Monitoring Food Insecurity using population non-representative survey

Sahoko Ishida

Department of Computer Science

University of Oxford

UNIVERSITY OF OXFORD

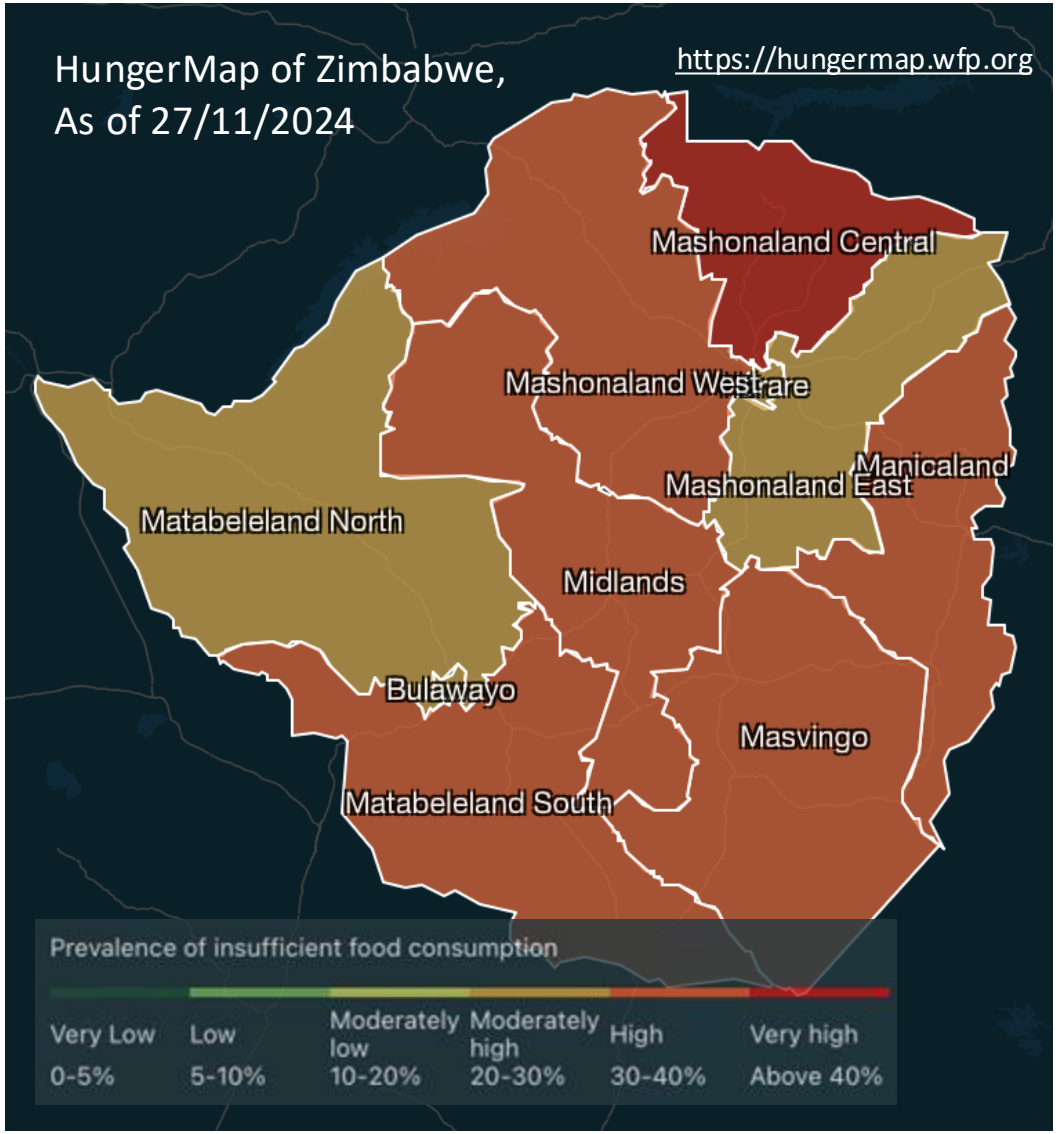
MACHINE LEARNING & GLOBAL HEALTH NETWORK

