Office of Data and Innovation

Making local population estimates more efficient to support resource allocation

Brittany Allen[1], Fennis Reed[2], Ian Rose[1], Jason Lally[1], Walter Schwarm[2], James Miller[2]

[1] California Office of Data and Innovation, 401 I Street, Ste 200, Sacramento CA 95814
[2] California Department of Finance, 915 L Street, Sacramento CA 95814

We helped the Department of Finance (Finance), Demographic Research Unit (DRU) make the process of creating small-area population estimates faster and more reliable to support critical state and local decision-making.

The opportunity

Small-area demographic estimates are critical to support decision-making across a range of applications. These estimates provide population and demographic insights at highly localized levels, allowing for accurate resource allocation, urban planning, and emergency response strategies. For example, local jurisdictions rely on these estimates to distribute funds effectively, forecast infrastructure needs, and plan services such as healthcare, education, and transportation. The Office of Data and Innovation (ODI) worked with the DRU to decrease repetitive data preparation by an analyst so they could focus on the estimations rather than processing the same source data for each analysis.

One of the key datasets for the small-area estimates is building footprints. Building footprints are outlines of physical structures derived from aerial imagery. Administrative units (like Census tracts) that cover large and irregular areas don't always have uniform distribution of population. Building footprints can provide indications of where people actually live. One excellent source for building footprints is the open source Microsoft's Building Footprints dataset. These data are generated from global imagery using computer vision algorithms, updated regularly, and used as a source for, among other things, Bing maps.

DRU had challenges they were seeking to solve in collaboration with ODI:

- The Microsoft Building Footprints dataset was hosted in an inefficient, non-cloud-friendly format, making it difficult to download and use in analyses.
- The dataset was frequently updated, but without helpful change tracking, making responding to these updates difficult.
- There was no way to preserve manual corrections and deduplication across each building dataset version.
- Combining footprint data with Census and parcel data was done using a single process for the entire state. This was both time-consuming and relatively crude: a good algorithm for an urban area was not necessarily appropriate for a rural area.

Overall, the process of footprint integration within an estimate could take up to 3 days.

