BALANCING QUANTITY AND REPRESENTATIVENESS IN CONSTRAINED GEOSPATIAL DATASET DESIGN

Livia Betti & Esther Rolf
Department of Computer Science
University of Colorado at Boulder
Boulder, CO 80302, USA
{Livia.Betti, Esther.Rolf}@colorado.edu

ABSTRACT

Effective geospatial machine learning (GeoML) relies on high-quality, large-scale datasets, yet geospatial data collection is often costly and logistically challenging. Creating new geospatial datasets frequently requires on-site labeling of data, including collecting data through surveys or scientific instruments, which leads to variable costs across different regions or groups. To address this, we propose a sampling method that jointly maximizes dataset size and representative composition with respect to cost constraints. We evaluate our proposed sampling method by training GeoML models on the selected subsets and comparing their performance to models trained on randomly sampled data. We find that our method leads to improved performance over standard data collection baselines. These findings provide guidance on when to prioritize representation or dataset size and highlight the need for further research into how sampling strategies can enhance model performance.

1 INTRODUCTION

Geospatial machine learning (GeoML) models are increasingly used to inform environmental and social policy, developing an interesting and pressing link between machine learning and real-world applications. Like all machine learning models, to operate at their full potential, GeoML models require a high-quality, representative, and sufficiently large dataset (Roscher et al., 2024). While the archive of unlabeled geospatial data is large and growing, it is infeasible to obtain ground truth labels in all geographic regions for most relevant prediction tasks.

Dataset composition is shown to be a consistent determinant of model performance (Rolf et al., 2021b; Rolf et al.; Ghorbani & Zou, 2019; Wang & Jia, 2023). In recognition of this, active learning and adaptive sampling methods aim to create a small, high-quality dataset, which can improve model performance above a large, randomly chosen set of samples (Tuia et al., 2009; Soman et al., 2023). However, it is difficult to use active learning methods to inform geospatial dataset design, as these methods assume (i) that there is existing training data to start with, and (ii) that there is a uniform cost in labeling all data (Settles, 2011). A significant amount of geospatial data requires on-site data collection, including conducting surveys and measuring with scientific instruments (Rolf et al., 2024). Geospatial data collection efforts thus have costs that vary across space, e.g., as accessing certain terrains can be more challenging and expensive. Here, we investigate approaches for composing high-quality training geospatial training sets in *cost-constrained environments*.

We work towards understanding when optimized sampling is necessary, by studying the impact different methods of sampling can have on model performance. To answer this, we (1) propose a novel algorithm by adapting ideas from general ML to a spatial ML setting, (2) evaluate this framework under different cost structures, representing different possible scenarios, and (3) compare performance of our sampling algorithm to simple random sampling and stratified random sampling baselines. Our results highlight that under cost-constraints, strategic data collection decisions can greatly affect GeoML model performance. This work takes steps toward studying effective geospatial data collection under cost constraints and encourages additional work to further optimize data composition in spatial settings.

2 OPTIMIZING REPRESENTATIVENESS AND QUANTITY OF SPATIAL DATA

Currently, a significant amount of training data in machine learning is collected using standardized methods of simple and stratified random sampling. However, these strategies for sampling data may not optimize predictive power of *geospatial* machine learning models, where cost structures of physical data collection induce a trade-off between collecting data that (i) are **representative**, containing enough data from relevant parts of the environment or region, and (ii) have a **high-quantity of data**, a significant factor in ML model performance across all domains.

To represent this challenge a general problem setting, we assume that were are given an unlabeled list of samples $S = \{s_1, \ldots, s_N\}$, and that each sample is associated with a group g, i.e. $s_i \in g$. We assume that the set of groups $\mathcal{G}$ are discrete and cover the whole population, with $\gamma_g = P_{s\sim\mathcal{D}}[s \in g]$, adopting notation from Rolf et al. (2021b). In a geospatial setting, relevant groups might include administrative boundaries, census tracts, or environmental groups such as ecoregions. To model the variable cost in labeling different samples, we assign a cost c_i to each sample s_i.