TIES 2024

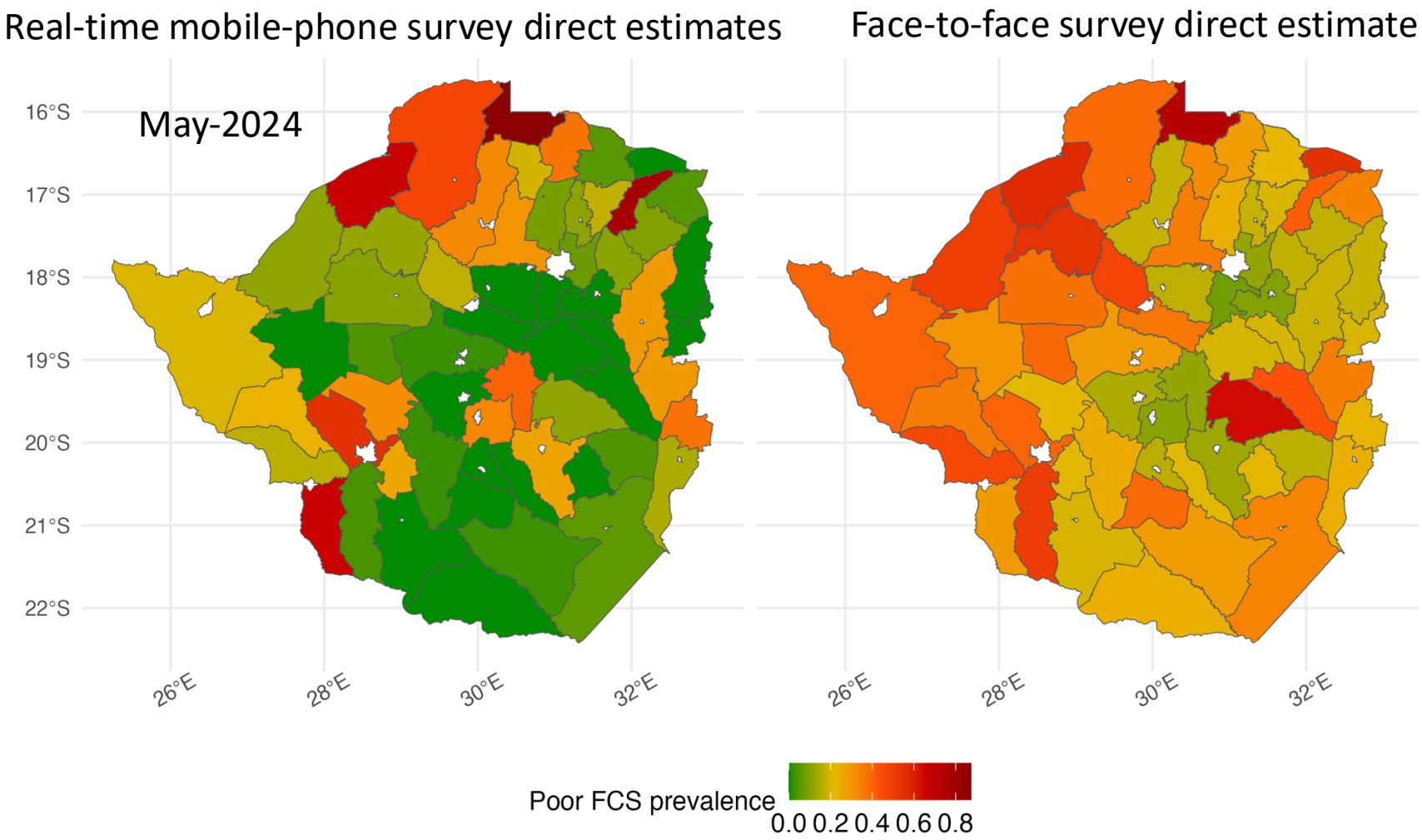
- **Dr. Adam Howes**, **Dr.Elizaveta Semenova** (Imperial College London)
- **Dr. Valerie Bradley**, **Prof. Seth Flaxman** (University of Oxford)
- **Prof. Dino Sejdinovic** (University of Adelaide)
- **WFP Rome** Hunger Monitoring Unit, Forecasting and Early Warning System
- **WFP Zimbabwe**
- And many more!

SMALL AREA DISAGGREGATION OF ZIMBABWE HUNGERMAP DATA

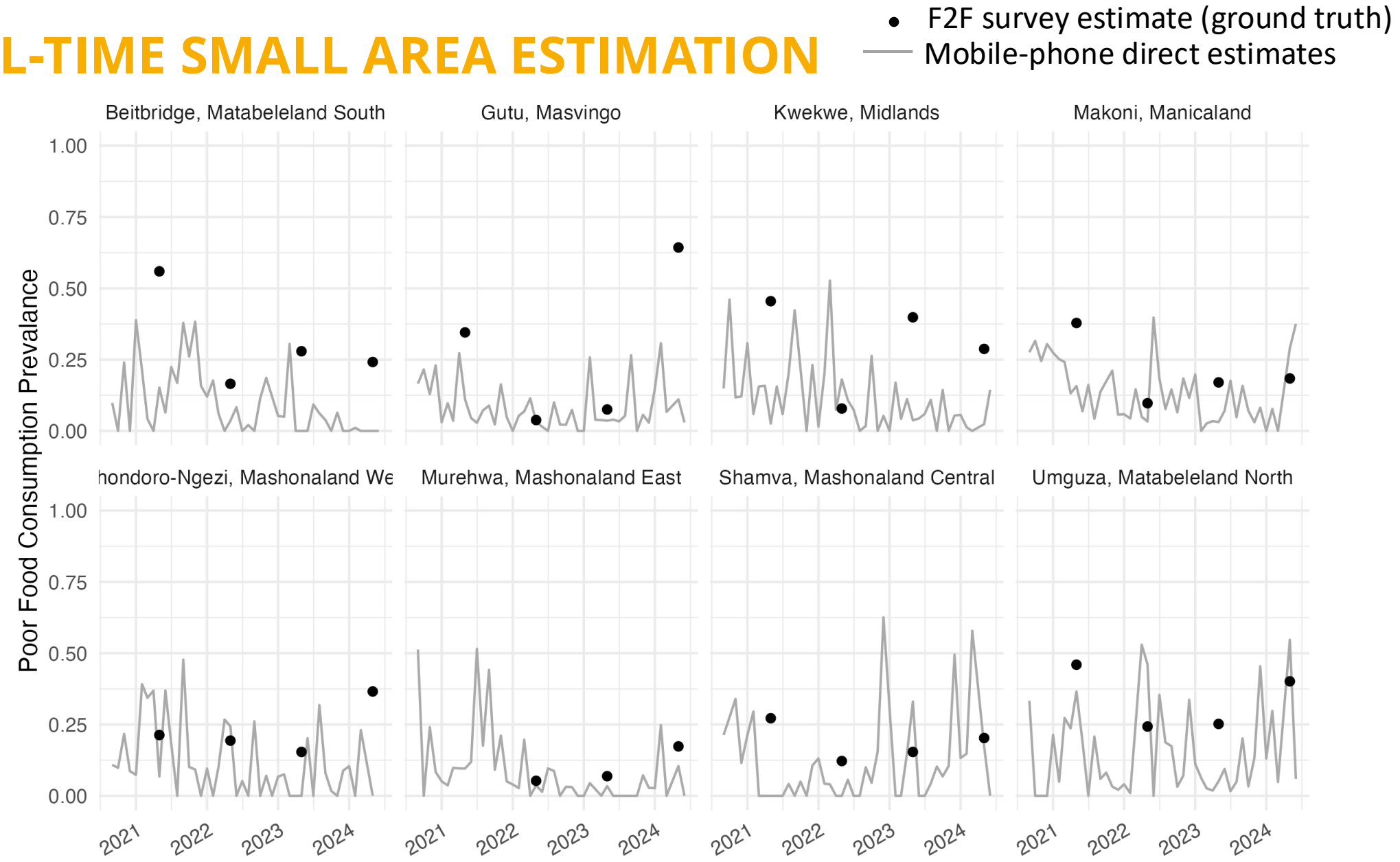
- **HungerMap:** WFP's main real-time monitoring survey and tool
  - Mobile phone survey, running in many countries Africa, South America, South (East) Asia
  - Collects data on household food consumption
  - n ≈ 1000 to ensure monthly representativeness at the **province** level
- **Challenges**
  - Decision making often made at the **district** level
  - Phone surveys prone to biases: sampling bias, modality bias
  - Face-to-face (F2F) survey representative at district level too costly to run frequently (once a year in Zimbabwe's case)



