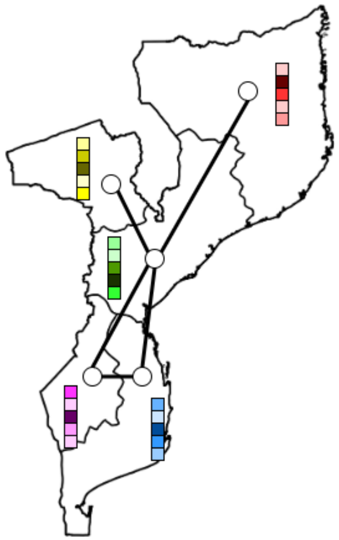
Estimating Population in Sub-Saharan Africa Using Graph Neural Networks and Earth Observation Data

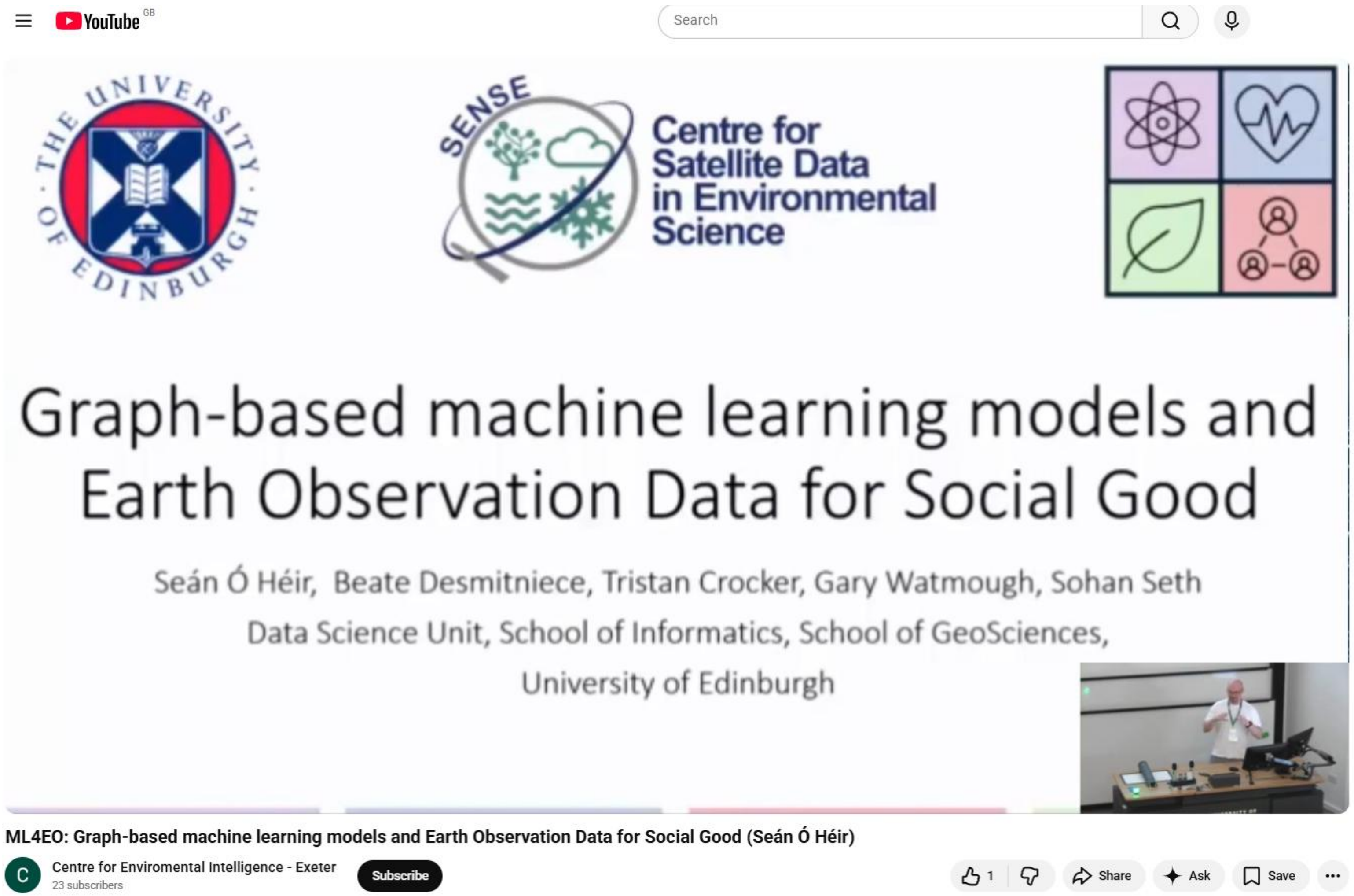
Seán Ó Héir, Gary Watmough, Sohan Seth

Data Science Unit, School of Informatics, School of GeoSciences,