To optimize S under budget constraints, we can determine a vector $\mathbf{x} \in \{0,1\}^N$ where the entry x_i = 1 if s_i is collected, and x_i is 0 if s_i is not collected. Then, we aim to solve the following problem:

$$\underset{x\in\{0,1\}^N}{\arg\min} \sum_{g\in\mathcal{G}} \gamma_g \left[\lambda \left(\sum_{i=1}^{N} x_i \mathbb{I}(s_i \in g)\right)^{-1} + (1-\lambda)\left(\sum_{i=1}^{N} x_i\right)^{-1}\right] \text{ subject to } \sum_{i=1}^{N} x_i c_i \leq B \quad (1)$$

where B represents a fixed budget, and $\lambda \in [0, 1]$ is fixed. Note that $\sum_{i=1}^{N} x_i \mathbb{I}(s_i \in g)$ represents the expected number of samples in group g, $\sum_{i=1}^{N} x_i$ the expected sample size, and $\sum_{i=1}^{N} x_i c_i$ the expected cost. Thus, minimizing Equation 1 amounts to jointly maximizing the number of points in each group g, relative to γ_g, and the overall dataset size, with respect to the budget. The parameter λ controls the importance of these objectives. Specifically, $\lambda = 0$ results in a greedy selection of the lowest-cost points, while $\lambda = 1$ encourages diversity before dataset size.

This framework is modified from Rolf et al. (2021b) and the optimal sample allocation problem (Neyman, 1934; Wright, 2020). In particular, the objective in Equation 1 adopts the form of the approximated population risk from Rolf et al. (2021b) and is modified so there are no group-dependent parameters. Furthermore, Equation 1 replicates the optimal sample allocation problem with the added objective of maximizing the overall dataset size.

To solve Equation 1, we relax the constraints $\mathbf{x} \in \{0,1\}^N$ to $\mathbf{x} \in [0,1]^N$, which turns this into a convex optimization problem that we can solve with standard tools. Note that in this relaxed setting, the *expected* sample cost must be less than the budget. The solution to this problem is a N-dimensional vector $\mathbf{x}$ with entries x_i representing the probability that sample i is in the training set. We refer to this process of sampling as OPT.

3 EXPERIMENTAL SETUP

The goal of the following experiments is to evaluate the effectiveness of our proposed data sampling procedure (optimizing Equation 1) for a cost-constrained setting. We evaluate performance with respect to two baselines: a simple random sample (SRS) geographically, and a stratified random sample (StRS), in which an equal number of samples is taken from each strata. Working with an already labeled dataset, we (1) collect a subset using each of the aforementioned methods according to a fixed budget, (2) train a model on the subset, and (3) compare the model performance on a test set across the methods.

We use the USAVars dataset (Rolf et al., 2021a), consisting of 1 km^2 crops of NAIP imagery resampled to 4m/pixel that are centered on 97876 points randomly sampled from the contiguous United States. Each sample is labeled with three outcomes: population density, tree cover percentage, and elevation. Images from the USAVars dataset are featurized using the random convolutional features (RCF) extractor from TorchGeo (Stewart et al., 2022). 4096-dimensional features are extracted using 4×4 patches drawn randomly from the empirical distribution of patches in the training data, and the bias of the convolutional layer is set to -1.0. A ridge regression is fit on the standardized features, using 5 fold cross-validation, and running over 10 values of the regularization parameter, chosen logarithmically from 10^{-5} to 10^{5}.

	C1					C2				
Budget	SRS	StRS	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	SRS	StRS	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$
1000	191	181	316	338	1000	91	73	322	343	1000
2000	373	363	633	675	2000	183	147	646	685	2000
3000	551	545	951	1014	3000	258	225	970	1027	3000
4000	738	727	1267	1351	4000	337	299	1293	1369	4000
5000	928	908	1584	1690	5000	421	377	1614	1713	5000

Table 1: Average number of samples obtained by each sampling method under cost-constraints (budget). Note that for C3, all sampling methods will result in the same number of samples.

