PUFFIN: PROBABILISTIC LEARNING OF SPATIAL DETAIL FROM COARSE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-resolution socioeconomic variables are important for applications such as urban planning, public health, disaster response, and resource allocation. In practice, however, these variables are often observed only at a coarse spatial resolution. We introduce PUFFIN, a probabilistic framework for statistical disaggregation that raises the resolution of coarse totals using high-resolution satellite embeddings as covariates. Instead of predicting a single value for each fine-resolution subregion, PUFFIN learns a probability distribution and is trained through an aggregation-aware likelihood. At inference, PUFFIN conditions these predictions on the observed regional total and splits it among the subregions. The resulting fine-scale estimates are consistent with the observed aggregate and come with calibrated uncertainty, without requiring fine-resolution labels for training. We evaluate PUFFIN on German and US census, employment, and election data across population, jobs, and other count variables, and study when statistical disaggregation succeeds or fails across regions, countries, and targets. Code is available at `anonymous.4open.science/r/Puffin`.

1 INTRODUCTION

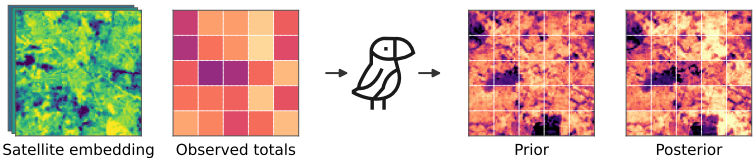

Figure 1: Overview: From a satellite embedding and one observed total per region, PUFFIN learns a prior distribution over every subregion and conditions it on the totals, which gives a posterior that splits each total exactly among its subregions (means shown).

Where people live, where they work, and where they are exposed to risk are questions that urban planning, public health, and disaster response must answer street by street. Censuses and registers publish such statistics only as one total per reporting unit, a census tract or a county, to protect the people behind the numbers and because surveying at a finer scale is expensive. Satellite imagery, by contrast, covers the whole world at 10 to 100 m and carries a strong signal about socioeconomic outcomes (Jean et al., 2016; Yeh et al., 2020; Rolf et al., 2021), and recent foundation models such as AlphaEarth turn every pixel into a short vector, an embedding (Brown et al., 2025; Cong et al., 2022). Recovering the fine-scale map behind a coarse total from such imagery is called spatial disaggregation (Metzger et al., 2022), and it is the problem we study.

We think of disaggregation as splitting a fixed amount. Each region reports one total, which we divide among the small subregions inside it so that every subregion gets a non-negative value and the pieces add back up to the reported total. This mass-preserving view goes back to Tobler's interpolation (Tobler, 1979), and we make the split learned and probabilistic. The only supervision is that single total, which makes this the count version of learning from label proportions (Quadrianto et al., 2008).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The gridded population products in widest use today, among them WorldPop (Stevens et al., 2015), GHS-POP (Schiavina et al., 2023), and LandScan (Lebakula et al., 2025), are still built by dasymetric mapping, which spreads each regional total over auxiliary layers assumed to indicate where people are, such as night lights, roads, or building footprints, with weights set by hand or fitted to the totals. Those layers are laborious to assemble and differ from country to country, and the published product is a single number per cell.

We introduce PUFFIN, a probabilistic framework for spatial disaggregation that removes all three limitations. It replaces the curated layer stack with a single satellite embedding, learns the split instead of fixing it by a rule, and returns a predictive distribution per subregion rather than a bare number, while still reproducing each published total exactly. Training needs no fine-resolution labels, which is what makes it usable where such labels do not exist.

PUFFIN treats disaggregation as probabilistic constraint satisfaction (Figure 1). A small network predicts a prior distribution over each subregion's count from its embedding, trained only through the aggregation constraint, and conditioning that prior on an observed total gives a posterior that satisfies the constraint exactly. Its only inputs are a regional total and a globally available satellite embedding, so it applies wherever totals are published, including the many countries with no fine-resolution census grid.

Contributions. PUFFIN frames spatial disaggregation as learning from aggregates under an exact aggregation constraint, and from this we make three contributions.

- **Disaggregation as Probabilistic Constraint Satisfaction.** We learn a per-subregion predictive distribution from regional totals alone, with the aggregation constraint as the only supervision, and cast inference as conditioning that distribution on the observed totals. No fine-resolution label is needed to train the model.
- **Mass Conservation With Calibrated Uncertainty.** The resulting posterior is closed-form and consistent with every aggregate. The estimate it yields is non-negative, reproduces each observed total exactly, and gives each subregion a share of the correction set by its uncertainty. A two-stage calibration, in which only a single global factor uses fine labels, turns the posterior variances into intervals with validated marginal coverage.
- **A Held-Out Region Evaluation.** On regions held out by spatial cross-validation and against three baselines, we evaluate the approach at many resolutions, on a city (Berlin) and a whole country (Germany), on a second country with irregular units (the United States), and on two variables (population and jobs). We also identify where a linear baseline is as good and where the approach does not help (precinct turnout and rare events).

2 NOTATION AND PROBLEM SETTING

The study area consists of N coarse regions, each divided into M subregions, and (i, j) denotes subregion j of region i. Regions are the level at which a total is observed (e.g., census tracts or counties), and subregions are the level we want to recover (grid cells in Germany, census blocks in the United States). For notational simplicity we assume every region has the same number M of subregions, all of the same size. Neither assumption is needed in practice, and both are dropped in Section 3.2.

The unobserved fine outcomes form a matrix $Y \in \mathbb{R}_{\geq 0}^{N \times M}$ whose entry $y_{i,j}$ is the outcome of subregion j of region i, for instance the number of residents of that grid cell or census block. The observed coarse outcomes form a vector $\widehat{Y} \in \mathbb{R}_{\geq 0}^{N}$ whose entry $\widehat{Y}_i$ is the single observable total of region i, for instance the number of residents of that census tract. Each subregion carries contextual covariates, collected in a tensor $C \in \mathbb{R}^{N \times M \times d}$ whose entries $\mathbf{c}_{i,j} \in \mathbb{R}^d$ are d-dimensional feature embeddings of the subregion (e.g., 64-dimensional AlphaEarth embeddings); nothing below depends on this choice. Fine and coarse outcomes are linked by sum aggregation,

$$\widehat{Y}_i = \sum_{j=1}^{M} y_{i,j} \quad \text{for } i = 1, \ldots, N, \tag{1}$$

so a single number per region is shared by all of its subregions, while Y itself is never observed. This is a weakly supervised, learning-from-aggregates setting.

Objective. Given the coarse totals $\widehat{Y}$ and the covariates C, our goal is to infer Y, consistently with the aggregates of Equation 1 and with uncertainty over the inferred values.