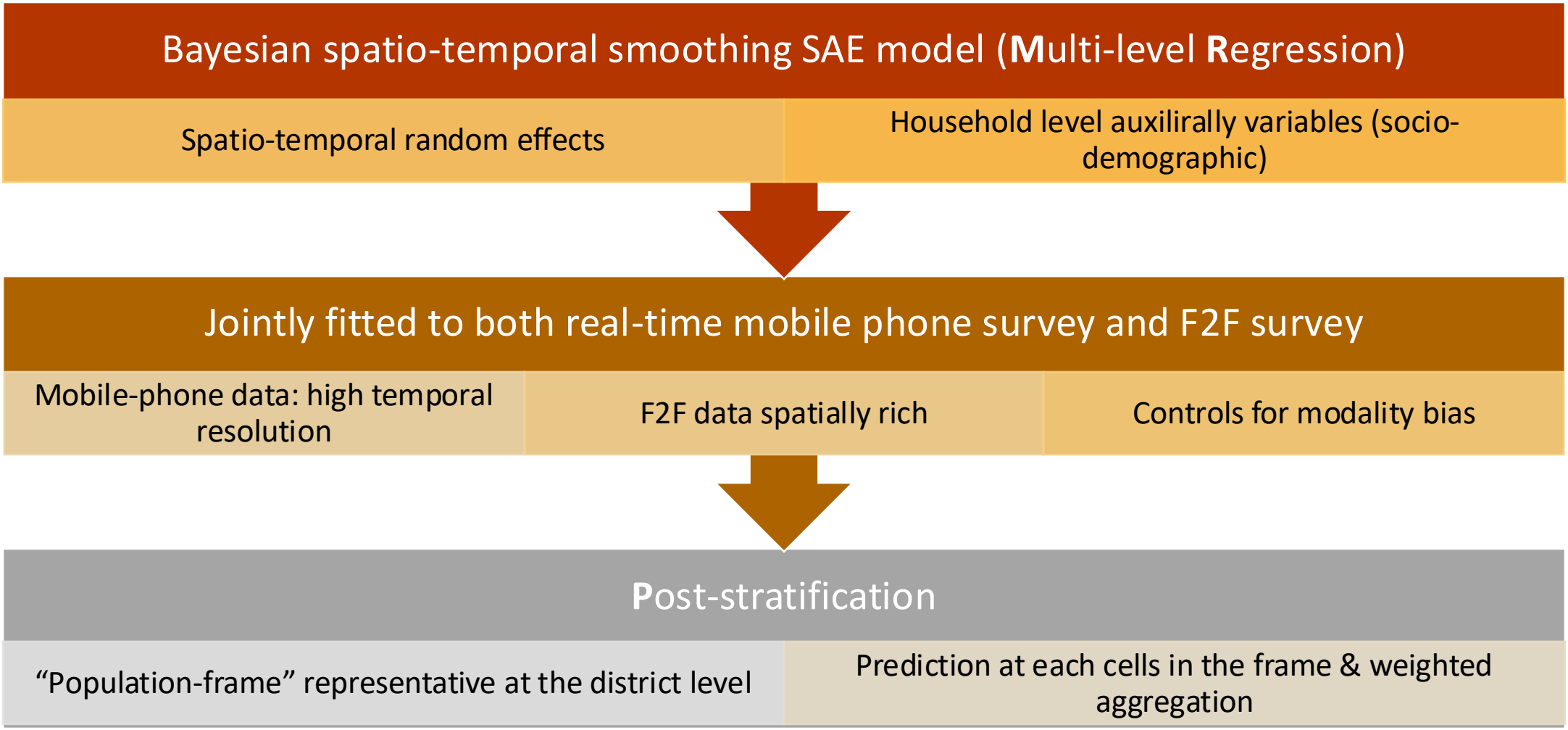
AIMS – REAL-TIME SMALL AREA ESTIMATION



REAL-TIME SMALL AREA ESTIMATION



OUR APPROACH TO (REAL-TIME) SMALL AREA ESTIMATION



MRP [Park et al. (2004)]

Bayesian spatio-temporal smoothing at household level

Modelling $p = p(Food\ Consumption\ Score(FCS) < c)$ at household level

$$logit(p) = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \boldsymbol{\theta_s} + \boldsymbol{\nu_t} + \boldsymbol{\psi_{rt}}$$

- Main spatial random effects: $\boldsymbol{\theta_s}$ for district $s = 1, \ldots S$

  BYM, BYM2, GP
- Main temporal random effects: $\boldsymbol{\nu_t}$ for time $t = 1, \ldots T$

  Combination of Random Walk of order 1 and i.i.d normal
- Spatio-temporal **interaction** random effects: $\boldsymbol{\psi_{rt}}$for time$t = 1, \ldots T$ and province $r = 1, \ldots R$

  i.i.d normal, can be replaced with tensor product of spatially and temporally structured prior

Bayesian spatio-temporal smoothing at household level

Modelling $p = p(Food\ Consumption\ Score(FCS) < c)$ at household level

$$logit(p) = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \boldsymbol{\theta_s} + \boldsymbol{\nu_t} + \boldsymbol{\psi_{rt}}$$

- 6 household-level covariates from survey - all categorical

  Head of Household education, sex / Household size / access to safe water sources / toilet types / phone ownership

- Survey type (mobile-phone / F2F) - we let the effect of household-level covariates differs for different surveys

- Similar to a smoothing model commonly seen in model-based unit-level small area estimation