To develop general-purpose groups that might apply across all three outcomes in the USAVars dataset, we cluster points by distribution of land cover in each 1 km^2 region (Figure 1). Specifically, we reference the 2016 National Land Cover Database (NLCD) data, which provides classification of 30m resolution areas. For each image, we determine a 16-dimensional simplex representing the distribution of land-cover classes across pixels. These simplices are grouped into 8 clusters (we refer to these as NCLD groups) using k-means clustering. We refer to these groups as NLCD groups.

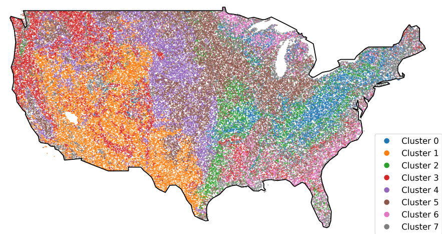

Figure 1: Distribution of NLCD groups, which capture large-scale ground and terrain conditions.

We study the comparison of sampling methods under three cost structures: In cost structure 1 (C1), we assign groups 0, 2, 5, 6, cost 1 and groups 1, 3, 4, 7 cost 10, to represent variable cost in groups. In cost structure 2 (C2), we assign groups 1 and 3 cost 50, and the remaining groups cost 1, to demonstrate a more distinct cost of collection between groups. In cost structure 3 (C3), we take a different approach and assign spatially relevant costs. We assign a sample cost 1 if it is located in the Eastern United States, and ∞ if it is located in the Western United States. This represents an extreme—but not unrealistic—circumstance, where it is infeasible to collect in some regions.

For each cost structure, we compare our method with two baseline strategies: Simple random sampling (SRS) and Stratified random sampling (StRS). For effective comparison, we run OPT with different values of λ. We vary λ across a sparse grid $[0, 0.05, 0.5, 0.95, 1]$; in our main results, we only report on $\lambda = 0, 0.5, 1.0$ for readability. We run each sampling method with respect to a budget B, which is varied across , $B \in [500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000]$ to simulate cost-constrained data collection. For SRS and StRS, sampling is halted once the budget is met. For our method, Equation 1 is solved with respect to B and the resulting sample cost is recorded.

4 Results

Figure 2 and 3 represent our findings for each sampling method on the cost structures from Section 3. In Figure 2 (top, corresponding to C1), OPT ($\lambda = 1$) offers an improvement above SRS and StRS for the population outcome with all costs and the treecover outcome after cost 2000 (See Tables 2, 3, 4 in Appendix). Table 1 shows that as λ decreases from 1 to 0, OPT results in larger training set sizes. However, larger training sets do not necessarily lead to increased model performance, as in Figure 2 (top), OPT ($\lambda = 1$) outperforms OPT ($\lambda = 0.05$) and OPT ($\lambda = 0$), **demonstrating the importance of having a representative training set**. In contrast, for C2 (Figure 2 (bottom)), for all values of λ, OPT leads to a significant improvement above SRS and StRS in the population and treecover outcomes and for the elevation outcome beyond budget 1500. Notably, in each outcome in Figure 2 (bottom), OPT with $\lambda = 0$, leads to a higher R^2 score than all other sampling methods. These results **demonstrate the importance of having a large dataset** when operating under cost constraints. Comparing the top and bottom rows of Figure 2, cost structure C2 leads to more dramatic differences in performance between OPT and the SRS and StRS baselines, implying that our method is particularly effective when some groups are significantly more expensive or difficult to sample. Lastly, in Figure 3, OPT ($\lambda = 1$) yields improvements over SRS and StRS for all budgets