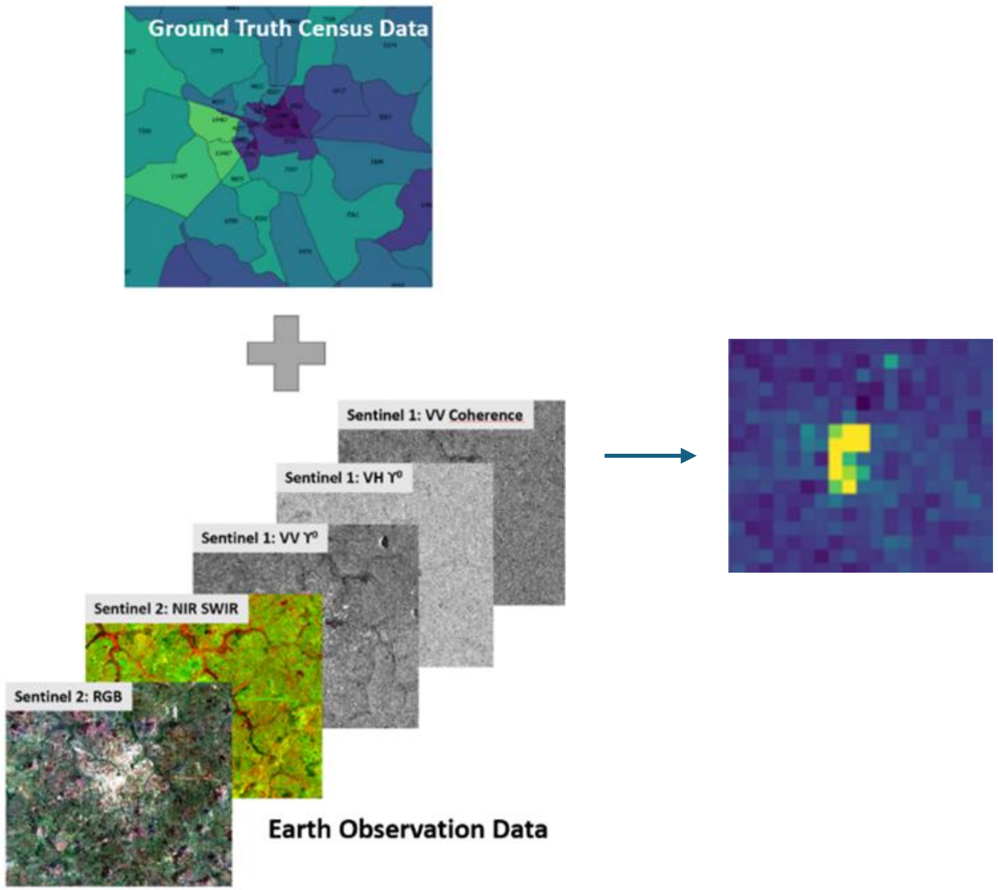
University of Edinburgh

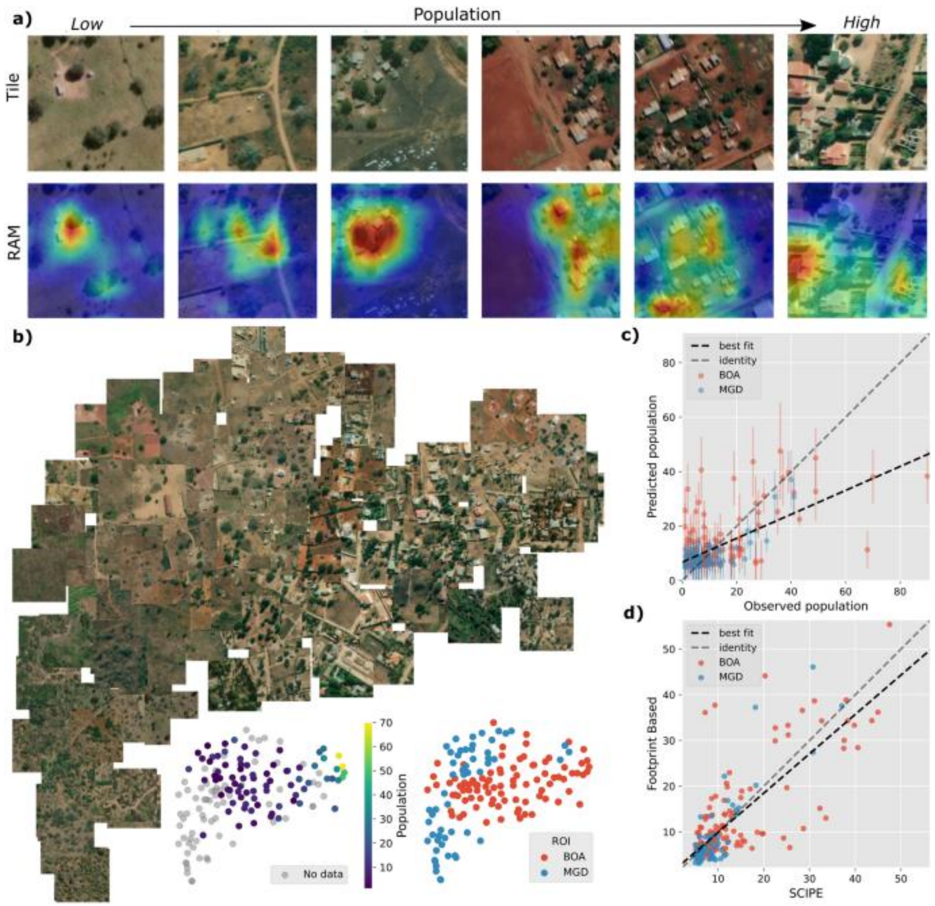


I was here last year… And now I’m back!

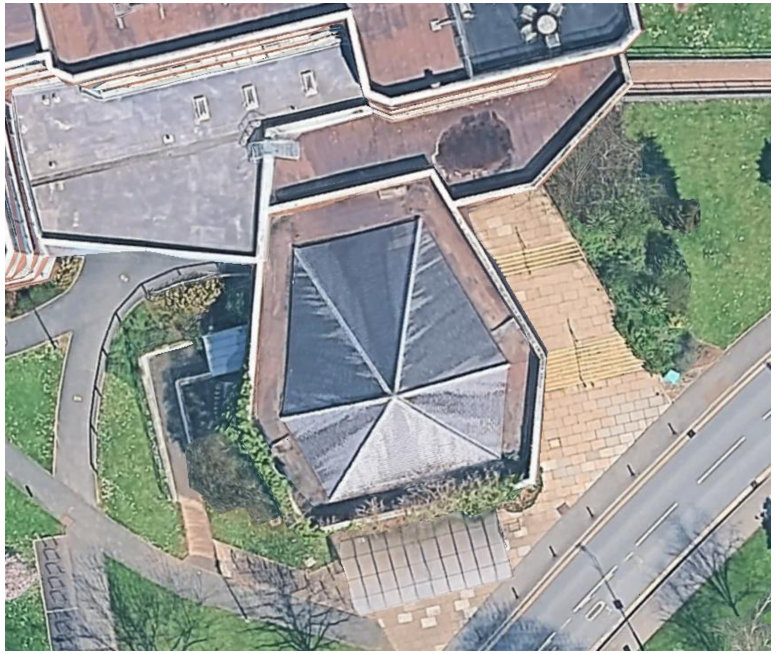


Estimating Population with Earth Observation

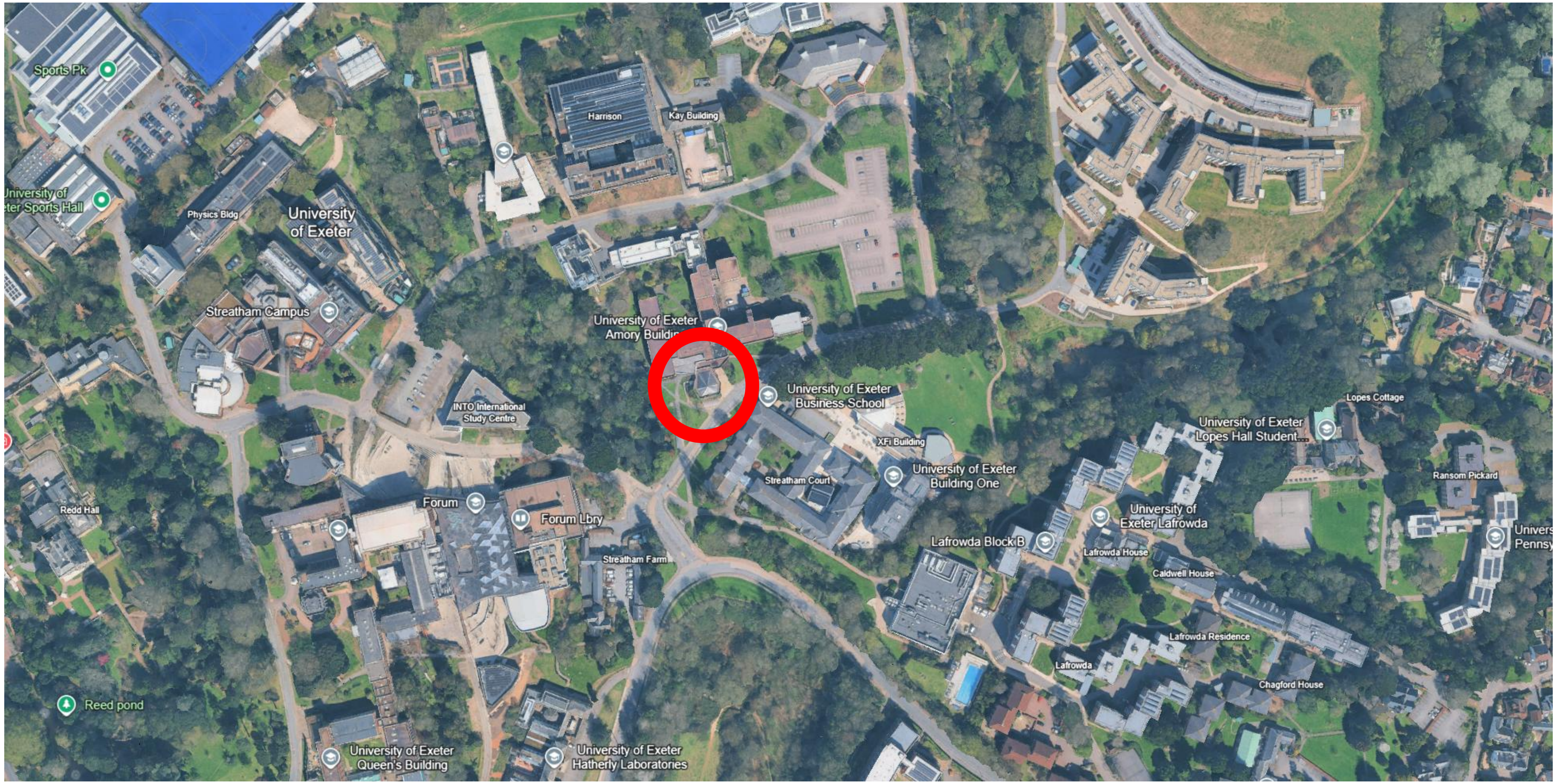




What kind of building is this?