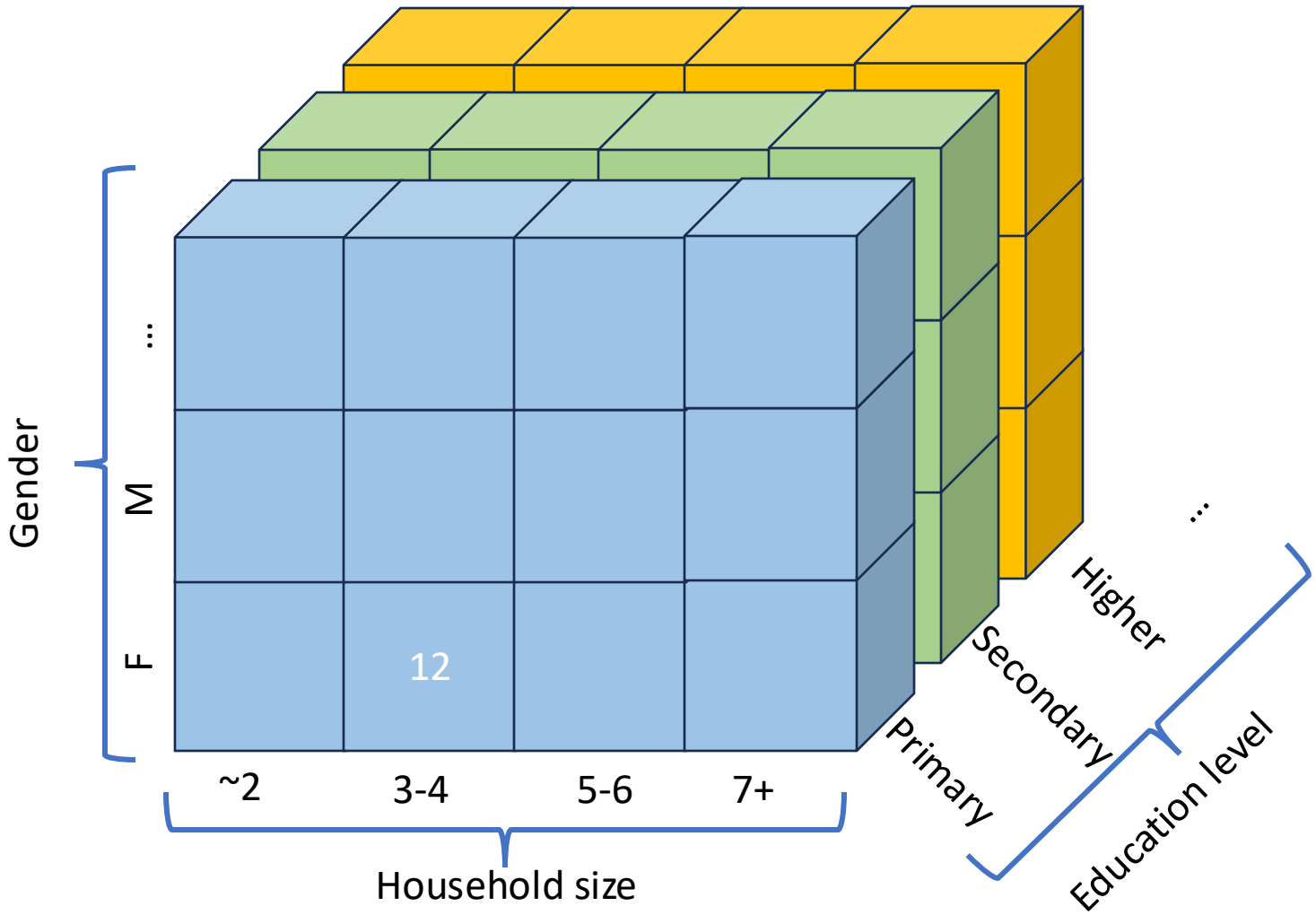
Post-stratification

- Construct population frame (tensor) at each district $\boldsymbol{s}$
- Each cell $\boldsymbol{j}$ reflecting the relative frequency n_j^s , given a combination of auxiliary variables, e.g.

(Gender = F) X

(Household size = 3) X

(Education level = Primary)