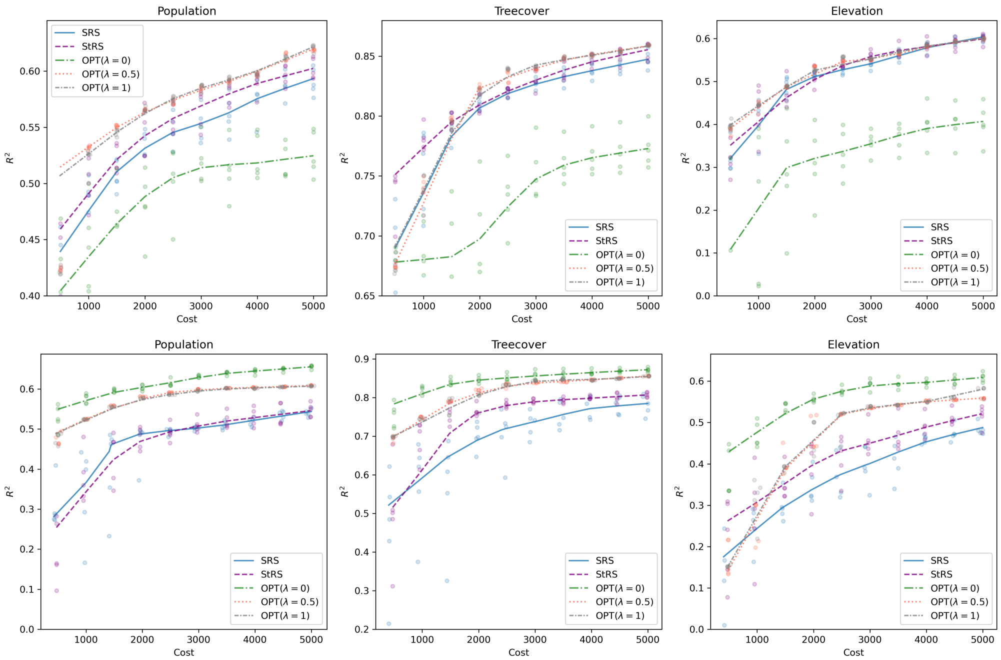

Figure 2: R^2 score vs. cost of collection for the three outcomes and two cost structures (top: C1, bottom: C2). Trendlines generated using locally weighted scatterplot smoothing (LOWESS).

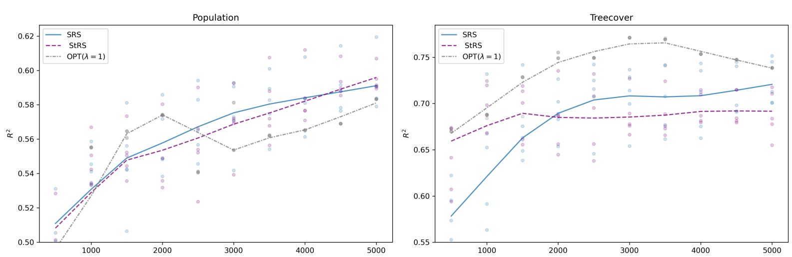

Figure 3: R^2 score vs. cost of collection (cost structure C3) for two outcomes (performance is very low for this out of distribution setting when predicting elevation). Note that for C3, all $\lambda \in (0, 1]$ yield equivalent samples and $\lambda = 0$ yields a simple random sample.

shown in the treecover outcome, but not as many improvements over these baselines in the population outcome. Due to the nature of the NLCD classes, it is possible that these groups are more useful for determining representative samples of treecover percentage than for population density.

5 Conclusion and Future Work

This work represents the first steps toward optimizing sampling of spatial data, a direction that is crucial to maximizing the performance of GeoML models used to study modern day climate change. We propose a novel framework for data collection, which optimizes with respect to group representation and dataset size. We test this framework across different parameters λ and demonstrate that under differing cost constraints, it can lead to improve model performance above simple and stratified random sampling. In this work, we only deployed our sampling framework on one dataset, which is limited to the United States, and one model. In continuing this work, we intend to test our methods on a variety of benchmark datasets with different models to further demonstrate the impact of strategic data collection, as well as on different cost structures, to more fully understand the robustness of our results under different realistic settings.

REFERENCES

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. ISSN 09528385. URL `http://www.jstor.org/stable/2342192`.

Esther Rolf, Ben Packer, Alex Beutel, and Fernando Diaz. Striving for data-model efficiency: Identifying data externalities on group performance. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

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, July 2021a. ISSN 2041-1723. doi: 10.1038/s41467-021-24638-z. URL `https://doi.org/10.1038/s41467-021-24638-z`.

Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pp. 9040–9051. PMLR, 2021b.

Esther Rolf, Lucia Gordon, Milind Tambe, and Andrew Davies. Contrasting local and global modeling with machine learning and satellite data: A case study estimating tree canopy height in African savannas, 2024. URL `https://arxiv.org/abs/2411.14354`.

Ribana Roscher, Marc Rußwurm, Caroline Gevaert, Michael Kampffmeyer, Jefersson A. dos Santos, Maria Vakalopoulou, Ronny Hänsch, Stine Hansen, Keiller Nogueira, Jonathan Prexl, and Devis Tuia. Better, Not Just More: Data-Centric Machine Learning for Earth Observation. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):335–355, December 2024. ISSN 2168-6831, 2473-2397, 2373-7468. doi: 10.1109/MGRS.2024.3470986. URL `http://arxiv.org/abs/2312.05327`. arXiv:2312.05327 [cs].

Burr Settles. From theories to queries: Active learning in practice. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov (eds.), *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pp. 1–18, Sardinia, Italy, 16 May 2011. PMLR. URL `https://proceedings.mlr.press/v16/settles11a.html`.

Satej Soman, Emily Aiken, Esther Rolf, and Joshua Blumenstock. Can strategic data collection improve the performance of poverty prediction models? International Conference on Learning Representations (ICLR 2023) Workshop on Practical Machine Learning for Developing Countries, 2023.

Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*, pp. 1–12, 2022.

Devis Tuia, FrÉdÉric Ratle, Fabio Pacifici, Mikhail F. Kanevski, and William J. Emery. Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, July 2009. ISSN 1558-0644. doi: 10.1109/TGRS.2008.2010404. URL `https://ieeexplore.ieee.org/document/4812037/?arnumber=4812037`. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Jiachen T. Wang and Ruoxi Jia. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 6388–6421. PMLR, April 2023. URL `https://proceedings.mlr.press/v206/wang23e.html`. ISSN: 2640-3498.

Tommy Wright. A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, 165:108829, October 2020. ISSN 01677152. doi: 10.1016/j.spl.2020.108829. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167715220301322`.

A Appendix

A.1 Tables

In this section, we provide tables representing the averaged R^2 score across the 5 random seeds used to sample according to the intended method. Note that as SRS and StRS are run until budget is reached but not exceeded, the true sample cost might be less then the budget. As OPT is solved with an expected total cost, then the true sample cost might be more or less than the budget. We allow a tolerance of ± 10.

	Budget				
Method	1000	2000	3000	4000	5000
SRS	0.49 ± 0.02	0.54 ± 0.00	0.55 ± 0.02	0.57 ± 0.01	0.59 ± 0.01
StRS	0.50 ± 0.05	0.55 ± 0.02	0.57 ± 0.02	0.59 ± 0.01	0.60 ± 0.01
OPT ($\lambda = 0$)	0.44 ± 0.04	0.48 ± 0.03	0.52 ± 0.02	0.52 ± 0.02	0.53 ± 0.02
OPT ($\lambda = 0.05$)	$\mathbf{0.53 \pm 0.00}$	0.56 ± 0.00	0.58 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
OPT ($\lambda = 0.5$)	$\mathbf{0.53 \pm 0.00}$	$\mathbf{0.57 \pm 0.00}$	$\mathbf{0.59 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.62 \pm 0.00}$
OPT ($\lambda = 0.95$)	$\mathbf{0.53 \pm 0.00}$	$\mathbf{0.57 \pm 0.00}$	0.58 ± 0.00	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.62 \pm 0.00}$
OPT ($\lambda = 1$)	$\mathbf{0.53 \pm 0.00}$	0.56 ± 0.00	0.58 ± 0.00	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.62 \pm 0.00}$

Table 2: Average R^2 score on the Test Set for Population with NLCD groups with cost structure C1.