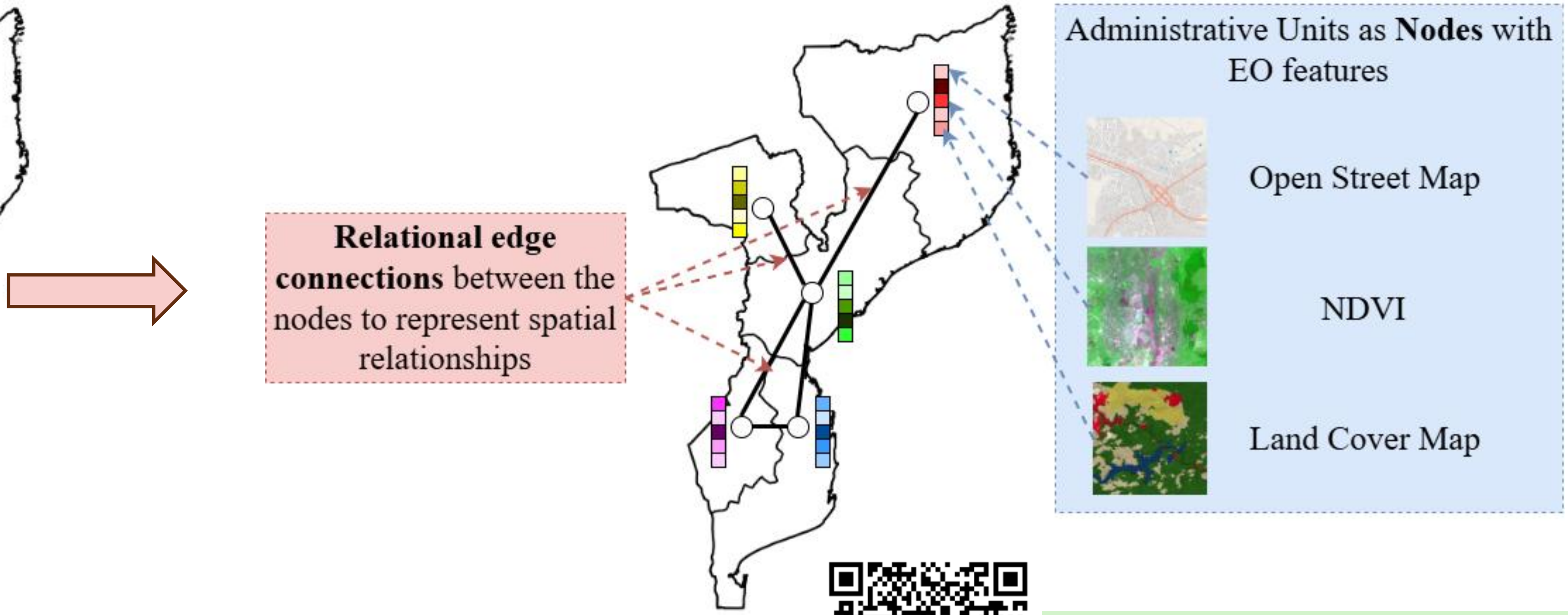


Zoom out to gain **spatial context**



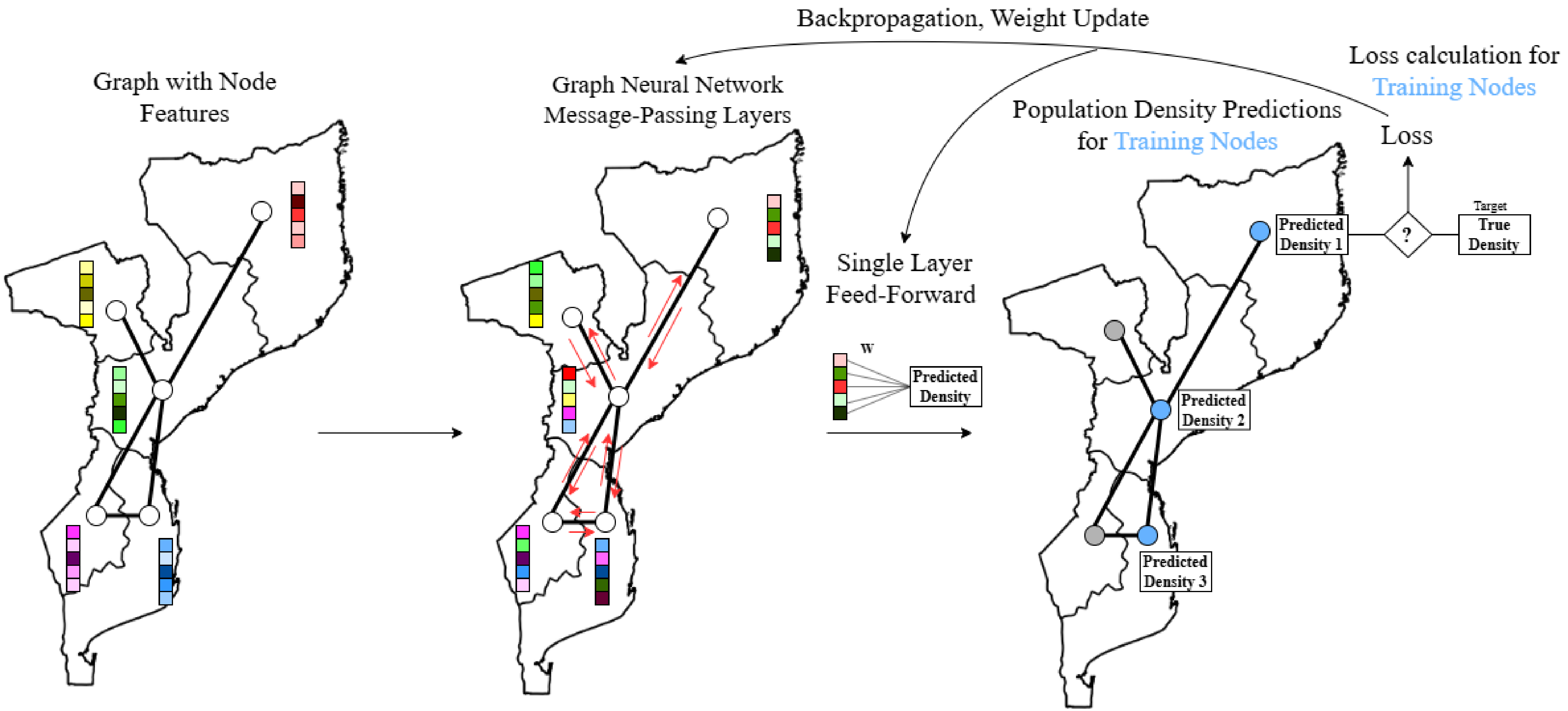
Graphs offer a natural representation of the geographical data.

Mozambique’s
Administrative level 1 areas

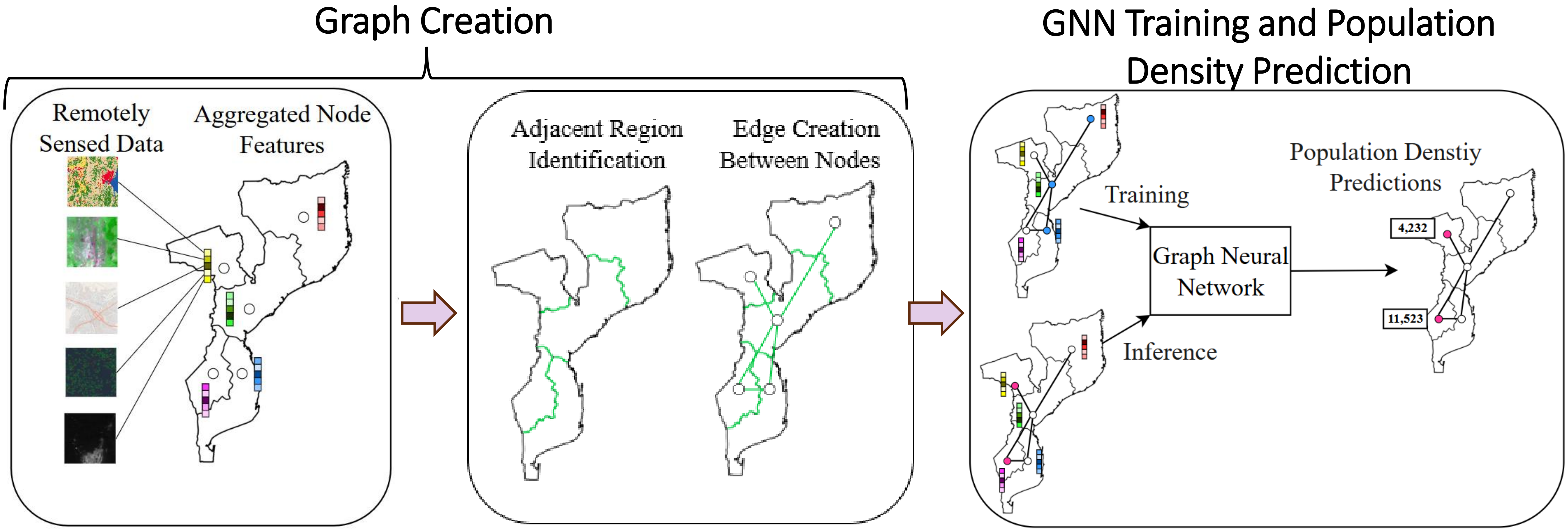


Extract & Aggregate EO data with GeoFlo! Visit Repo

Using graphs with a Graph Neural Network (GNNs) to incorporate spatial context



*To **predict population density** across administrative areas in Mozambique, using **satellite-based features** and **spatial relationships** with a GNN.*



No method wins everywhere…
Why do GNNs win in some countries, not others?