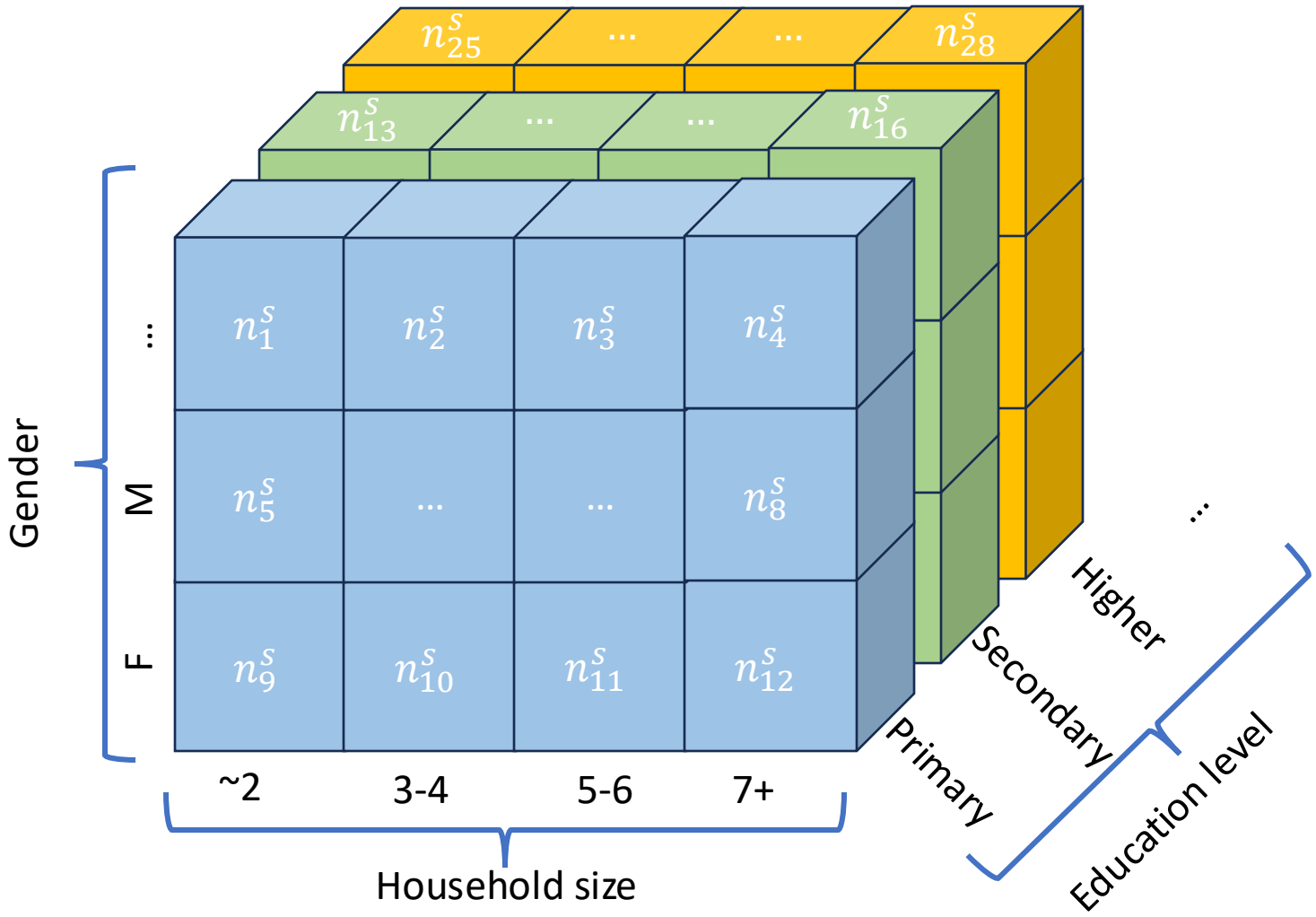
Post-stratification

- Construct population frame (multidimensional tensor) for each district $\boldsymbol{s}$
- Each cell $\boldsymbol{j}$ reflecting the relative frequency n_j^s of a specific subgroup in the population, e.g.,

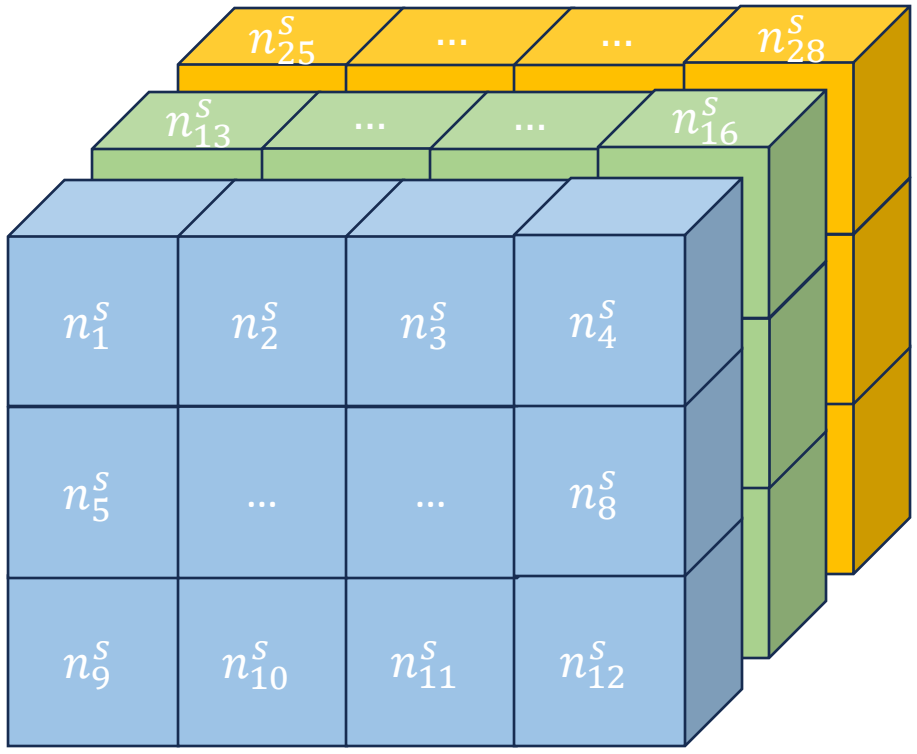
(Gender = F) X

(Household size = 3) X

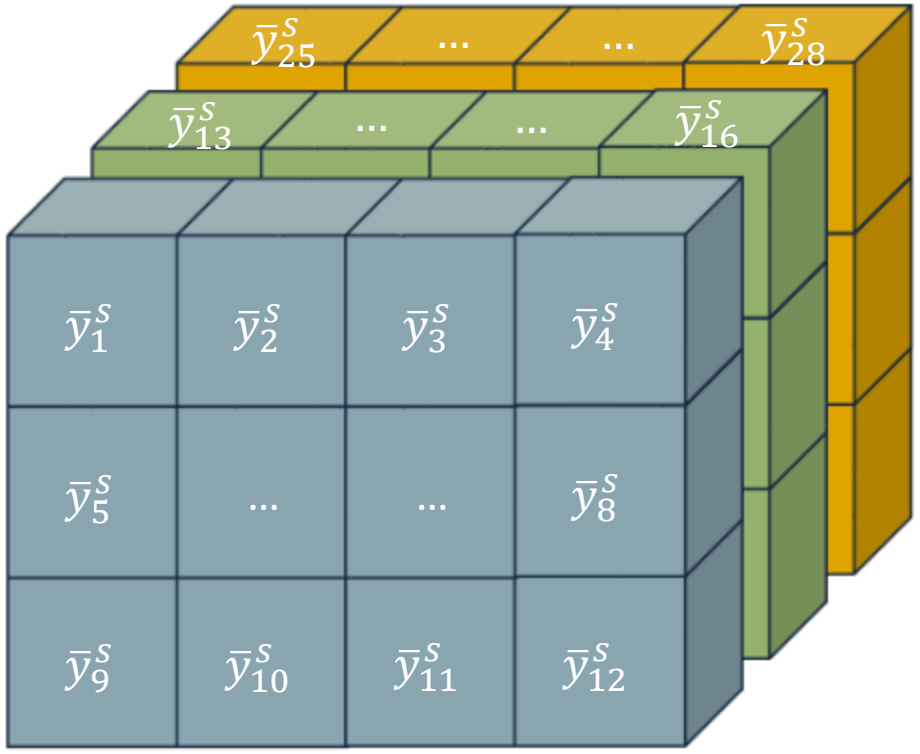
(Education level = Primary)



