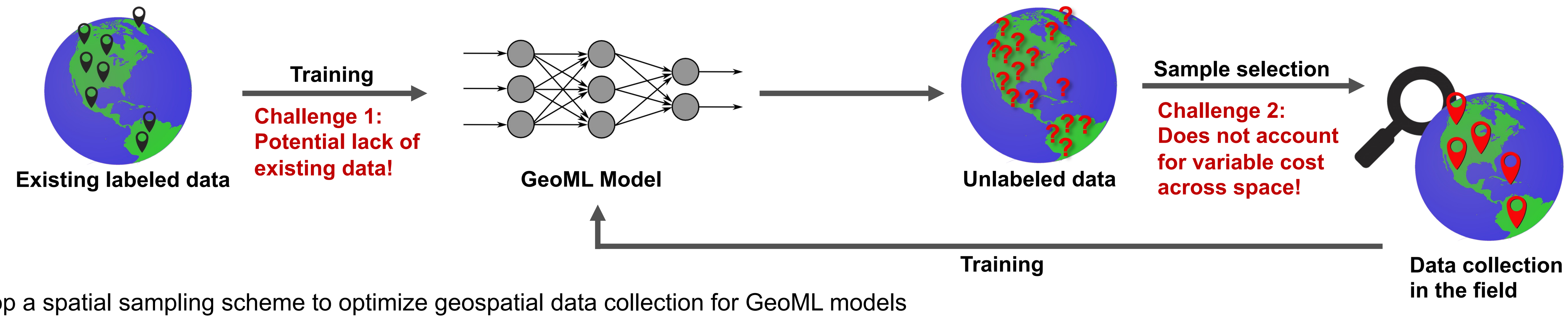
Balancing Quantity and Representativeness in Constrained Geospatial Dataset Design

Livia Betti[1]; Esther Rolf[1]
[1]University of Colorado Boulder
livia.betti@colorado.edu

Motivation

Problem:

⚠ Traditional active learning & subset selection paradigms present challenges for GeoML.



Long-term goal: Develop a spatial sampling scheme to optimize geospatial data collection for GeoML models

→ **Step 1 (Workshop paper focus):** Understand how factors of dataset composition effect GeoML model performance.

Optimizing Representativeness and Quantity

Cost structures of physical data collection induce a trade-off between collecting datasets that

1. **representative**, containing enough data from relevant parts of the region of interest, and
2. have a **high-quantity of data**, a significant factor in ML model performance across all domains.

Objective:

tunable hyperparameter

$$\underset{x\in\{0,1\}^N}{\arg\min} \sum_{g\in\mathcal{G}} \gamma_g \left[\lambda \underbrace{\left(\sum_{i=1}^{N} x_i \mathbb{I}(s_i\in g)\right)^{-1}}_{\text{representative}} + (1-\lambda)\underbrace{\left(\sum_{i=1}^{N} x_i\right)^{-1}}_{\text{high data quantity}}\right] \text{ subject to } \underbrace{\sum_{i=1} x_i c_i}_{\text{total cost}} \le B$$

sample inclusion vector

set of groups covering entire population

group proportions

sample cost

budget

Methods

Objective: Evaluate the effectiveness of our proposed sampling method in constrained settings.

Steps:

1. Obtain sample subset according to sampling method with respect to budget
2. Train model on selected sample
3. Compare performance across sampling method

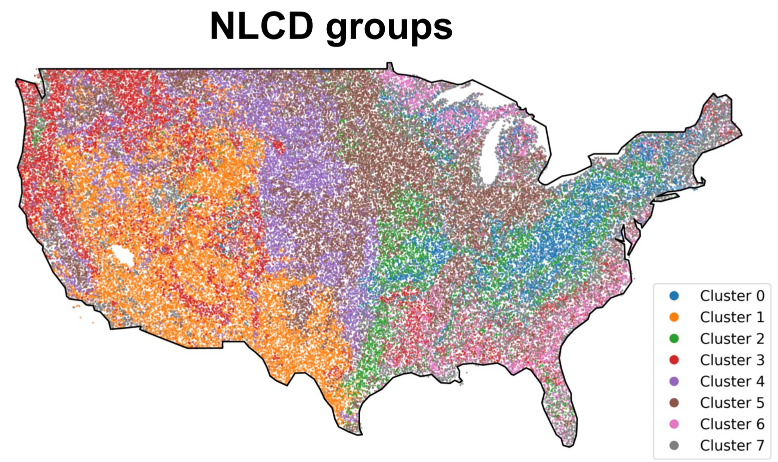
Dataset: USAVars [1]