3 Our Method: Puffin

Puffin works in two stages. It learns a prior over every subregion's outcome from the covariates, fitted only through the aggregation constraint, and then conditions that prior on the observed total, which gives a closed-form posterior that conserves the total exactly.

3.1 Overall Idea

For each subregion j of region i, a neural network takes the context $\mathbf{c}_{i,j}$ as input and predicts the mean and variance $(\mu_{i,j}, \sigma^2_{i,j}) = f_{\boldsymbol{\theta}}(\mathbf{c}_{i,j})$ of a Gaussian distribution over the target variable.

Prior. We interpret the joint distribution induced by these predictions as a *prior* over the fine-resolution outcomes. In particular,

$$\tilde{y}_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j}), \qquad j = 1, \ldots, M, \tag{2}$$

with the subregions assumed to be independent. Here, $\tilde{y}_{i,j}$ denotes the model's random variable for the unobserved ground-truth outcome $y_{i,j}$. This prior represents what the model expects for each subregion based only on its covariates, before taking the observed regional aggregate into account.

Posterior. For each region, we observe only the aggregate $\widehat{Y}_i = \sum_{j=1}^{M} y_{i,j}$. We obtain the *posterior* distribution over the model's subregion outcomes by conditioning the prior on this observed aggregate:

$$p_{\boldsymbol{\theta}}\left(\tilde{y}_{i,1}, \ldots, \tilde{y}_{i,M} \;\middle|\; \sum_{j=1}^{M} \tilde{y}_{i,j} = \widehat{Y}_i\right).$$

For the Gaussian prior, the posterior is again Gaussian (restricted to the aggregation constraint), and its maximum a posteriori (MAP) estimate is equal to its mean and available in closed form (Equation 5 in Section 3.2). Intuitively, if the observed aggregate is larger than the sum of the predicted means, the individual predictions must be corrected upward, and if it is smaller, they must be corrected downward. The correction is distributed across subregions according to their uncertainty, such that subregions with larger variance are adjusted more.

Training. Because independent Gaussian variables add to another Gaussian, the model's prior over the regional aggregate is again Gaussian, with mean $\mu_i = \sum_j \mu_{i,j}$ and variance $\sigma_i^2 = \sum_j \sigma^2_{i,j}$. We train the neural network through the prior only, never through the posterior, by maximizing the likelihood that the prior assigns to the observed regional aggregates, or equivalently by minimizing the negative log-likelihood (NLL)

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \left[\frac{\left(\widehat{Y}_i - \sum_{j=1}^{M} \mu_{i,j}\right)^2}{\sum_{j=1}^{M} \sigma^2_{i,j}} + \log\left(\sum_{j=1}^{M} \sigma^2_{i,j}\right) \right], \tag{3}$$

over the set $\mathcal{T}$ of training regions, up to an additive constant and an overall positive factor. It reduces to the squared loss used by CLAM (Großmann et al., 2026), a related disaggregation model, when all variances are equal. The prior does not take the total as input, so its mean may well miss the observed $\widehat{Y}_i$. That residual is exactly what the loss penalizes, and reproducing a total exactly is the task of the posterior, not of the prior. A different additive variable just swaps in its own aggregate likelihood.

3.2 IMPLEMENTATION

Four things separate the idea above from a working method. Subregions differ in area, the split has to stay non-negative, the posterior variances have to be turned into trustworthy intervals, and the parametrizable function $f_{\boldsymbol{\theta}}$ needs a concrete form.

Uneven Subregion Sizes. Real subregions are not equally large, and a large block should be able to hold more than a small one that looks the same. We assume an area matrix $A \in \mathbb{R}_{>0}^{N\times M}$, where $a_{i,j}$ contains the area of subregion (i, j), measured in units of 100 m cells. The network then predicts a mean and a variance per unit area, and the area turns both into a count,

$$(u_{i,j},\, v_{i,j}) = f_{\boldsymbol{\theta}}(\mathbf{c}_{i,j}), \qquad \mu_{i,j} = a_{i,j}\, u_{i,j}, \qquad \sigma_{i,j}^2 = a_{i,j}\big(\mathrm{softplus}(v_{i,j}) + \varepsilon\big), \tag{4}$$

with $\varepsilon = 10^{-6}$ keeping the variance positive. Everything above is unchanged, because prior, loss, and posterior only ever use $\mu_{i,j}$ and $\sigma_{i,j}^2$. On a uniform grid the area is constant and cancels. On US census blocks, whose areas span about six orders of magnitude, it is what makes the split work. Scaling the variance with the area lets a large block absorb a correspondingly larger share of the region's correction, and without it the split degrades sharply. A varying number of subregions per region needs no change at all, beyond reading M as M_i.

Closed-Form Split and Non-Negative Estimates. Conditioning a Gaussian on the value of a sum has a closed form, so the split needs no optimization at inference,

$$\tilde{y}_{i,j}^{\mathrm{MAP}} = \mu_{i,j} + \frac{\sigma_{i,j}^2}{\sum_{k=1}^{M} \sigma_{i,k}^2} \left(\widehat{Y}_i - \sum_{k=1}^{M} \mu_{i,k} \right). \tag{5}$$

Counts cannot be negative, but this estimate can be, wherever the prior over-predicts a region. We clip negative $\tilde{y}_{i,j}^{\mathrm{MAP}}$ to zero and rescale the region so that the estimates still sum to $\widehat{Y}_i$, falling back to an even split if all of a region's estimates vanish. We write $\tilde{\mu}_{i,j}$ for the resulting value and call it the PUFFIN *estimate*. It is the posterior mean after clipping, it is our estimate of $y_{i,j}$, and it is what every map and table below reports, while the posterior variances give the intervals.

Calibrated Intervals. Conditioning on the total also gives each subregion a posterior variance, $\tilde{\sigma}_{i,j}^2 = \sigma_{i,j}^2 - \sigma_{i,j}^4/\sigma_i^2$ with $\sigma_i^2 = \sum_k \sigma_{i,k}^2$. They come from a prior fitted to totals alone and are smaller than the errors the estimate actually makes, so we rescale them in two steps. Within each 10 km tile t we first use the coarse residuals, which need no fine labels, to set $\gamma_t^2 = \sum_{i\in t}(\widehat{Y}_i - \mu_i)^2 / \sum_{i\in t} \sigma_i^2$ and $\bar{\sigma}_{i,j}^2 = \gamma_{t(i)}^2\, \tilde{\sigma}_{i,j}^2$. On a set of validation regions we then choose a single factor κ as the 90% quantile of the standardized errors $|y_{i,j} - \tilde{\mu}_{i,j}|/\bar{\sigma}_{i,j}$, which gives the interval $\tilde{\mu}_{i,j} \pm \kappa\, \bar{\sigma}_{i,j}$. Only κ uses fine labels, and only on validation regions, which are always disjoint from the regions we evaluate on (Section 4).