12 / 28
countries where a GNN method gave the best estimate

Ran several test to understand why GNNs outperformed baseline, and vice versa.

None of these tests separated the GNN winners from the losers — **when GNNs help remains an open question**.

Graphs work for geographic data… in data-rich settings

Google Research

2025-01-29

General Geospatial Inference with a Population Dynamics Foundation Model

Mohit Agarwal[◦,1], Mimi Sun[◦,1], Chaitanya Kamath[1], Arbaaz Muslim[1], Prithul Sarker[1,2], Joydeep Paul[1], Hector Yee[1], Marcin Sieniek[1], Kim Jablonski[1], Swapnil Vispute[1], Atul Kumar[1], Yael Mayer[1], David Fork[1], Sheila de Guia[1], Jamie McPike[1], Adam Boulanger[1], Tomer Shekel[1], David Schottlander[1], Yao Xiao[1], Manjit Chakravarthy Manukonda[1], Yun Liu[1], Neslihan Bulut[1], Sami Abu-el-haija[1], Bryan Perozzi[1], Monica Bharel[1], Von Nguyen[1], Luke Barrington[1], Niv Efron[1], Yossi Matias[1], Greg Corrado[1], Krish Eswaran[3], Shruthi Prabhakara[1], Shravya Shetty[1] and Gautam Prasad[1]
[◦]equal contribution, [1]Google Research, [2]University of Nevada, Reno, [3]PURI AI

Supporting the health and well-being of dynamic populations around the world requires governmental agencies, organizations, and researchers to understand and reason over complex relationships between human behavior and local contexts. This support includes identifying populations at elevated risk and gauging where to target limited aid resources. Traditional approaches to these classes of problems often entail developing manually curated, task-specific features and models to represent human behavior and the natural and built environment, which can be challenging to adapt to new, or even related tasks. To address this, we introduce the Population Dynamics Foundation Model (PDFM), which aims to capture the relationships between diverse data modalities and is applicable to a broad range of geospatial tasks. We first construct a geo-indexed dataset for postal codes and counties across the United States, capturing rich aggregated information on human behavior from maps, busyness, and aggregated search trends, and environmental factors such as weather and air quality. We then model this data and the complex relationships between locations using a graph neural network, producing embeddings that can be adapted to a wide range of downstream tasks using relatively simple models. We evaluate the effectiveness of our approach by benchmarking it on 27 downstream tasks spanning three distinct domains: health indicators, socioeconomic factors, and environmental measurements. The approach achieves state-of-the-art performance on geospatial interpolation across all tasks, surpassing existing satellite and geotagged image based location encoders. In addition, it achieves state-of-the-art performance in extrapolation and super-resolution for 25 of the 27 tasks. We also show that the PDFM can be combined with a state-of-the-art forecasting foundation model, TimesFM, to predict unemployment and poverty, achieving performance that surpasses fully supervised forecasting. The full set of embeddings and sample code are publicly available for researchers. In conclusion, we have demonstrated a general purpose approach to geospatial modeling tasks critical to understanding population dynamics by leveraging a rich set of complementary globally available datasets that can be readily adapted to previously unseen machine learning tasks.

US postcodes · proprietary behavioural + search data

A graph neural network for small-area estimation: integrating spatial regularisation, heterogeneous spatial units, and Bayesian inference

Pengyuan Liu[a,b], Yang Chen[c,d], Xiucheng Liang[c], Hao Li[e], Filip Biljecki[c,f] and Rudi Stouffs[c]

[a]Urban Analytics Subject Group, Urban Studies & Social Policy Division, University of Glasgow, Glasgow, UK; [b]Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, Glasgow, UK; [c]Department of Architecture, National University of Singapore, Singapore, Singapore; [d]School of Geography, Nanjing Normal University, Nanjing, China; [e]Department of Geography, National University of Singapore, Singapore, Singapore; [f]Department of Real Estate, National University of Singapore, Singapore, Singapore

ABSTRACT
Fine-resolution spatial analytics are essential for urban planning and policy-making, yet traditional small-area estimation often struggles with sparse, hierarchical, or imbalanced data. This paper introduces a Spatially Regularised Bayesian Heterogeneous Graph Neural Network (SR-BHGNN) that integrates multiple census tract levels within a unified framework. The model builds a heterogeneous graph where nodes represent spatial units at different scales, edges encode adjacency or membership, and Bayesian inference quantifies uncertainty in parameters and predictions. A spatial regularisation term, inspired by Tobler's First Law of Geography, penalises large discrepancies between neighbouring nodes, reducing errors in imbalanced datasets and ensuring coherent local estimates. We evaluate SR-BHGNN through two London case studies, population estimation and $PM_{2.5}$ prediction, comparing it against random forests, single-level GNNs, and spatial hierarchical Bayesian estimation. SR-BHGNN achieves strong performance gains, with classification accuracies of 0.85 for population estimation and 0.81 for $PM_{2.5}$ prediction. Its Bayesian design produces posterior distributions that capture uncertainty, enabling policy-relevant insights into vulnerable neighbourhoods or priority intervention zones (e.g. low-emission areas). These results demonstrate that SR-BHGNN advances the state of the art in small-area estimation, offering a flexible, uncertainty-aware framework for diverse urban analytics applications.

ARTICLE HISTORY
Received 7 August 2025
Accepted 27 November 2025

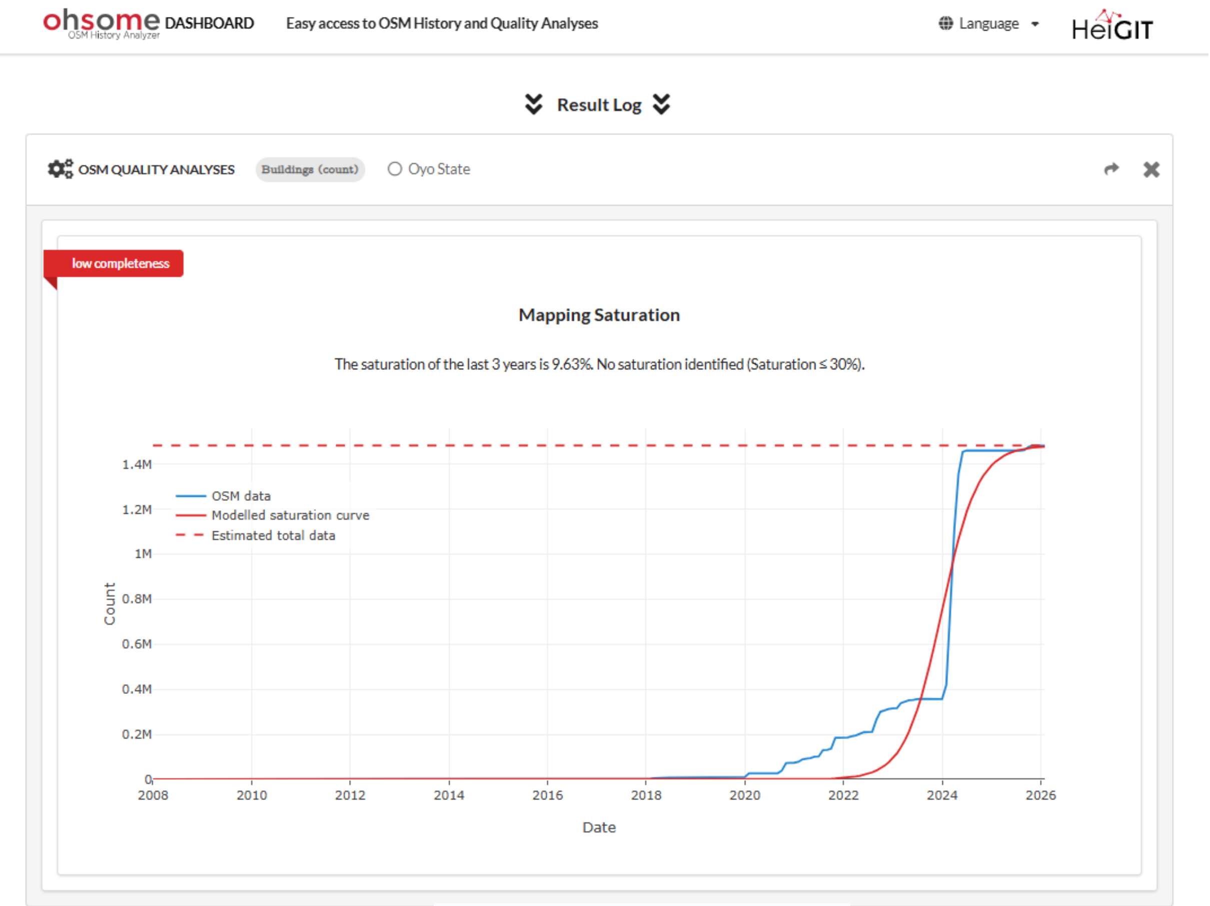
KEYWORDS
Small-area estimation; spatial regularisation; uncertainty quantification; Tobler's first law