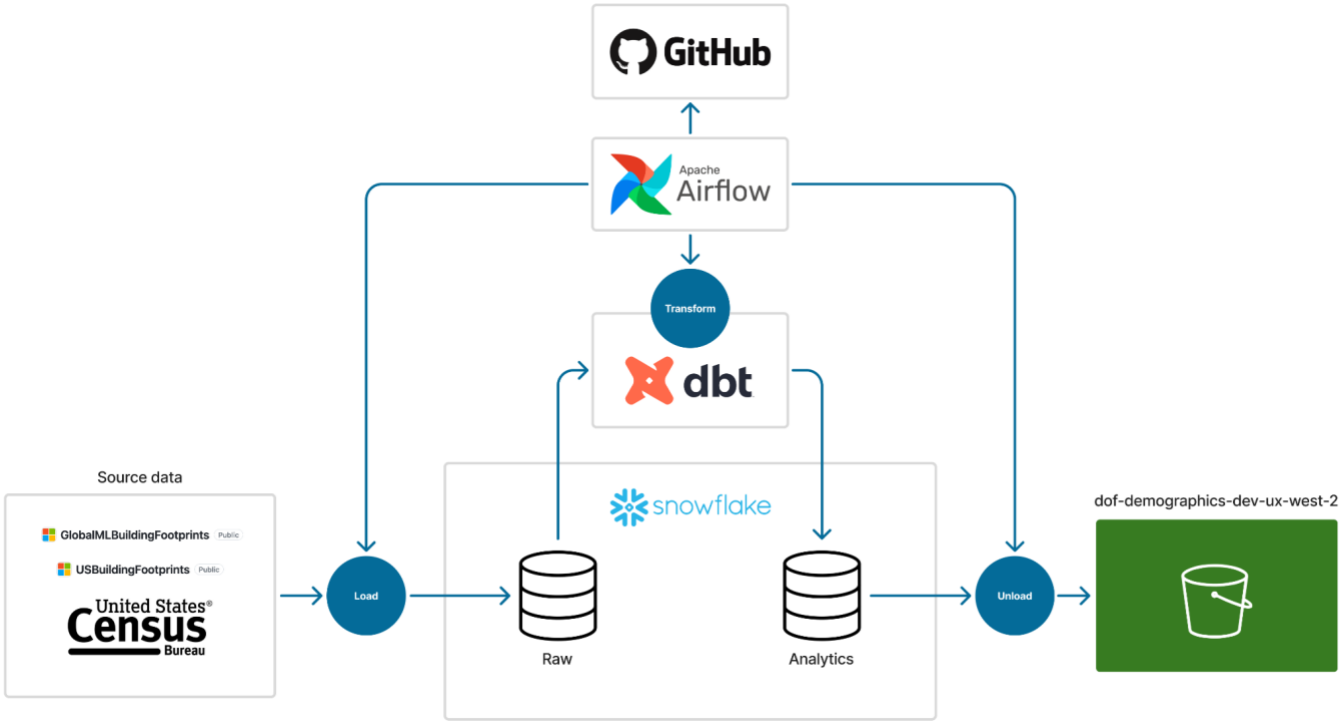
Our approach

ODI developed a derived dataset from the Microsoft footprints dataset. We joined it with US Census TIGER data to make it more useful for demographic and social science research. We also partitioned the data by county, which enables end users to access data for only the counties they need, saving time and computational resources.

Method

ODI developed a data pipeline that automates:

- Geospatial joins with Census TIGER/Line data (representing Census boundaries)
- Deduplication operations on footprints
- Footprint assignment to Census geometry when they intersect more than one.



ODI also developed 2 Python workflows with Finance, one for joining footprints with parcel data (or other geospatial datasets) and the other for removing incorrectly identified footprints. This is important because the algorithms that produce the building footprints from imagery can misidentify objects on the ground as buildings. Once an analyst identifies one, it will keep it out of the dataset even if the source data maintains that footprint, also saving repeat data cleaning over multiple analyses.

Distribution

The data are stored in a publicly accessible AWS S3 bucket which is low-cost, reliable, and fast. To allow for efficient storage, the data are delivered in both GeoParquet and zipped Shapefile formats. Zipped Shapefiles are universally supported by Geographic Information System (GIS) tools, and GeoParquet files are cloud-friendly, provide sensible data types, and allow users to read in only the columns they need. The data is also available via API access, which allows for programmatic retrieval. For an analyst working with source data covering the entire state, this enables much more efficient processing of data and speedier analysis using only the data required to complete a small area demographic estimate.

These data are freely downloadable at Building Footprints Dataset.

Impact

The collaboration between DRU and ODI successfully transformed the building footprint integration process, yielding the following key results:

Significant time reduction

Turned a days-long footprint integration process into a minutes-long one.

Automated updates with data retention

Developed a system for periodic automated updates in the cloud that retains revisions, such as the removal of incorrectly identified footprints, across every new dataset version.

Enhanced accessibility and efficiency

Data are now publicly available via a low-cost, reliable, and fast AWS S3 bucket. End-users can access it through multiple methods, including API access and direct downloads, and can download data partitioned by county in cloud-friendly formats like GeoParquet, saving significant time and computational resources.

Improved data integration and customization

Created a specialized toolbox that automates the calculation of Census TIGER joins and allows for customizable parameters during integration with other geospatial data (like parcels), which accounts for the variability of housing types across the state.

Code

GitHub repository: data transformation

GitHub repository: data loading

Authors

Brittany Allen
Senior Analytics Engineer
brittany.allen@innovation.ca.gov
California Office of Data and Innovation | 401 I Street, Ste 200, Sacramento, CA 95814
Roles: Software, methodology, data analysis, user research, writing – original draft
https://orcid.org/0009-0003-4005-8848

Fennis Reed
Research Data Specialist
fennis.reed@dof.ca.gov
California Depart of Finance, Demographic Research Unit | 915 L Street, Sacramento CA 95814
Roles: Conceptualization, Methodology, data analysis, writing – original draft

Ian Rose
Principal Data Engineer
ian.rose@innovation.ca.gov
California Office of Data and Innovation | 401 I Street, Ste 200, Sacramento, CA 95814
Roles: Software, methodology, data analysis, user research, writing – original draft

Jason Lally
Chief Data Officer
jason.lally@innovation.ca.gov
California Office of Data and Innovation | 401 I Street, Ste 200, Sacramento, CA 95814
Roles: Conceptualization, Project administration, resource, writing – original draft

Walter Schwarm
Chief, Demographer
walter.schwarm@dof.ca.gov
California Depart of Finance, Demographic Research Unit | 915 L Street, Sacramento CA 95814
Roles: Conceptualization, resources, supervision

James Miller
Assistant Chief
james.miller@dof.ca.gov
California Depart of Finance, Demographic Research Unit | 915 L Street, Sacramento CA 95814
Roles: Conceptualization, resources, supervision

Roles use the CRediT taxonomy