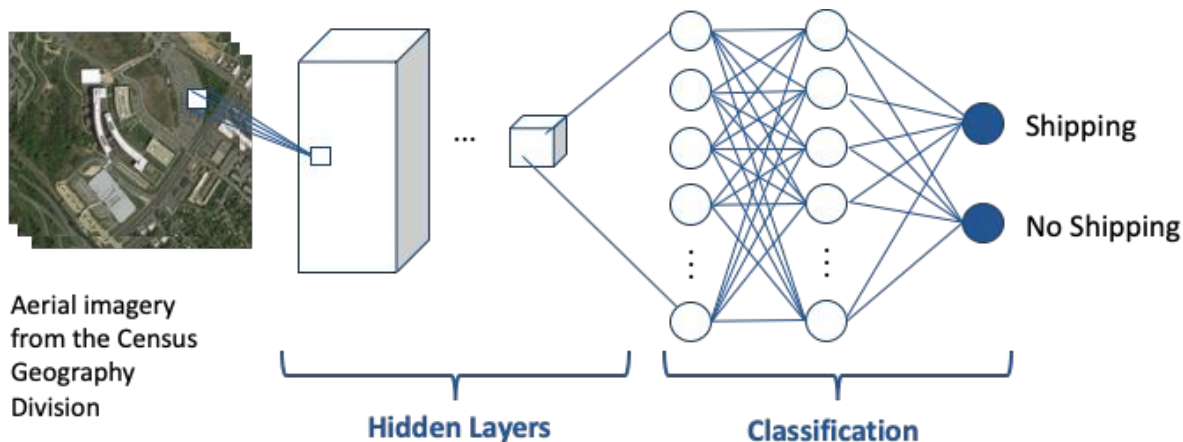


xx% accuracy



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.

CNN Results

xx% validation
accuracy

Model predictions on public data
(green = correct; red = incorrect)

[placeholder]

Future Impact

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

Acknowledgements

Special thanks to

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