Population frame for district s



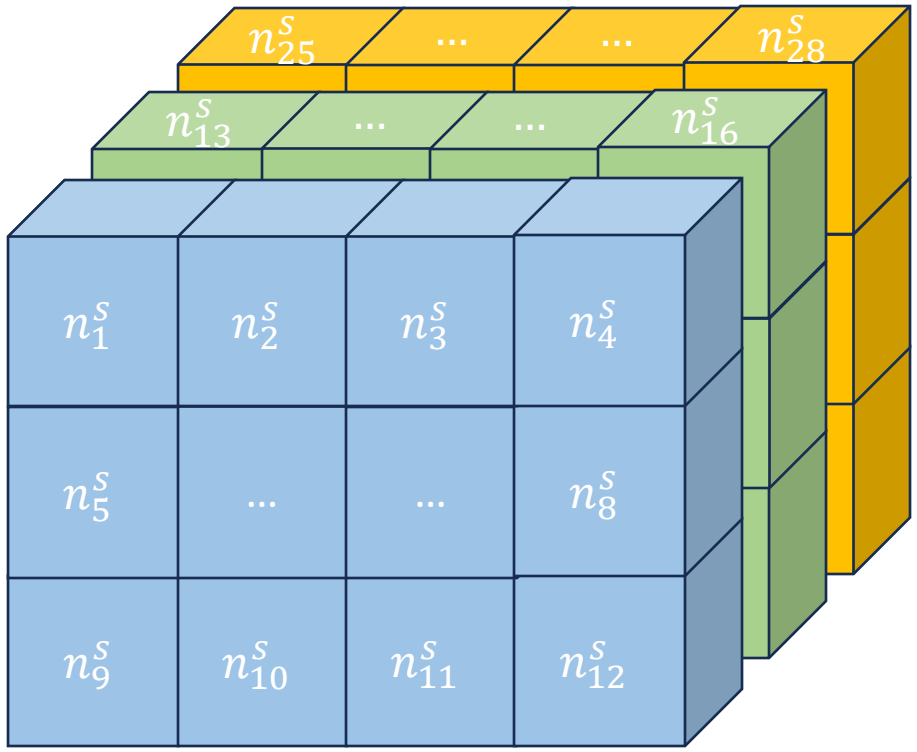
A frame of sample mean for district s



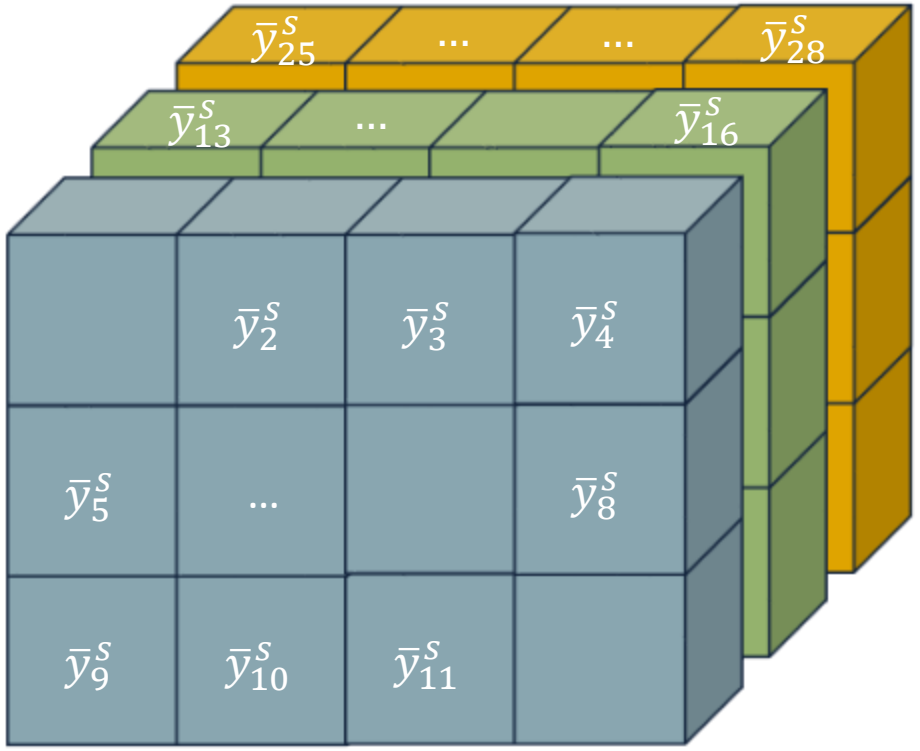
Estimates of prevalence for district s

$$\hat{p}_s = \frac{\sum_j n_j^s \bar{y}_j{}^s}{\sum_j n_j^s}$$

Population frame for district s



A frame of sample mean for district s

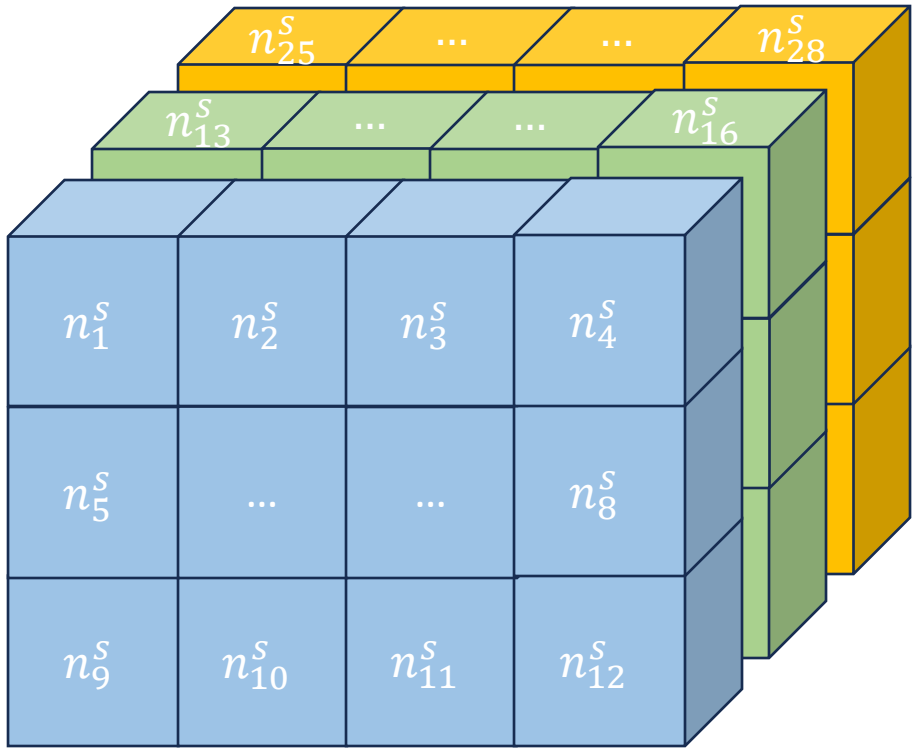


Estimates of prevalence for district s

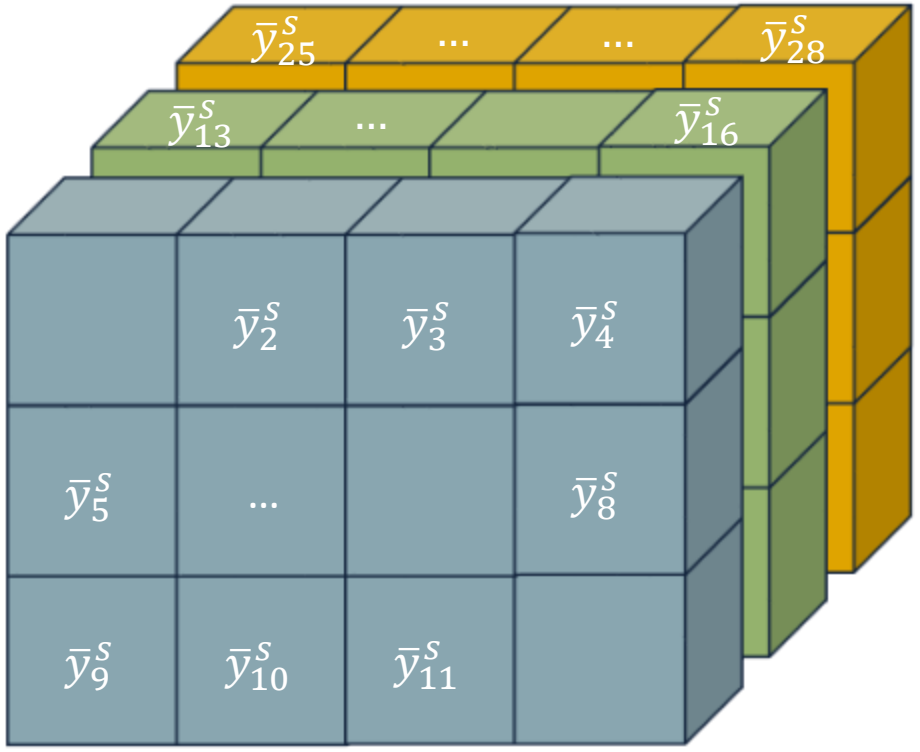
$$\hat{p}_s = \frac{\sum_j n_j^s \, \hat{p}_j^s}{\sum_j n_j^s}$$