single city · census + street-view · *explicit spatial regularisation*

Our Setting:

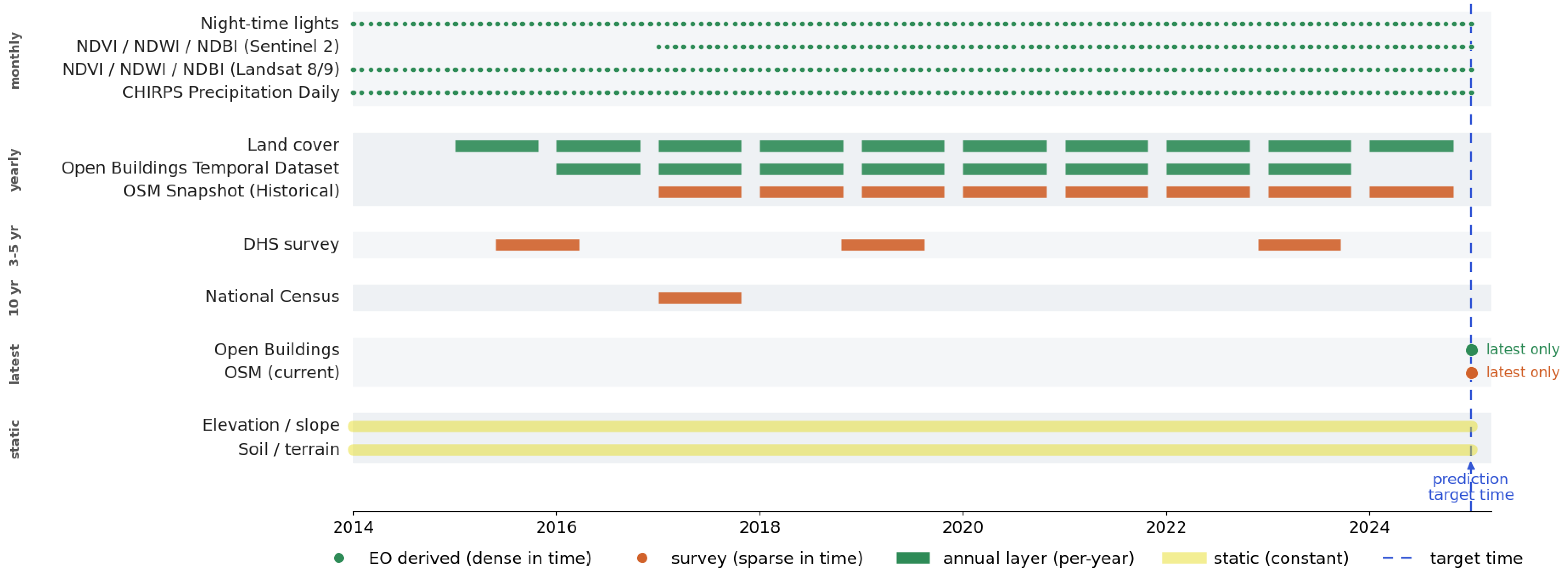
- 28 sub-Saharan African countries
- open EO + OSM only → sparse and noisy.

Issue #1: Data Quality / Missingness Problem

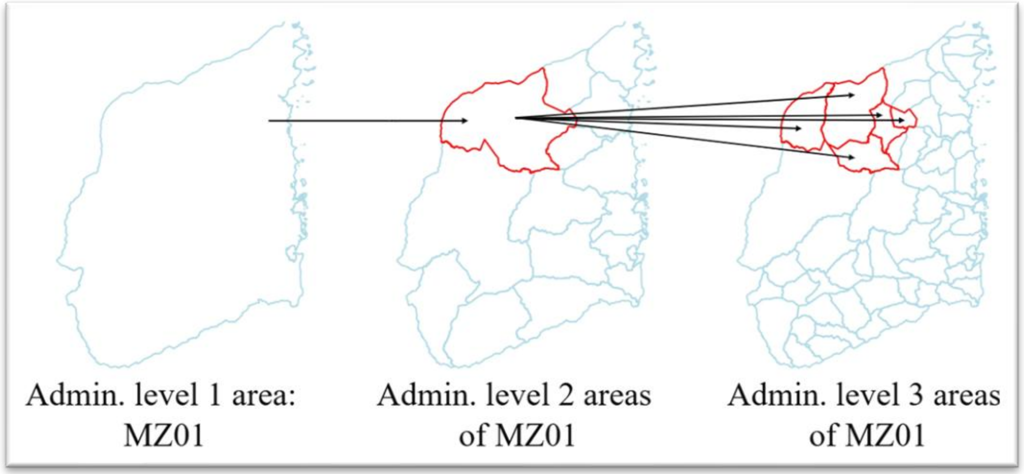


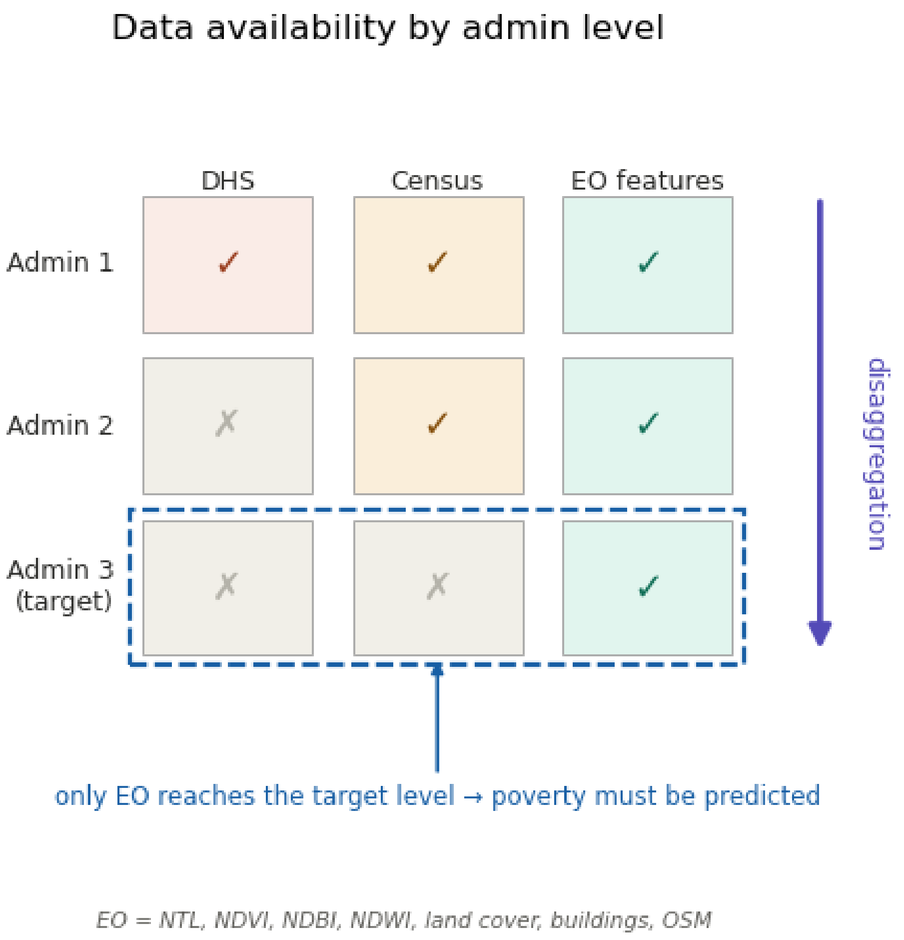
HeiGIT HEIDELBERG INSTITUTE FOR GEOINFORMATION TECHNOLOGY