Network and Optimization. The network $f_{\boldsymbol{\theta}}$ is a single hidden layer of width 32 with Gaussian error linear unit (GELU) activations and about 2,000 parameters, and a larger network overfits the totals. We minimize Equation 3 as written, standardizing the totals by the average training density so the loss does not depend on the scale of the variable, and batching whole regions at country scale, since a region's total needs all of its subregions in the batch. Appendix A gives the remaining settings.

4 EXPERIMENTS

We run three experiments on real census and election data, where the fine-scale values are counted rather than modeled, so unlike gridded population products they can validate a disaggregation. We first give the protocol and the baselines shared by all three.

Prior Versus Posterior. The prior (Equation 2) is what one would use where no total is available, and it enters the experiments only through the training loss. Since every region's total is observed here, every reported number uses the posterior (Equation 5), through the PUFFIN estimate for point accuracy and the posterior variances for the intervals. The exceptions are the PUFFIN (prior) rows, which score the prior alone against its own regional mean.

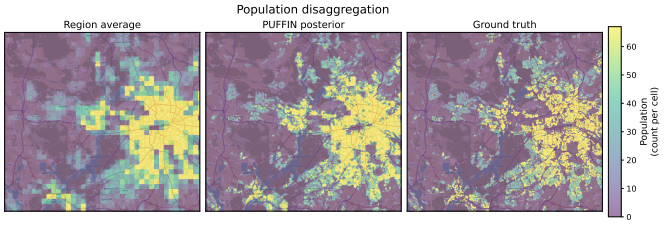

Figure 2: **Qualitative disaggregation (Berlin).** Region-average baseline (left), PUFFIN estimate (middle), and ground truth (right), for population. PUFFIN recovers within-region structure that the flat baseline misses, from the embeddings and totals alone.

Metric. We compare the PUFFIN estimate $\tilde{\mu}_{i,j}$ with held-out true values using within-region R^2, which removes each region's mean $\bar{y}_i = \widehat{Y}_i/M$ from both truth and estimate,

$$R^2_{\text{within}} = 1 - \frac{\sum_{i,j} \left(\dot{y}_{i,j} - \dot{\mu}_{i,j}\right)^2}{\sum_{i,j} \left(\dot{y}_{i,j} - \bar{\dot{y}}\right)^2}, \qquad \dot{y}_{i,j} = y_{i,j} - \bar{y}_i, \quad \dot{\mu}_{i,j} = \tilde{\mu}_{i,j} - \bar{y}_i, \tag{6}$$

so the region-average baseline scores exactly zero and the metric measures only how well each total is split among its subregions. For comparison with other work, Appendix B also reports subregion-R^2, the usual R^2 against the global mean. It mostly measures the totals rather than the split, since even the region average reaches 0.83 on Berlin (Table 5), the observed totals alone already separating dense regions from empty ones.

Protocol. All results use 5-fold spatial cross-validation. Regions are grouped into 5 km spatial groups that are assigned to folds as a whole, so test regions sit away from the training ones. Since training uses only totals, one would in practice fit on all regions, so this is a stricter test than deployment requires. For each test fold the preceding fold is used for validation (checkpoint selection, the ridge strength of the regression baseline below, and κ) and the remaining three for training, so every subregion is scored once, out of fold. We report within-region R^2 (Equation 6) on all subregions with the coverage of the 90% intervals, and Appendix B adds subregion-R^2 and the inhabited-only comparisons. PUFFIN results are 5-seed means with fixed folds. Irregular units use the area term of Equation 4.

Baselines. We compare PUFFIN with three baselines, all using the same regions, subregions, and folds, never using fine labels, and reproducing every observed total exactly, so the comparison isolates how each method splits a total. The **region average** gives every subregion the same share $\widehat{Y}_i/M$ and defines the zero point of within-region R^2. **Bicubic interpolation** reads the regional densities as samples of a smooth surface and evaluates a cubic interpolant of it at every subregion, so it tests whether spatial smoothness alone explains the fine structure, using neighboring totals but no covariates. **Goodman-style ecological regression** (Goodman, 1953; 1959) infers individual-level relationships from aggregate data by regressing region-level rates on region-level covariates, which we apply to the same embeddings as PUFFIN, so it tests whether a nonlinear probabilistic model is needed at all. Both fitted densities are multiplied by the subregion area, clipped at zero, and rescaled to the observed total. The baselines are deterministic and therefore have no seeds. Appendix A gives their exact form.

4.1 EXPERIMENT 1: GERMAN CENSUS, CITY TO COUNTRY

Data. The German register-based Zensus 2022 counts residents on a 100 m grid (Statistisches Bundesamt, 2024), so the fine values are measured rather than modeled. We use every 100 m land cell in a 100×100 km box around Berlin (1.0M cells, referred to as Berlin below) and in Germany (35.7M), and form nested square grids. The covariates are AlphaEarth embeddings averaged over each subregion. The regions are the coarse squares whose totals we observe, with sizes from 200 m to 10 km, and the subregions are the finer squares inside them, 2 to 100 per side, which gives 18 configurations. Appendix A lists the cell counts and the configurations.

Data	Region size R	Fine cells per side (R/s) 2	5	10	20	50	100
Berlin	200 m	.14	–	–	–	–	–
	1 km	.43	.38	.28	–	–	–
	2 km	.50	.49	.42	.30	–	–
	4 km	.54	.51	.47	.39	–	–
	10 km	.70	.59	.52	.46	.38	.28
Germany	200 m	.18	–	–	–	–	–
	1 km	.59	.48	.35	–	–	–
	2 km	.70	.61	.51	.37	–	–
	4 km	.77	.69	.61	.50	–	–
	10 km	.84	.76	.70	.61	.48	.36

Mean over the 18 configurations	Reg.-avg.	Bicubic	Goodman	PUFFIN (prior)	PUFFIN
Berlin	.000	.142	**.448**	.346	.432
Germany	.000	.100	.544	.485	**.562**

Table 1: Within-region R^2 on the German census, all land cells. **Top**: PUFFIN per configuration, a row being the region size R and a column the number of subregions per side, with blanks where subregions would be smaller than 100 m. **Bottom**: the mean over the 18 configurations per method, best in bold. The region average scores 0 by construction, and PUFFIN (prior) is the prior before conditioning, which the posterior improves on everywhere. Tables 3 and 4 give every method per configuration. PUFFIN: 5-seed mean (std ≤ 0.03).