- Number of cells could be thousands, hence some cells might have very few or no observations
- Use predicted values instead

Population frame for district s



A frame of sample mean for district s



- The weights can be calibrated to ensure that n_j^s is not too small/large
- If no survey/census available to get joint distribution, raking ratio estimation can be used. Only marginal totals/proportion required

Estimates of prevalence for district s

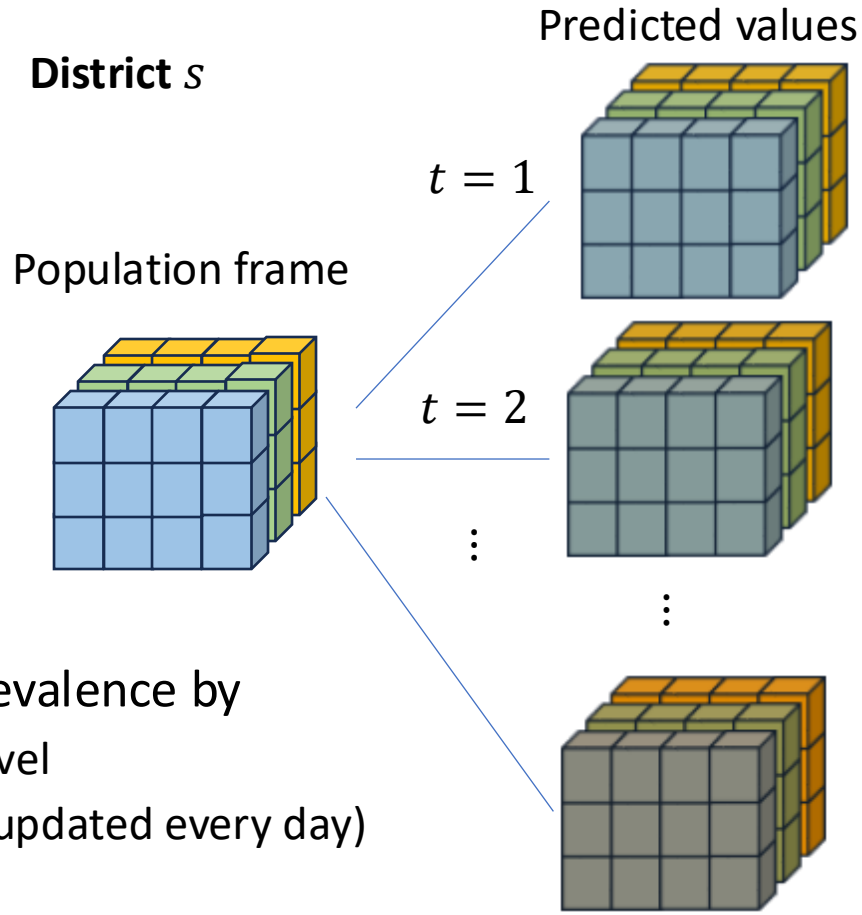
$$\hat{p}_s = \frac{\sum_j \hat{w}_j^s \, \hat{p}_j^s}{\sum_j \hat{w}_j^s}$$