Issue #2: Data availability across **time**

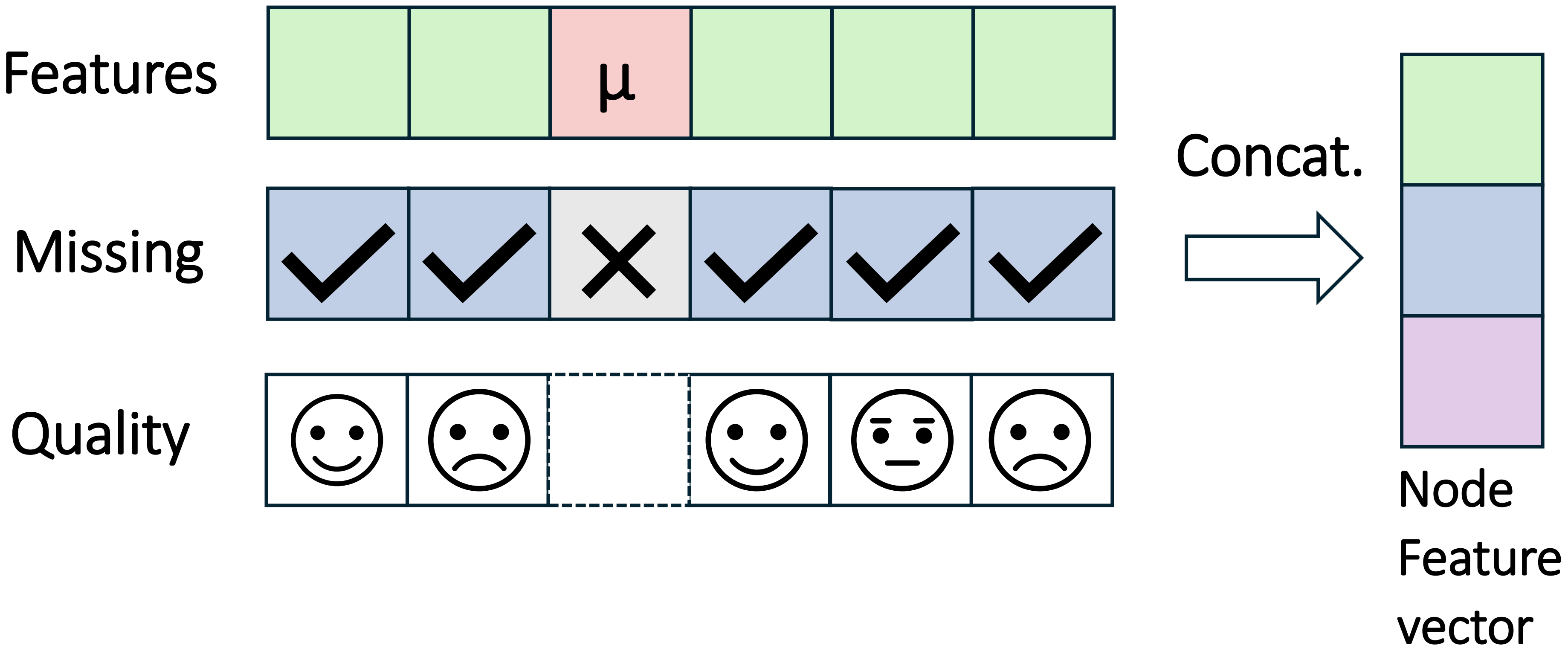


Issue #2: Data availability across **hierarchies**

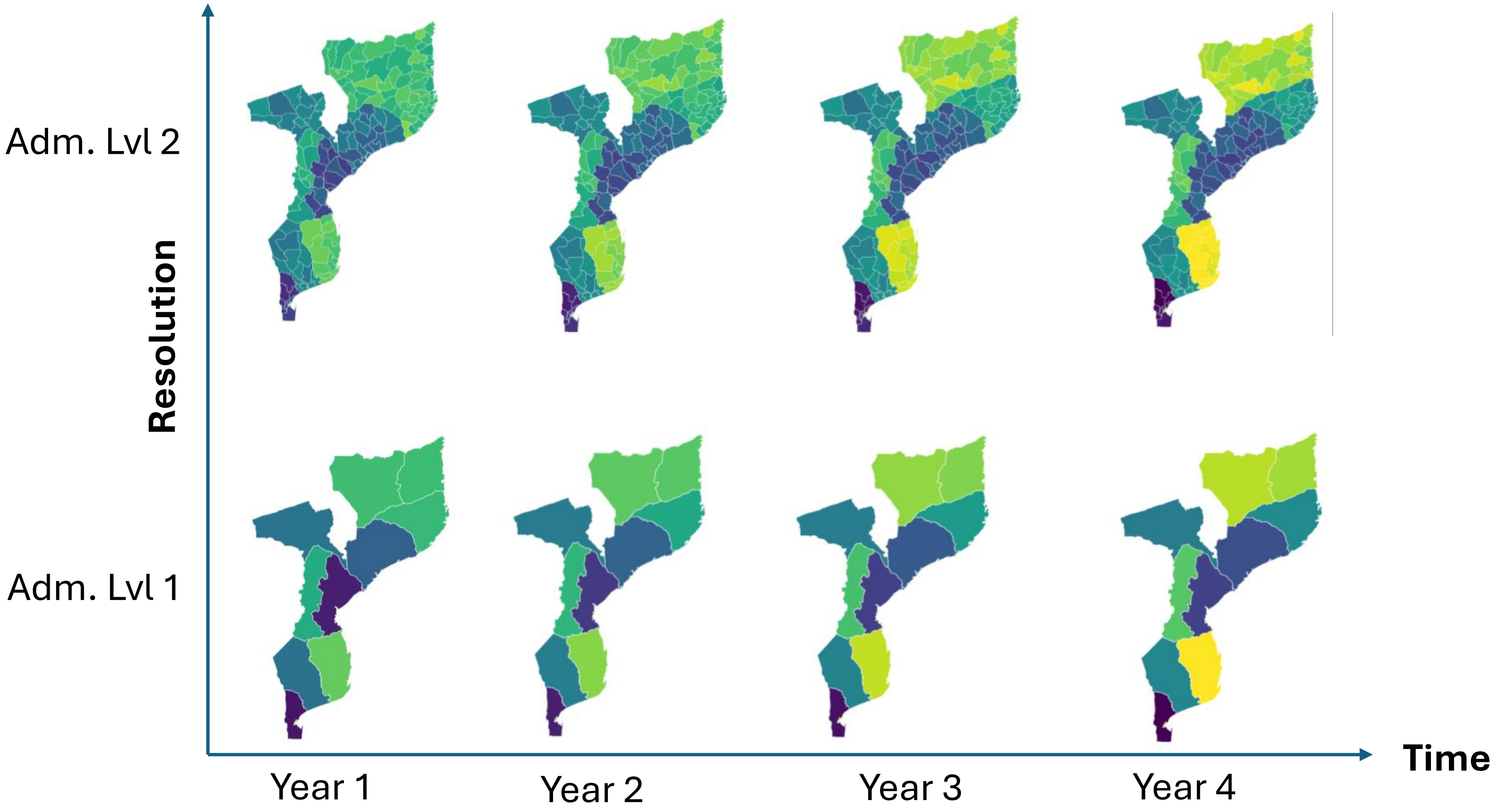