Within a City (Berlin). Berlin tests the core question in the simplest setting, namely whether, within one urban region, the embeddings can tell which parts of a region hold its people. The uniform grid removes the area term, and varying the region and subregion sizes shows how the answer depends on scale. PUFFIN answers yes in every one of the 18 configurations (Table 1, top), where the region average scores zero by construction, and it does better the larger the subregion, from 0.14 at 200 m regions with 100 m subregions to 0.70 at 10 km regions with 5 km subregions. Bicubic interpolation reaches only 0.142 on average, so the totals of neighboring regions alone do not reveal the fine structure, pointing to the embeddings as the source of the gain. Goodman's regression keeps up on a single city and is slightly ahead on average (0.448 versus 0.432), winning 12 of the 18 configurations, but its lead comes from the largest regions and the coarsest subregions. The 10 km row of the grid holds six configurations, one per subregion size, and Berlin has only 73 regions of that size to train on. Goodman's ridge-regularized linear model wins all six of them, and the fine-scale error of PUFFIN starts to rise after about 100 of the 2,000 epochs (Figure 4), which is what overfitting so few totals looks like. Germany has 2,317 regions of the same size, and there PUFFIN wins all six. Over the remaining twelve configurations the two are on par in Berlin, 0.004 apart on average. Figure 2 shows the disaggregation on a map.

Scaling to a Country (Germany). Next, we train on all of Germany, which provides about 36× more data than Berlin. With this much data, PUFFIN reaches within-region R^2 between 0.18 and 0.84 (Table 1), higher than on Berlin in every configuration, and it now beats Goodman's regression in 16 of 18 configurations (mean 0.562 versus 0.544). The margins are small, at most 0.04, but consistent. The prior alone stays below Goodman (0.485 versus 0.544), so what puts PUFFIN ahead is the variance-weighted split rather than the nonlinearity of the prior. Bicubic interpolation again stays far below, at 0.100 on average. The per-configuration tables for both datasets, the inhabited-only cells, and subregion-R^2 are in Appendix B.

4.2 EXPERIMENT 2: US CENSUS, POPULATION AND JOBS

Data. The US 2020 census reports the population of 8.0M census blocks (U.S. Census Bureau, 2021), irregular polygons nested in 83,774 census tracts (50 states and DC), and the Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) report the number of jobs located in each block (U.S. Census Bureau, 2024). The regions are census tracts, whose totals we observe, and the subregions are the census blocks inside them. The covariates are

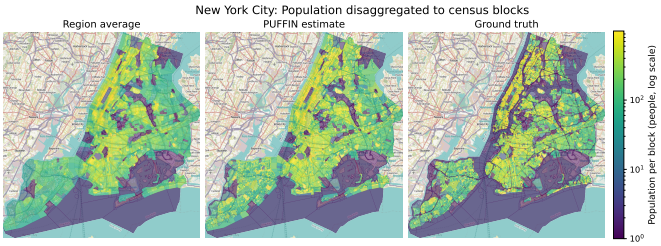

Figure 3: **US population in New York City**, with census tracts as regions and census blocks as subregions. Region average (left), the PUFFIN estimate (middle), and ground truth (right). The region average gives every block of a tract the same count, the tract total divided by its number of blocks, while PUFFIN recovers the block-level structure.

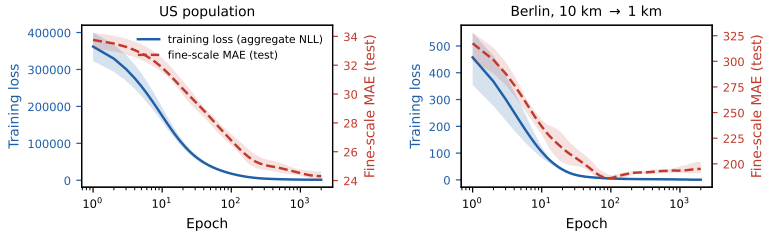

Figure 4: **Training on totals improves the fine-scale split.** Training loss (blue, left axis) and the mean absolute error (MAE) of the PUFFIN estimate on held-out test subregions (red, right axis) over training, as mean and range over 5 seeds. No subregion value ever enters training, yet the fine-scale error falls with the aggregate loss. On Berlin at 10 km (right), where only 73 regions are available for training, it starts to rise again after about 100 epochs, while the validation totals keep improving until epoch 485, so selecting the checkpoint on them does not catch this.

AlphaEarth embeddings averaged over each block polygon, and the area term uses each block's area, which spans about six orders of magnitude.

A Second Country and a Second Variable (US). This experiment changes two things at once. The fine units are irregular blocks instead of grid cells, and the variables include jobs, which concentrate in a few small commercial blocks instead of following where people live. PUFFIN works here too (Table 2), with population at $+0.377$, inside the range seen in Germany, and jobs at $+0.209$. Here bicubic interpolation and Goodman's regression both fail. On population they reach -0.41 and -1.04, worse than the region average, and on jobs only $+0.06$ and $+0.05$. The likely reason is that a smooth or linear density cannot fall to near zero on large empty blocks, so multiplying it by area pushes each tract's total into its largest blocks, whereas the nonlinear prior learns that such land is empty. Figure 4 shows that this skill is learned from the totals alone. Subregion-R^2 and inhabited blocks show the same ordering (Tables 6 and 9 in Appendix B). Figure 3 shows population in New York City, and Figure 7 in Appendix B shows jobs.

Calibration. The per-subregion intervals reach a marginal coverage between 0.88 and 0.92 at the nominal 90% level in every configuration, across resolutions, countries, and targets, including the negative cases below. Most subregions are empty and κ is a quantile of validation errors, so this is a weak test, read only as evidence that the calibration is not mis-scaled. Gridded population products are usually published as one number per cell, without per-subregion intervals (Figure 5), although recent work has begun to attach uncertainty to them (Adams et al., 2025).

4.3 EXPERIMENT 3: WHERE PUFFIN DOES NOT HELP

Data. For the 2024 US presidential election, we use the turnout, that is, the total number of votes cast, in each of 163,690 precincts (Datar et al., 2025), with AlphaEarth embeddings averaged over each precinct polygon. Here the regions are the 3,096 counties and the subregions are their precincts.

Target	Region → subregion	Reg.-avg.	Bicubic	Goodman	PUFFIN	90% cov.
population	tract → block	+0.000	−0.407	−1.037	**+0.377**	0.90
jobs	tract → block	+0.000	+0.057	+0.050	**+0.209**	0.90
turnout	county → precinct	**+0.000**	−0.104	−0.313	−0.434	0.90
gun violence	tract → block	+0.000	−0.079	−0.075	**+0.014**	0.90

Table 2: Within-region R^2 on the **US**, all fine units. The top rows are Experiment 2, with census tracts as regions and census blocks as subregions (83,774 tracts, 8.0M blocks), where every method uses the area term. The bottom rows are Experiment 3, where precinct turnout does not follow what the embedding sees and 97% of blocks record no gun-violence incident, so no method finds much to place. Best per row in bold. The baselines are deterministic, and PUFFIN is a 5-seed mean (std ≤ 0.03) reported with its 90% interval coverage.