Model:

1. Feature extraction to create 4096-dimensional features.
2. Ridge regression fit on standardized features.

Groupings: Points are clustered by land cover distribution in each 1 km^2 region using the 2016 National Land Cover Database (NLCD) 30m classifications.

NLCD groups



Cost Structures:

- Cost Structure 1 (Moderate cost difference): Groups 0, 2, 5, 6 cost 1; Groups 1, 3, 4, 7 cost 10.
- Cost Structure 2 (Extreme cost difference): Groups 1 and 3 cost 50; other groups cost 1.

Results

	Cost Structure 1					Cost Structure 2				
Budget	Simple Random	Stratified Random	Ours ($\lambda=1$)	Ours ($\lambda=0.05$)	Ours ($\lambda=0$)	Simple Random	Stratified Random	Ours ($\lambda=1$)	Ours ($\lambda=0.05$)	Ours ($\lambda=0$)
1000	191	181	316	528	1000	91	73	322	510	1000
2000	373	363	633	1054	2000	183	147	646	1018	2000
3000	551	545	951	1581	3000	258	225	970	1529	3000
4000	738	727	1267	2109	4000	337	299	1293	2037	4000
5000	928	908	1584	2637	5000	421	377	1614	2550	5000

***Table 1:* Average number of samples obtained by each sampling method under budget constraints.** Cost Structure 1 (moderate cost difference) and Cost Structure 2 (extreme cost difference).