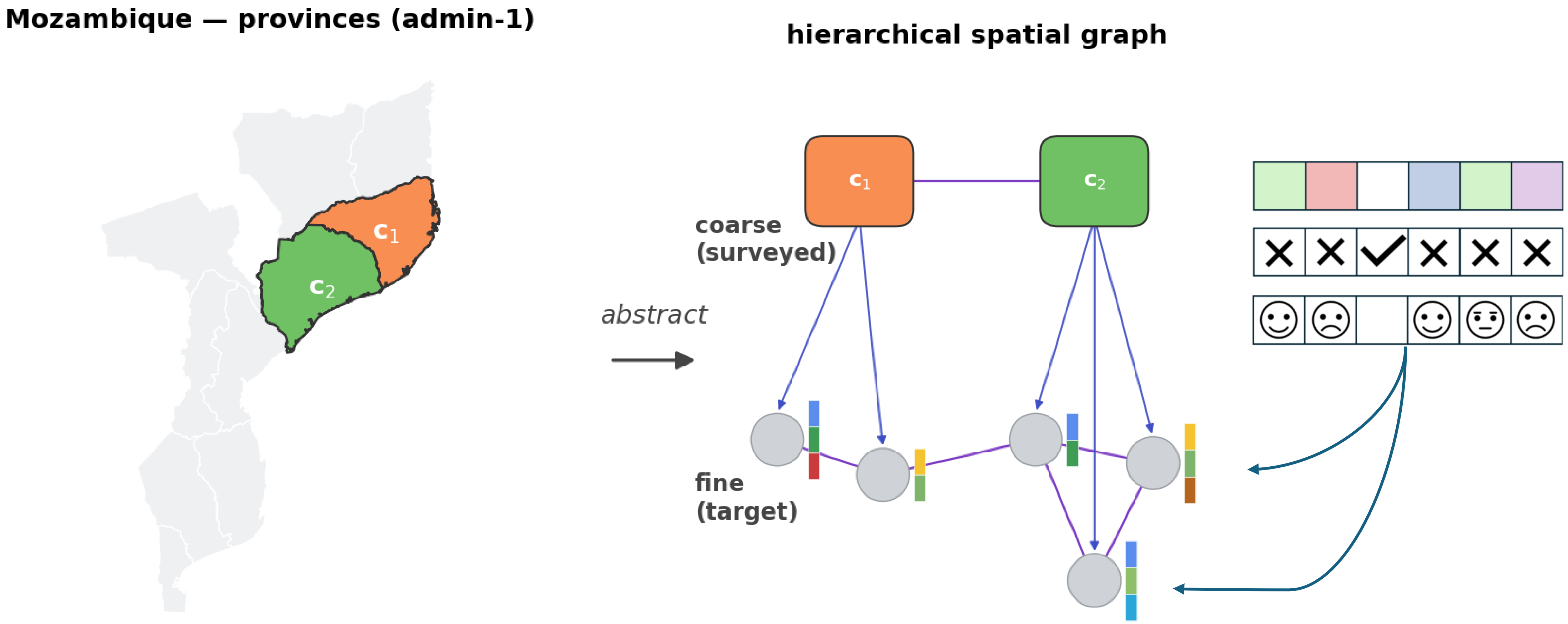
Fix #1: Turn data problems into extra signal the model can learn from



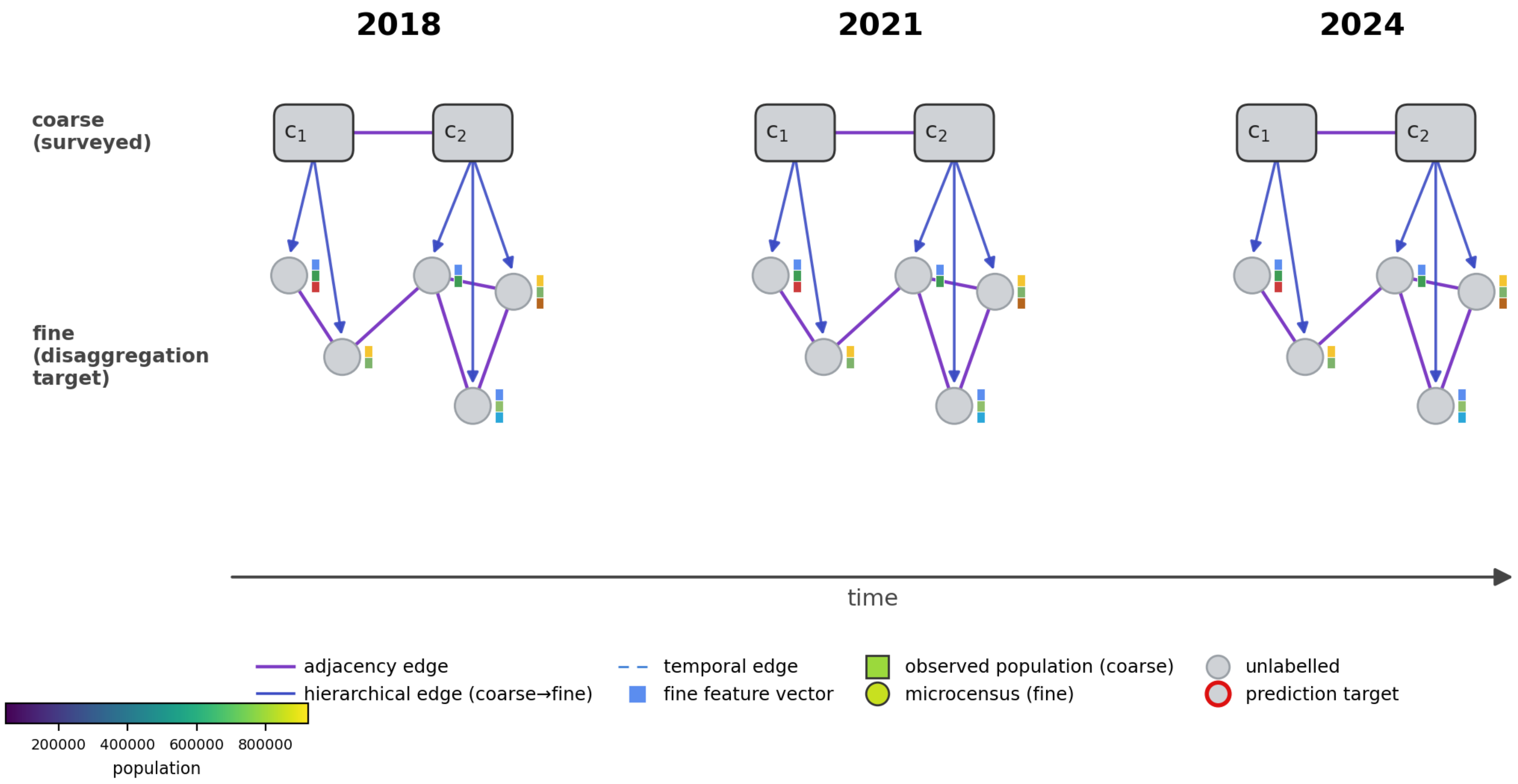
Fix #2: Using both Spatial & Temporal Context