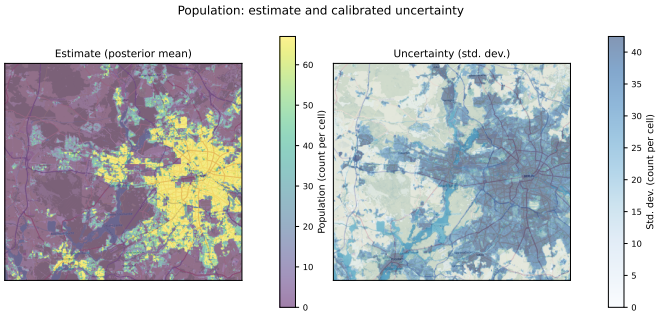

Figure 5: **Calibrated per-subregion uncertainty (Berlin).** PUFFIN estimate (left) and calibrated per-subregion standard deviation (right). Uncertainty concentrates at populated transitions, while empty areas are confidently near zero. Gridded population products commonly provide the left map alone.

The Gun Violence Archive records geocoded incidents (Gun Violence Archive, 2013), which we count per census block, with census tracts as regions and census blocks as subregions, as in Experiment 2.

Turnout. Precinct turnout does vary within counties, where about half of its variance lies, but not in a way that follows what the embedding sees. Precincts are delineated around voters, so a dense urban precinct and a vast rural one can have similar turnout, and within a county areas differ by about 110× while turnout differs by only about 9×. An area term therefore makes the split much worse and we use none here, yet every method that redistributes turnout by the embedding still adds error, and the less it redistributes the better it does (Table 2). Bicubic interpolation stays closest to the region average (−0.10), while Goodman's regression (−0.31) and PUFFIN (−0.43) fall below it.

Gun Violence. Gun violence is a rare event. Only 3% of census blocks record any incident, so within a tract nearly every block has the same value of zero and there is little to redistribute. PUFFIN accordingly stays close to the region average (+0.01), while bicubic interpolation and Goodman's regression fall slightly below it. For turnout and gun violence alike, the 90% intervals keep their nominal marginal coverage, so a failure of the point estimate is not hidden by a mis-scaled interval.

5 Discussion and Limitations

The broad lesson is that a single coarse total, read together with a satellite embedding, already pins down a large part of the fine-scale map, but only where the fine variation is real and tied to what the embedding can see. The rest of our findings follow from this. The learned splits reach within-region R^2 between 0.14 and 0.84, far above bicubic interpolation, which points to the embeddings rather than to spatial smoothness as the source of the gain. A linear density with the same embeddings keeps up on a single city and pulls ahead where training regions are few, and on irregular units

both it and bicubic interpolation fall below the region average on population. Beyond accuracy, PUFFIN returns a calibrated posterior, which the baselines do not. Skill grows with the size of the subregion, which fits a hard limit at the finest scale, where one sees roofs from above but not how many people are inside. Because skill depends this much on scale, we report the whole grid rather than a single number. Where the fine-scale variation does not follow what the embedding sees, as for precinct turnout, a learned split does worse than the region average, while its intervals remain calibrated. The model also has structural limits. The Gaussian prior places mass on negative counts, which we remove by clipping and rescaling rather than by a count likelihood. It treats subregions as independent, and it describes each subregion by a single averaged embedding, which can hide a small settlement inside a large rural unit. Its variance scales with area, which suits sparse land but understates the uncertainty of small, dense subregions, and it disaggregates one variable at a time.

6 RELATED WORK

Population Products and Dasymetric Mapping. WorldPop (Stevens et al., 2015), GHS-POP (Schiavina et al., 2023), and LandScan (Lebakula et al., 2025) split census totals using curated map layers. POMELO (Metzger et al., 2022) maps coarse counts to 100 m with a deep network over open geodata, and related work maps population from Sentinel imagery (Metzger et al., 2024). Tobler's mass-preserving interpolation (Tobler, 1979) conserves the total but splits by a fixed rule. Closest in spirit, CLAM (Großmann et al., 2026) likewise recovers a fine field by using the coarse total as the training signal, but targets causal local effects through a structural causal model, whereas we disaggregate observed counts with exact conservation and calibrated uncertainty on real census and satellite data.

Aggregates, Embeddings, and Calibration. Training from totals alone is an instance of learning from aggregate observations (Zhang et al., 2020), specifically the count version of learning from label proportions (Quadrianto et al., 2008). Inferring fine-level quantities from aggregates is also the subject of ecological inference, whose classical regression estimator (Goodman, 1953; 1959) we use as a baseline. We build on AlphaEarth (Brown et al., 2025), whose 64-dimensional 10 m embeddings are roughly additive, so averaging them over a subregion is meaningful and a small model suffices. Predicting socioeconomic outcomes from satellite imagery is well established (Jean et al., 2016; Yeh et al., 2020; Rolf et al., 2021; Hou et al., 2026), but those methods predict a value at or within the reporting unit by supervised regression, whereas we disaggregate a coarse total below the unit under weak supervision. Conformal prediction (Romano et al., 2019) usually needs labels at the target resolution, whereas the step of our calibration that adapts to a new area uses coarse residuals alone, and fine labels enter only through a single global factor fitted on validation regions.

7 CONCLUSIONS AND FUTURE WORK

PUFFIN casts spatial disaggregation as probabilistic constraint satisfaction. Trained only on coarse totals, it learns a prior over every subregion and conditions it on the observed totals, which gives an estimate that is non-negative, aggregates back to each total exactly, and carries a calibrated uncertainty. From one satellite embedding and no fine labels for training, it holds across resolutions, two countries, and several variables, and its failures follow a simple pattern. Where the fine-scale signal is almost absent or does not follow what the embedding sees, it falls back to or below the regional average. Needing only a published total and a globally available embedding, it applies where no fine-resolution reference grid exists, which is where disaggregation is needed most.

Future Work. The clearest next step is a head-to-head comparison against deployed gridded products such as WorldPop (Stevens et al., 2015), GHS-POP (Schiavina et al., 2023), and LandScan (Lebakula et al., 2025) on the same held-out regions, and an extension to more countries. On the modeling side, describing large units by a grid of cells instead of one averaged embedding, letting the variance grow with the predicted count instead of the area, and conditioning several correlated variables jointly on all of their aggregation constraints are natural extensions of the same posterior.

