ETHER DATA · TECHNICAL GUIDE

One description of place carries across models and markets

Ether Data's place embeddings give every small area of the United States outside Alaska 128 numbers that describe what kind of place it is. This guide shows how to join them to your records and how we evaluated them. In states held out of the fit, they keep skill that census features lose.

ETHER DATA
NOVEMBER 2026 · A GUIDE AND AN EVALUATION ON PUBLIC TARGETS, NOT A CLIENT RESULT
ETHERDATA.AI

WHERE THIS STARTED

Ether Data built the place embeddings so that one consistent description of place can sit beside the features a team already uses. The same 128 numbers describe every small area of the country outside Alaska, so a model can carry its sense of place from one market to the next. We designed and trained the model behind them in house.

This guide covers what the embeddings contain, how to join them to records with a location, how we evaluated them and where they have added signal so far.

01 · WHAT THEY ARE

Each small area gets 128 numbers that describe what kind of place it is

Each cell of the H3 grid at resolution 8, about 0.74 km² on average, carries one vector of 128 numbers.[1]

- **Our own model, self-supervised.** Ether Data designed and trained it in house, from census attributes and geographic context. No device IDs.[2]
- **Standardized.** Each of the 128 dimensions has zero mean and unit variance across cells.
- **Read as geometry, not as columns.** The dimensions carry no labels. The information lives in distances and directions: places with similar profiles sit close together, and a direction in the space can line up with a property you care about.

Coverage is 49 states and DC (Alaska is not covered), about 7.6 million cells, one row per H3 cell. The evaluation sample covers ten Northeast states (CT, DE, MD, MA, NH, NJ, NY, PA, RI, VT): 485,701 cells.

COLUMN	TYPE	DESCRIPTION
`h3`	STRING	H3 resolution 8 cell index. The join key.
`state`	STRING	Two-digit state FIPS code.
`H000 … H127`	FLOAT	The 128-number place vector, standardized per dimension.

You can receive the embeddings two ways: Parquet files that load into any platform, or a shared BigQuery table you query in place. The BigQuery table also carries the vector as a single array column for vector and machine learning operations, and keeps every cell on one national scale, so standardize on your side if a model needs it. You need no SDK or API.

02 · THE JOIN

A shared H3 cell attaches the embeddings to any record with a location

1. Geocode each record to latitude and longitude, if it is not already.
2. Convert each point to its H3 resolution 8 cell.
3. Join on `h3`.

```
import h3
import pandas as pd

emb = pd.read_parquet("ether_place_embeddings_NE10.parquet")

df["h3"] = [h3.latlng_to_cell(lat, lng, 8) for lat, lng in zip(df.lat, df.lng)]
df = df.merge(emb, on="h3", how="left")
```

Coarser units are averages of the cells inside them: child cells for a parent H3 cell, or a weighted average of the cells inside a ZIP code, county or territory polygon, weighted by population or record counts.

03 · HOW WE EVALUATED

We score the embeddings only on areas held out of training

Neighbouring cells resemble each other, so we hold out whole areas rather than random rows. We group cells by their H3 resolution 5 parent for cross-validation, and for transfer we leave out whole states. We score census features and the embeddings on the same folds, alone and together, and we validate against public targets the embeddings were not built from: night-time lights, terrain elevation and cell-tower locations.[3,4,5] The blocked probe we use:

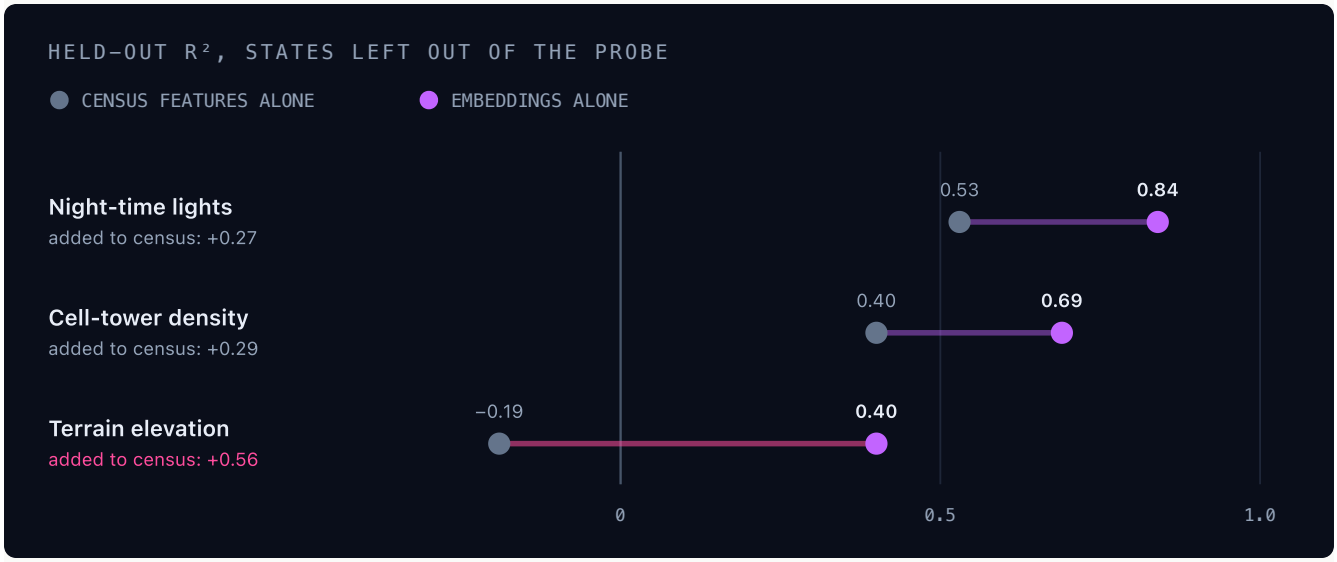
```
import numpy as np
from sklearn.linear_model import RidgeCV
from sklearn.model_selection import GroupKFold, cross_val_score

cols = [f"H{i:03d}" for i in range(128)]
groups = df["h3"].map(lambda c: h3.cell_to_parent(c, 5))
model = RidgeCV(alphas=np.logspace(-2, 3, 20))
scores = cross_val_score(model, df[cols], df["target"], groups=groups,
                         cv=GroupKFold(n_splits=5), scoring="r2")
```

04 · THE RESULTS

The embeddings keep their skill in states where census features fail

The embeddings add the most as spatial context that transfers. In our evaluations they add the most on targets that vary smoothly across space, and when the probe predicts in states it was not fitted on.



The embeddings score well above census features in held-out states on all three public targets, most of all on elevation. Linear probe fitted with each state left out in turn and scored on that state's cells; pooled out-of-fold R^2 on 120,000 sampled cells, current release. The dots show each feature set alone: census features and the embeddings. The figure under each target is the gain from adding the embeddings to census features, not the gap between the dots. Jurisdictions held out: night-time lights 13, cell-tower density 15, terrain elevation 50 (49 states and DC).[6]

Census features alone score below zero on elevation in held-out states, worse than predicting the average, while the embeddings keep most of their skill. That gap is the transferable context.

WHAT THESE RESULTS DO NOT SHOW

Linear probes are a lower bound on what the vectors carry, not a ceiling. The targets are public measurements, so these results do not establish performance on any team's own business outcomes. The embeddings themselves were trained on all states, without these targets; only the probe holds states out. For targets driven mainly by a cell's own density, such as counts of points of interest, raw census counts do about as well as the embeddings. The comparison that matters is lift over existing features, not lift over nothing.

NOTES

Sources, with the strength of each

Each note names the publisher and the kind of evidence: public statistics, open-source reference, crowdsourced data or Ether Data's own evaluation.

1. Uber Technologies, H3 hexagonal hierarchical geospatial indexing system. Open-source reference; resolution 8 cells average about 0.74 km². h3geo.org
2. US Census Bureau, American Community Survey. Federal statistics, one of the inputs to the embeddings. census.gov/programs-surveys/acs
3. Earth Observation Group, VIIRS night-time lights. Public satellite-derived data, used as an evaluation target and not as an input. eogdata.mines.edu/products/vnl
4. Public terrain elevation data, used as an evaluation target and not as an input.
5. OpenCellID, open database of cell-tower locations. Crowdsourced, used as an evaluation target and not as an input. opencellid.org
6. Ether Data, held-out linear-probe evaluation of the current embedding release, November 2026. Ether Data's own evaluation: each state left out in turn, pooled out-of-fold R^2 on 120,000 sampled cells.

About Ether Data. Ether Data describes what kind of place each small area is and how it changes by hour across the year. It uses no device IDs. This guide reports evaluations on public targets and no client results. etherdata.ai