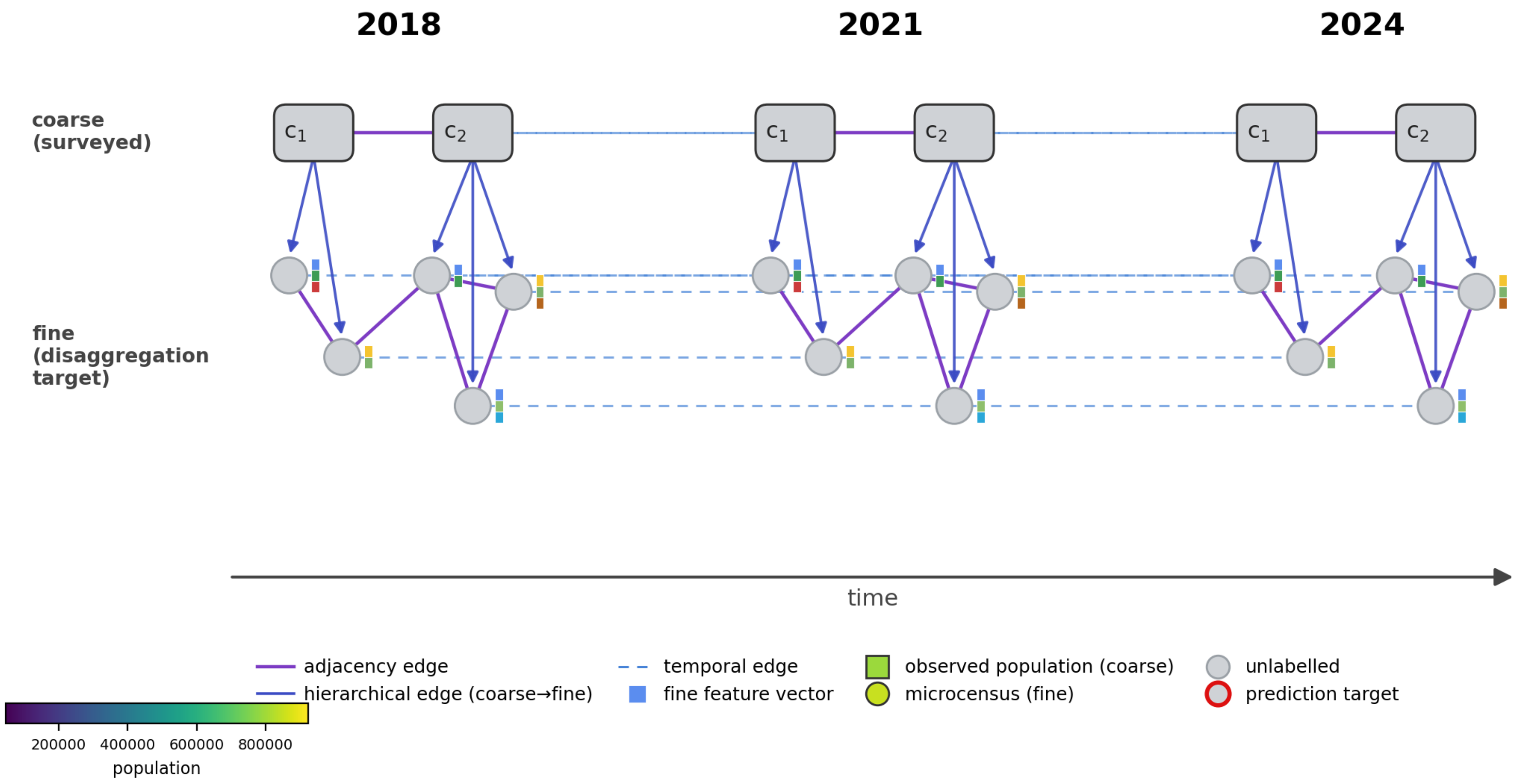
Hierarchical Spatial Graph



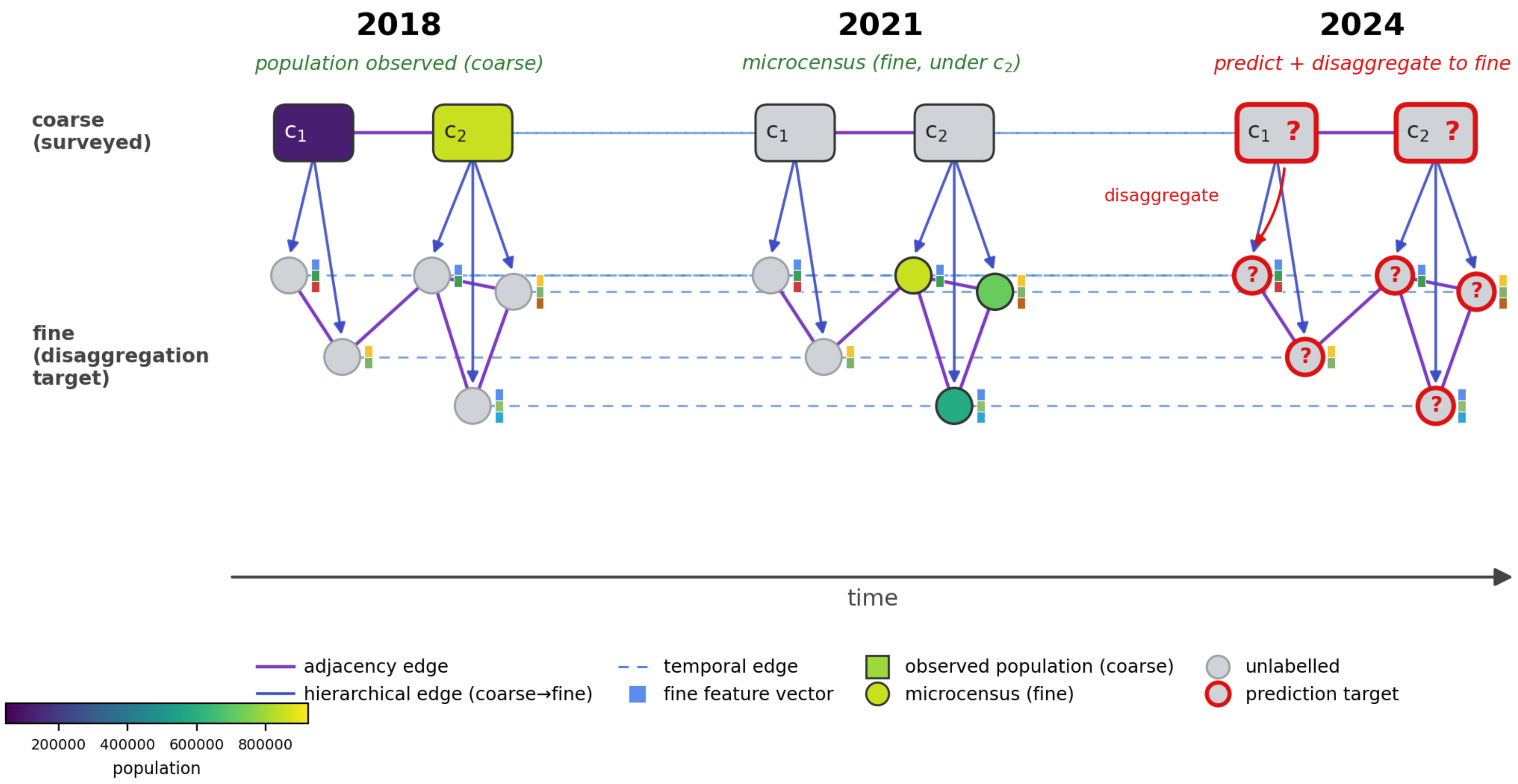
Spatial graphs at multiple points in time



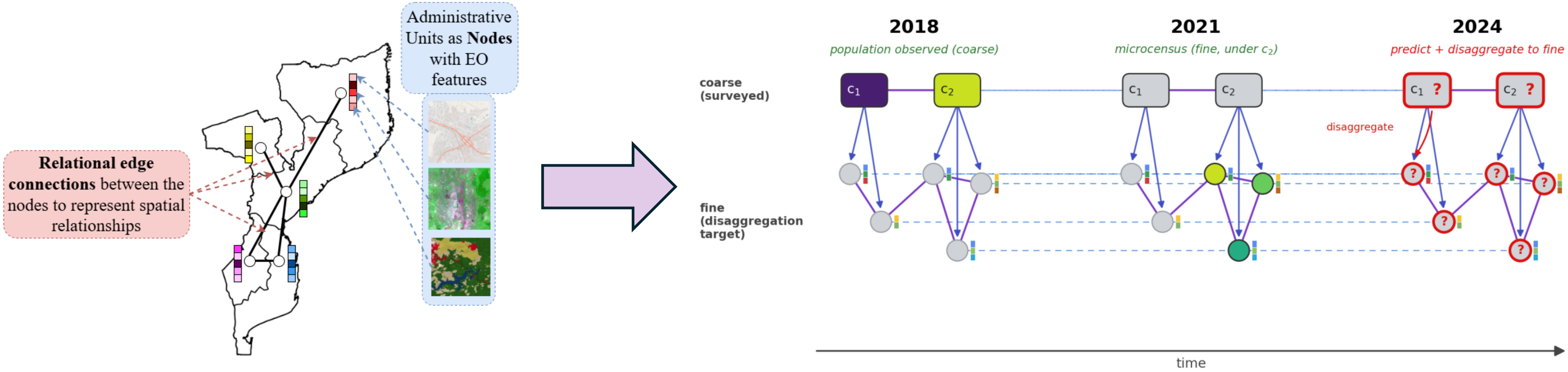
Spatio-temporal graph~~s at multiple points in time~~



Spatio-temporal graph with population labels



Summary



- Using ALL available data for a country across a large time
- Using data quality and data missingness issues with the features
- Turning out-of-date coarse census population information into finer-grained estimates with both **spatial & temporal context**

GeoFlo Repository