AI USE STATEMENT

We used generative AI tools in two ways: to aid and polish the writing, and for retrieval and discovery, that is, to cross-check our literature search for relevant work we might have overlooked. Generative AI was not used to generate the research idea or the method, and it produced no experimental result. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Every data source is public: the Zensus 2022 grid (Statistisches Bundesamt, 2024), the 2020 US census (U.S. Census Bureau, 2021), LODES (U.S. Census Bureau, 2024), the Gun Violence Archive (Gun Violence Archive, 2013), the 2024 precinct returns (Datar et al., 2025), and the AlphaEarth embeddings (Brown et al., 2025). Section 3.2 states the model, the loss, the split, and the calibration, and Appendix A gives the dataset construction, the fold assignment, and every optimization setting. The anonymized code that produced every number in this paper is linked in the abstract.

ETHICS STATEMENT

This work uses only publicly released data (census counts, workplace job counts, publicly reported incident records, and precinct turnout) together with public satellite embeddings. PUFFIN redistributes totals that are already published, uses no data about identified individuals, and keeps every official total unchanged. Fine-scale estimates of sensitive variables such as election turnout or violent incidents can nevertheless be mistaken for observations or misused to single out small areas, and Experiment 3 shows that such estimates can be unreliable. This is why every estimate comes with a calibrated uncertainty, and why the method is intended for planning and research at an aggregate level, not for inferring anything about individuals.

REFERENCES

Daniel S Adams, Andrew Zimmer, Joseph Tuccillo, Jessica Moehl, Robert Stewart, Sarah Walters, Budhendra Bhaduri, Peter Li, and Marie Urban. LandScan mosaic enables high-resolution gridded population estimates with explicit uncertainty. *Scientific Reports*, 15(1):44493, 2025.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Ray W Clough and James L Tocher. Finite element stiffness matrices for analysis of plate bending. In *Proc. of the First Conf. on Matrix Methods in Struct. Mech.*, pp. 515–545, 1965.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Saurabh Datar, Alex Lemonides, Ilana Marcus, Eli Murray, Ethan Singer, and Christine Zhang. Precinct-Level Results and Shapefiles for the 2024 U.S. Presidential General Election. `https://www.nytimes.com/2025/01/15/us/elections/2024-election-map-data.html`, 2025. URL `https://github.com/nytimes/presidential-precinct-map-2024`.

Leo A Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 18(6):663–664, 1953.

Leo A Goodman. Some alternatives to ecological correlation. *American Journal of Sociology*, 64 (6):610–625, 1959.

Gerrit Großmann, Sumantrak Mukherjee, and Sebastian J Vollmer. CLAM: Causal spatial disaggregation to infer local effects from coarse data. *arXiv preprint arXiv:2608.08064*, 2026.

Gun Violence Archive. Gun Violence Archive, 2013. URL https://www.gunviolencearchive.org/. Accessed: 2026-09-22.

Shuyang Hou, Ziqi Liu, Haoyue Jiao, Lutong Xie, Yaxian Qing, Xiaopu Zhang, Qingyang Xu, Zhangyan Xu, Xuefeng Guan, and Huayi Wu. AEF-Econ: Toward plug-and-play socioeconomic foundation embeddings from AlphaEarth for urban remote sensing. *arXiv preprint arXiv:2606.20697*, 2026.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794, 2016.

Viswadeep Lebakula, Kelly Sims, Andrew Reith, Amy Rose, Jake McKee, Phil Coleman, Jason Kaufman, Marie Urban, Chris Jochem, Carrie Whitlock, et al. LandScan global 30 arcsecond annual global gridded population datasets from 2000 to 2022. *Scientific Data*, 12(1):495, 2025.

Nando Metzger, John E Vargas-Muñoz, Rodrigo C Daudt, Benjamin Kellenberger, Thao Ton-That Whelan, Ferda Ofli, Muhammad Imran, Konrad Schindler, and Devis Tuia. Fine-grained population mapping from coarse census counts and open geodata. *Scientific Reports*, 12(1):20085, 2022.

Nando Metzger, Rodrigo Caye Daudt, Devis Tuia, and Konrad Schindler. High-resolution population maps derived from Sentinel-1 and Sentinel-2. *Remote Sensing of Environment*, 314:114383, 2024.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine learning*, pp. 776–783, 2008.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Marcello Schiavina, Sergio Freire, Alessandra Carioli, and Kytt MacManus. GHS-POP R2023A – GHS population grid multitemporal (1975–2030), 2023. URL http://data.europa.eu/89h/2ff68a52-5b5b-4a22-8f40-c41da8332cfe.

Statistisches Bundesamt. Zensus 2022 – Ergebnisse des registergestützten Zensus. https://ergebnisse.zensus2022.de/, 2024. Accessed: 2026-09-22.

Forrest R. Stevens, Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, 10(2):e0107042, 2015.

Waldo R Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979.

Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

U.S. Census Bureau. 2020 Census Redistricting Data (Public Law 94-171) Summary File and 2020 TIGER/Line Shapefiles, 2021. URL https://www.census.gov/geographies/mapping-files/2020/geo/tiger-line-file.html.

U.S. Census Bureau. LEHD Origin-Destination Employment Statistics (LODES). U.S. Census Bureau, Longitudinal Employer-Household Dynamics Program, 2024. URL https://lehd.ces.census.gov/data/lodes/. Accessed: 2026-09-22.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1):2583, 2020.

Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. *Advances in Neural Information Processing Systems*, 33:7993–8005, 2020.

A Experimental Details

German Grids. The Berlin bounding box holds $979{,}200$ land cells, of which $92{,}724$ are inhabited, and Germany holds 35.7M land cells, of which 3.1M are inhabited. Regions are square grids of size $R \in \{200\,\mathrm{m}, 1\,\mathrm{km}, 2\,\mathrm{km}, 4\,\mathrm{km}, 10\,\mathrm{km}\}$, each split into 2, 5, 10, 20, 50 or 100 subregions per side, keeping the 18 of those whose subregions are at least $100\,\mathrm{m}$ wide. A subregion's outcome is the sum of the $100\,\mathrm{m}$ census cells it contains, and its covariates are the AlphaEarth embeddings (Brown et al., 2025) averaged over it. For the US the covariates are averaged over each block or precinct polygon instead.

Empty Subregions. Nothing in the method excludes subregions that hold nobody, and the model treats them like any other subregion, so where the embedding looks like forest or water the observed totals push the predicted mean toward zero. What does not come for free is having such subregions in the data. The German Zensus grid lists only cells with at least one resident, so uninhabited cells are missing from the file rather than recorded as zero, and a model trained on the file alone could only rank cells that are already known to be inhabited. We therefore build the grid over every $100\,\mathrm{m}$ land cell and label the cells the census does not list as zero. Absent means zero only inside the surveyed country, so for Germany, whose bounding box reaches into neighboring countries, cells outside the national border are dropped rather than labeled zero. US census blocks need none of this, because the file already lists every block, with a count of zero where nobody lives.