	Budget*				
Method	1000	2000	3000	4000	5000
SRS	0.42 ± 0.05	0.52 ± 0.00	0.54 ± 0.01	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.60 \pm 0.01}$
StRS	0.43 ± 0.06	0.51 ± 0.02	$\mathbf{0.56 \pm 0.01}$	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.60 \pm 0.01}$
OPT ($\lambda = 0$)	0.22 ± 0.18	0.31 ± 0.08	0.36 ± 0.03	0.40 ± 0.05	0.40 ± 0.05
OPT ($\lambda = 0.05$)	$\mathbf{0.46 \pm 0.00}$	0.50 ± 0.00	0.53 ± 0.00	0.54 ± 0.00	0.56 ± 0.00
OPT ($\lambda = 0.5$)	0.45 ± 0.00	$\mathbf{0.54 \pm 0.00}$	0.55 ± 0.00	$\mathbf{0.58 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$
OPT ($\lambda = 0.95$)	0.44 ± 0.00	0.53 ± 0.00	0.55 ± 0.00	0.57 ± 0.00	$\mathbf{0.60 \pm 0.00}$
OPT ($\lambda = 1$)	0.45 ± 0.00	0.53 ± 0.00	0.55 ± 0.00	$\mathbf{0.58 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$

Table 3: Average R^2 score on the Test Set for Elevation with NLCD groups with cost structure C1.

	Budget*				
Method	1000	2000	3000	4000	5000
SRS	0.75 ± 0.04	0.81 ± 0.00	0.83 ± 0.00	0.84 ± 0.01	0.85 ± 0.01
StRS	$\mathbf{0.78 \pm 0.01}$	0.81 ± 0.01	0.83 ± 0.01	0.81 ± 0.01	0.85 ± 0.01
OPT ($\lambda = 0$)	0.65 ± 0.08	0.69 ± 0.05	0.75 ± 0.02	0.77 ± 0.02	0.77 ± 0.02
OPT ($\lambda = 0.05$)	0.77 ± 0.00	0.81 ± 0.00	0.83 ± 0.00	$\mathbf{0.85 \pm 0.00}$	$\mathbf{0.86 \pm 0.00}$
OPT ($\lambda = 0.5$)	0.74 ± 0.01	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.84 \pm 0.00}$	$\mathbf{0.85 \pm 0.00}$	$\mathbf{0.86 \pm 0.00}$
OPT ($\lambda = 0.95$)	0.73 ± 0.01	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.84 \pm 0.00}$	$\mathbf{0.85 \pm 0.00}$	$\mathbf{0.86 \pm 0.00}$
OPT ($\lambda = 1$)	0.73 ± 0.02	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.84 \pm 0.00}$	$\mathbf{0.85 \pm 0.00}$	$\mathbf{0.86 \pm 0.00}$

Table 4: Average R^2 score on the Test Set for Treecover with NLCD groups with cost structure C1.

A.2 Additional Plots

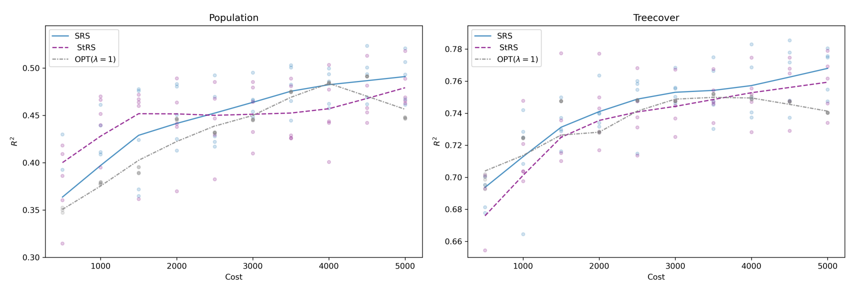

Figure 4: R^2 score vs. cost of collection (cost structure C3 variant with cost 1 assigned if the sample is in the Western United States, and ∞ if the sample is in the Eastern United States). Elevation is excluded, as performance is very low for this out of distribution setting when predicting this outcome. Note that for C3, all $\lambda \in (0, 1]$ yield equivalent samples and $\lambda = 0$ yields a simple random sample.