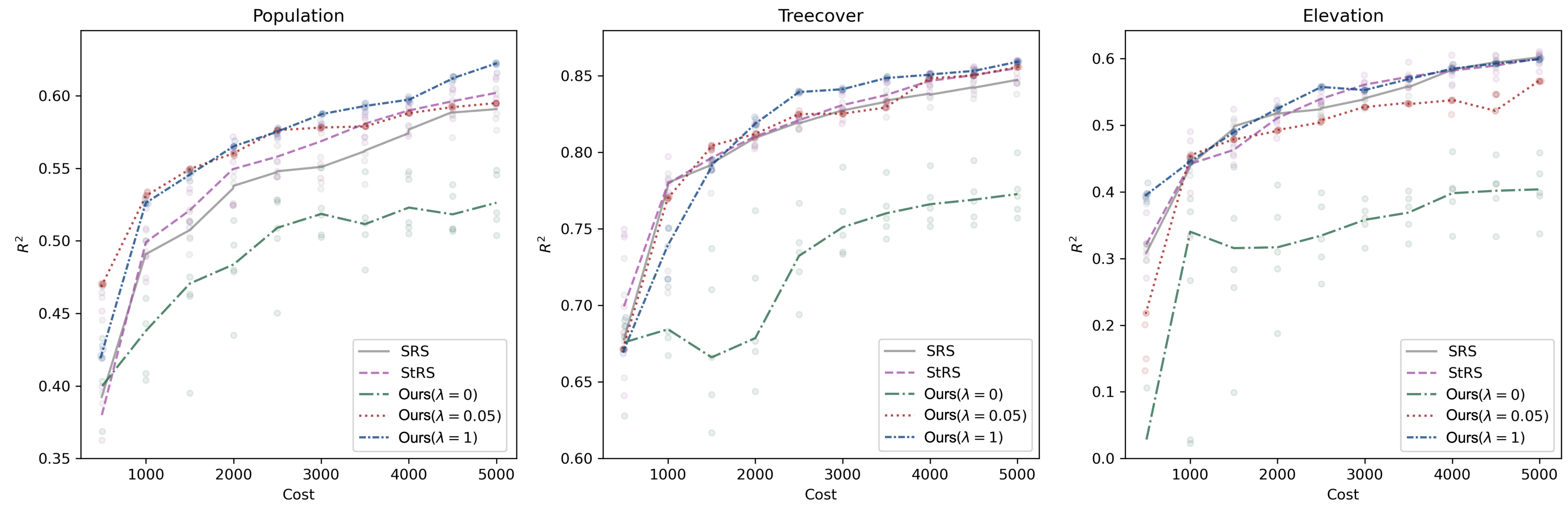


***Figure 1:* R^2 vs. cost of collection for Cost Structure 1.**

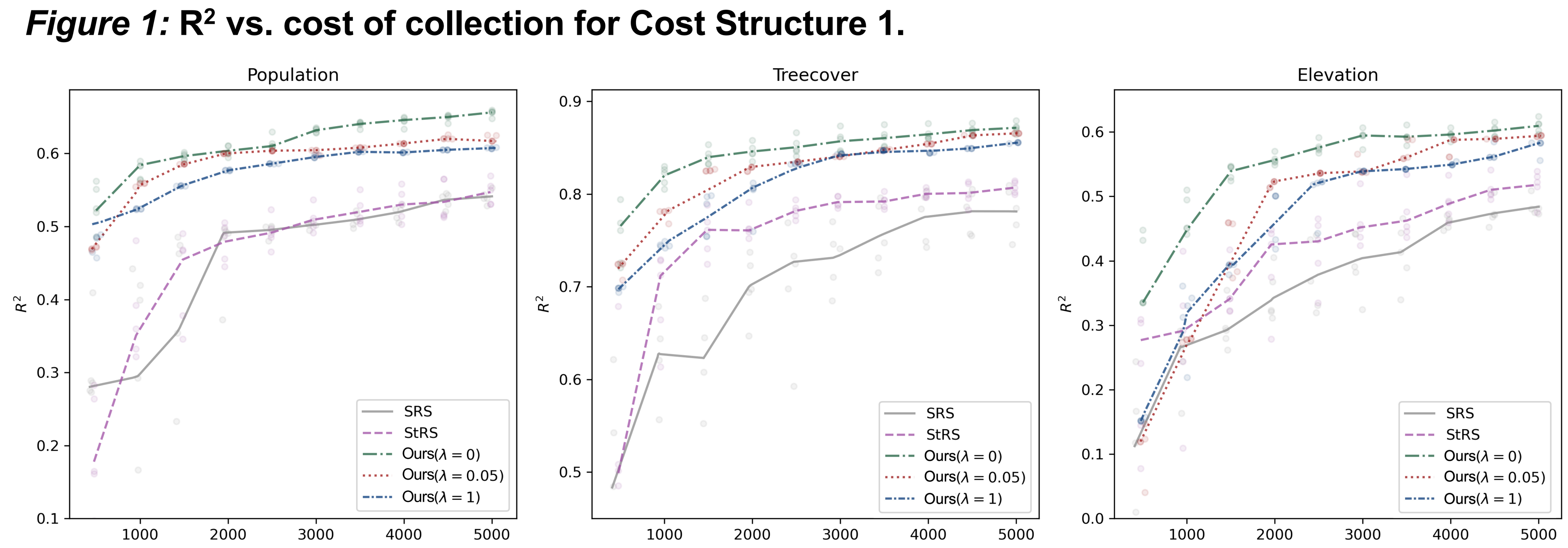


***Figure 2:* R^2 vs. cost of collection for Cost Structure 2.**

Takeaways

Takeaway 1. Larger training sets do not necessarily lead to increased model performance, as for cost structure 1, our method with λ=1 outperforms λ=0.05 and λ=0. This demonstrates **the importance of having a representative training set.**

Takeaway 2. For cost structure 2, our method with all values of λ leads to significant improvements above simple random and stratified random sampling in the population and treecover outcomes. This demonstrates **the importance of having a large dataset** when operating under cost constraints.

Takeaway 3. Our method is particularly effective when some groups are significantly more expensive or difficult to sample.

References

1. 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, 2021.
2. Esther Rolf, Theodora T. Worledge, Benjamin Recht, and Michael Jordan. Representation matters: Assessing the importance of subgroup allocations in training data. ICML, 2021.