Baselines. Bicubic interpolation treats the regional densities $\widehat{Y}_i / \sum_j a_{i,j}$ as samples of a smooth surface and evaluates it at every subregion centroid. On regular grids we use cubic B-spline interpolation over the region centers (Unser, 1999). For irregular US units, whose centers do not form a grid, we use the piecewise-cubic Clough–Tocher interpolant over the region centroids (Clough & Tocher, 1965). Goodman's ecological regression is a ridge regression of the regional density on the area-weighted mean of the standardized embeddings in the region, fitted on the training regions, with the ridge strength chosen on the totals of the validation regions. In both cases the fitted density is applied to every subregion, multiplied by its area, clipped at zero, and rescaled to each regional total.

Optimization. The embeddings are standardized per dimension with statistics computed on the training regions only. We train for $2{,}000$ epochs of AdamW (peak learning rate $3 \cdot 10^{-3}$, weight decay 10^{-4}) with a cosine-annealed schedule, evaluating the aggregate likelihood of the validation regions each epoch and keeping the best checkpoint. Germany and the US use minibatches of whole regions, and the smaller settings are full-batch. Folds are fixed across seeds and methods, so PUFFIN and the baselines see exactly the same splits, and the 5 seeds vary only the network initialization.

B Additional Results

Interval coverage is 0.898 to 0.902 across the German configurations and 0.880 to 0.918 across the Berlin ones. The main paper reports within-region R^2 on all subregions, summarized over configurations, and Figure 6 adds the learning curves of the remaining settings. Tables 3 and 4 give the per-configuration results behind Table 1, for every method. This appendix then collects the two other views of the same runs: the inhabited-only results and the subregion-R^2 results. Tables 7 and 8 repeat the within-region comparison on inhabited cells only, and Table 9 repeats Table 2 on inhabited blocks. Tables 5 and 6 report subregion-R^2, which unlike within-region R^2 also credits the between-region differences that the observed totals give every method, so even the region average scores high. The ordering of the methods is the same throughout.

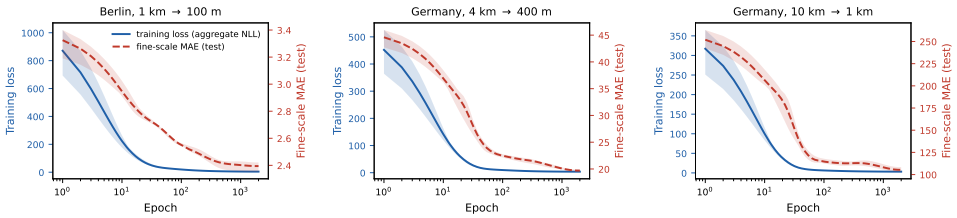

Figure 6: Training loss and fine-scale test error over training, read as Figure 4, for the remaining settings.

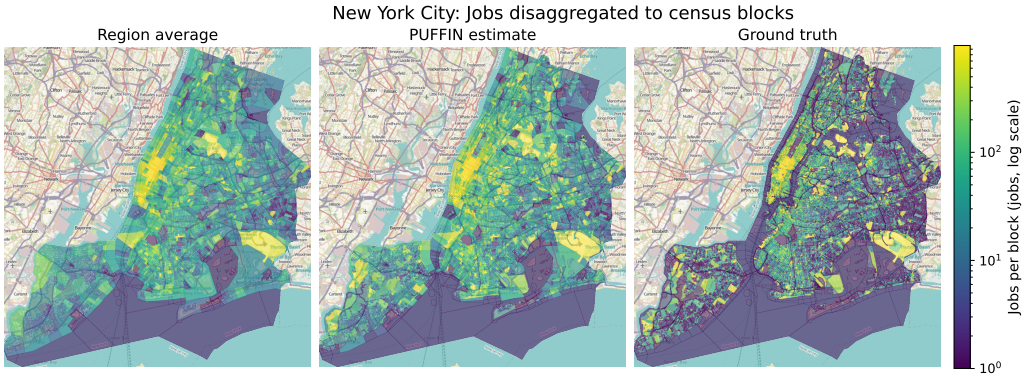

Figure 7: **US jobs in New York City**, with census tracts as regions and census blocks as subregions. Region average (left), the PUFFIN estimate (middle), and ground truth (right). Jobs concentrate in a few small commercial blocks, which the region average cannot express and PUFFIN recovers from the embeddings and the tract totals alone.

		Fine cells per side (R/s)					
Method	Region size R	2	5	10	20	50	100
Bicubic	200 m	.11	–	–	–	–	–
	1 km	.21	.11	.06	–	–	–
	2 km	.17	.11	.07	.04	–	–
	4 km	.20	.11	.07	.05	–	–
	10 km	.54	.27	.18	.12	.08	.05
Goodman	200 m	**.17**	–	–	–	–	–
	1 km	**.49**	**.40**	**.28**	–	–	–
	2 km	**.53**	.46	.38	.28	–	–
	4 km	**.57**	.50	.44	.36	–	–
	10 km	**.71**	**.70**	**.59**	**.50**	**.39**	**.29**
PUFFIN (prior)	200 m	.07	–	–	–	–	–
	1 km	.27	.29	.21	–	–	–
	2 km	.39	.41	.34	.24	–	–
	4 km	.46	.44	.40	.32	–	–
	10 km	.56	.49	.44	.38	.30	.21
PUFFIN	200 m	.14	–	–	–	–	–
	1 km	.43	.38	**.28**	–	–	–
	2 km	.50	**.49**	**.42**	**.30**	–	–
	4 km	.54	**.51**	**.47**	**.39**	–	–
	10 km	.70	.59	.52	.46	.38	.28

Table 3: Within-region R^2 on **Berlin**, all land cells, per configuration and method. The PUFFIN rows are those of Table 1. Each row is a coarse region size R and each column the number of fine cells per side, so the fine-cell size is R divided by that number (100 means 100×100 fine cells), and blanks are configurations with sub-100 m cells. The region-average baseline scores 0 by construction. Best method per configuration in bold. PUFFIN: 5-seed mean (std $\leq$ 0.03), and the baselines are deterministic. PUFFIN (prior) is the prior alone, before conditioning on the totals.

Method	Region size R	Fine cells per side (R/s) 2	5	10	20	50	100
Bicubic	200 m	.14	–	–	–	–	–
	1 km	.23	.13	.08	–	–	–
	2 km	.16	.10	.07	.05	–	–
	4 km	.16	.09	.06	.04	–	–
	10 km	.22	.12	.07	.05	.03	.02
Goodman	200 m	**.19**	–	–	–	–	–
	1 km	**.62**	**.48**	.34	–	–	–
	2 km	.69	.59	.49	.36	–	–
	4 km	.73	.65	.58	.48	–	–
	10 km	.80	.74	.66	.59	.47	.35
PUFFIN (prior)	200 m	.09	–	–	–	–	–
	1 km	.44	.39	.28	–	–	–
	2 km	.59	.53	.44	.32	–	–
	4 km	.67	.63	.54	.44	–	–
	10 km	.71	.70	.65	.55	.43	.32
PUFFIN	200 m	.18	–	–	–	–	–
	1 km	.59	**.48**	**.35**	–	–	–
	2 km	**.70**	**.61**	**.51**	**.37**	–	–
	4 km	**.77**	**.69**	**.61**	**.50**	–	–
	10 km	**.84**	**.76**	**.70**	**.61**	**.48**	**.36**

Table 4: Within-region R^2 on **Germany**, all 35.7M land cells, per configuration and method, read as Table 3. Each row is a coarse region size R and each column the number of fine cells per side, so the fine-cell size is R divided by that number (100 means 100×100 fine cells), and blanks are configurations with sub-100 m cells. The region-average baseline scores 0 by construction. Best method per configuration in bold. PUFFIN: 5-seed mean (std ≤ 0.01), and the baselines are deterministic. PUFFIN (prior) is the prior alone, before conditioning on the totals.

Config (R / s)	inhabited Reg.-avg.	Bicubic	Goodman	PUFFIN	all land cells Reg.-avg.	Bicubic	Goodman	PUFFIN
10 km / 100 m	0.103	0.168	**0.409**	0.402	0.247	0.287	**0.466**	0.460
2 km / 100 m	0.310	0.339	0.499	**0.511**	0.385	0.411	0.557	**0.569**
1 km / 100 m	0.410	0.444	0.557	**0.558**	0.468	0.501	**0.619**	0.617
1 km / 200 m	0.581	0.621	**0.734**	0.727	0.622	0.663	**0.775**	0.767
10 km / 1 km	0.506	0.594	**0.791**	0.753	0.527	0.611	**0.805**	0.772
10 km / 5 km	0.828	0.920	**0.950**	0.949	0.829	0.920	**0.950**	0.949

Table 5: Subregion-R^2 on **Berlin** for representative configurations, inhabited cells and all land cells. Unlike within-region R^2, this metric also credits the between-region differences that the observed totals give every method, so even the region average scores high. Best per group in bold, and PUFFIN std ≤ 0.01.

Target	Reg.-avg.	Bicubic	Goodman	PUFFIN
population	0.222	−0.095	−0.584	**0.515**
jobs	0.087	0.139	0.133	**0.277**

Table 6: Subregion-R^2 on the **US**, census tracts as regions and census blocks as subregions, all blocks, for the runs of Table 2. Best per row in bold, and PUFFIN std ≤ 0.02.

Method	Region size R	Fine cells per side (R/s) 2	5	10	20	50	100
Bicubic	200 m	.06	–	–	–	–	–
	1 km	.19	.07	.03	–	–	–
	2 km	.16	.09	.05	.03	–	–
	4 km	.19	.11	.07	.04	–	–
	10 km	.54	.27	.18	.13	.09	.07
Goodman	200 m	**.10**	–	–	–	–	–
	1 km	**.47**	**.33**	**.16**	–	–	–
	2 km	**.52**	.43	.32	.18	–	–
	4 km	**.57**	.48	.41	.31	–	–
	10 km	**.71**	**.69**	**.57**	**.47**	**.37**	**.26**
PUFFIN (prior)	200 m	.06	–	–	–	–	–
	1 km	.30	.27	.14	–	–	–
	2 km	.39	.40	.31	.17	–	–
	4 km	.46	.44	.40	.31	–	–
	10 km	.56	.49	.43	.37	.32	.24
PUFFIN	200 m	**.10**	–	–	–	–	–
	1 km	.41	.31	**.16**	–	–	–
	2 km	.49	**.45**	**.34**	**.19**	–	–
	4 km	.54	**.49**	**.44**	**.33**	–	–
	10 km	.70	.58	.50	.42	.35	.25

Table 7: Within-region R^2 on **Berlin**, inhabited cells only, read as Table 3. Each row is a coarse region size R and each column the number of fine cells per side, so the fine-cell size is R divided by that number (100 means 100×100 fine cells), and blanks are configurations with sub-100 m cells. The region-average baseline scores 0 by construction. Best method per configuration in bold. PUFFIN: 5-seed mean (std ≤ 0.03), and the baselines are deterministic. PUFFIN (prior) is the prior alone, before conditioning on the totals.

Method	Region size R	Fine cells per side (R/s) 2	5	10	20	50	100
Bicubic	200 m	.09	–	–	–	–	–
	1 km	.20	.08	.03	–	–	–
	2 km	.16	.08	.05	.03	–	–
	4 km	.16	.08	.05	.03	–	–
	10 km	.22	.12	.07	.05	.03	.03
Goodman	200 m	**.11**	–	–	–	–	–
	1 km	**.59**	**.37**	.16	–	–	–
	2 km	.67	.53	.37	**.18**	–	–
	4 km	.73	.63	.52	.37	–	–
	10 km	.80	.73	.65	.54	.36	.20
PUFFIN (prior)	200 m	.08	–	–	–	–	–
	1 km	.44	.32	.14	–	–	–
	2 km	.59	.49	.36	.16	–	–
	4 km	.68	.62	.51	.36	–	–
	10 km	.72	.70	.65	.54	.37	.20
PUFFIN	200 m	**.11**	–	–	–	–	–
	1 km	.56	**.37**	**.17**	–	–	–
	2 km	**.69**	**.55**	**.39**	**.18**	–	–
	4 km	**.77**	**.67**	**.55**	**.39**	–	–
	10 km	**.84**	**.75**	**.69**	**.58**	**.40**	**.22**

Table 8: Within-region R^2 on **Germany**, inhabited cells only, read as Table 3. Each row is a coarse region size R and each column the number of fine cells per side, so the fine-cell size is R divided by that number (100 means 100×100 fine cells), and blanks are configurations with sub-100 m cells. The region-average baseline scores 0 by construction. Best method per configuration in bold. PUFFIN: 5-seed mean (std ≤ 0.01), and the baselines are deterministic. PUFFIN (prior) is the prior alone, before conditioning on the totals.

Target	within-region R^2 Reg.-avg.	Bicubic	Goodman	PUFFIN	subregion-R^2 Reg.-avg.	Bicubic	Goodman	PUFFIN
population	$+0.000$	-0.297	-0.901	$\mathbf{+0.429}$	0.231	0.035	-0.406	$\mathbf{0.568}$
jobs	$+0.000$	$+0.068$	$+0.073$	$\mathbf{+0.183}$	0.094	0.174	0.179	$\mathbf{0.278}$

Table 9: **US**, census tracts as regions and census blocks as subregions, inhabited blocks only, read as Table 2. PUFFIN std ≤ 0.02.