Real-time monitoring

- We assume that the weights (set of w_j^s) do not change for the period of our study, and only the predicted values (set of $\hat{p}_j^s$) evolve over time

MRP-SAE for district s at time t

$$\hat{p}_{st} = \frac{\sum_j \hat{w}_j^s \hat{\boldsymbol{p}}_{\boldsymbol{j}}^{\boldsymbol{st}}}{\sum_j \hat{w}_j^s}$$



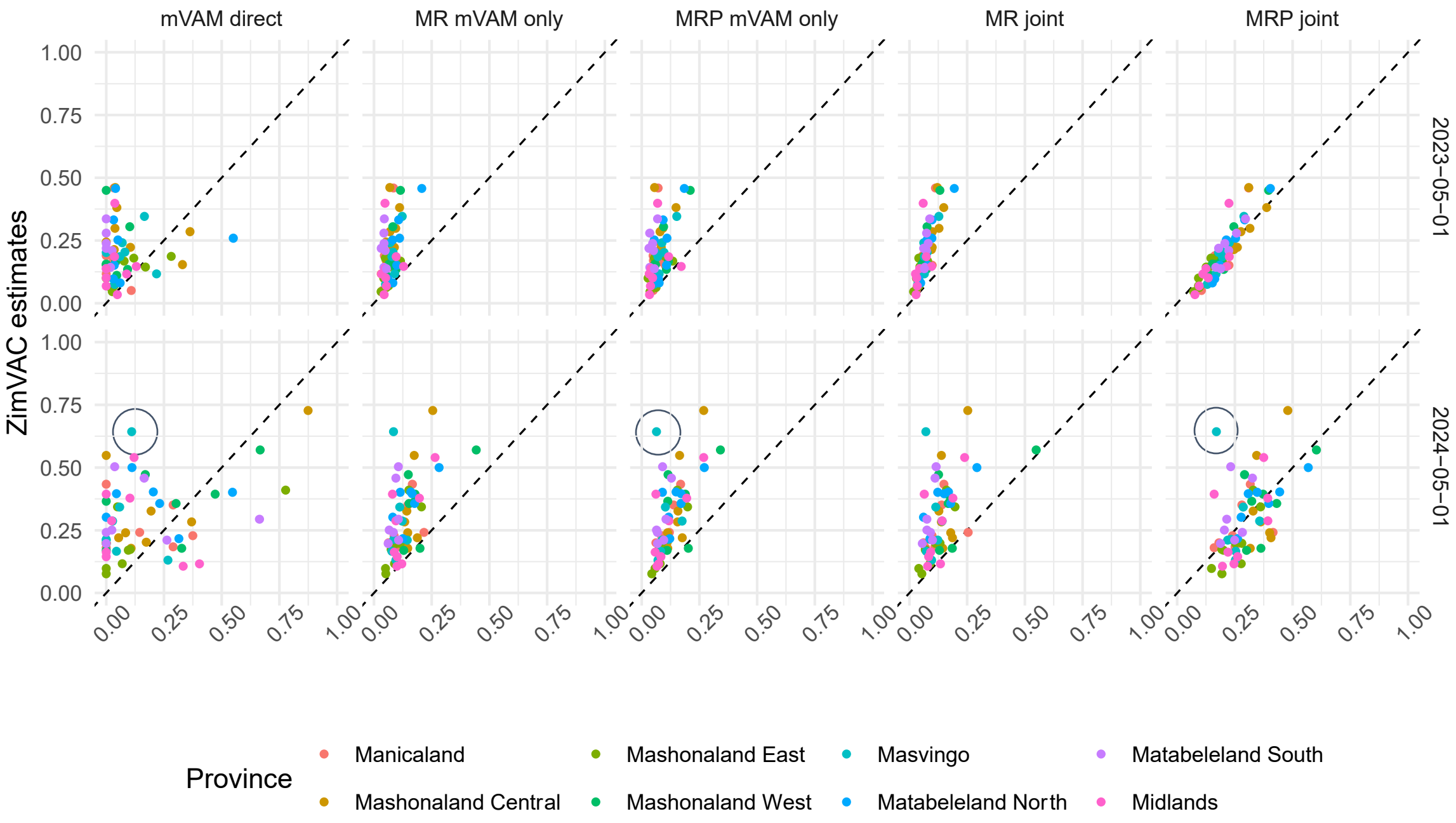
- Now-casting attempt, aiming to get reliable estimates of the prevalence by
  1. Using temporally sparse but representative at granular district level
  2. To correct for bias and variance of real-time phone survey data (updated every day)

Application - Zimbabwe

- 8 provinces and 90> districts – focuses on 61 rural districts
- Data source
  - Mobile Vulnerability Analysis and Mapping (mVAM) survey
    - The number of respondents is $n_r \approx 150$ per province per month
    - Our study period September 2020 – June 2024 (Ongoing),
    - Data collected daily - total of 43, 139 obserbations
  - Zimbabwe Vulnerability Assessment Committee (ZimVAC) rural survey
    - F2F survey conducted every annually (May)
    - The number of respondents for the month of study is $n_s \approx 250$ per district
    - We have access to 2023 and 2024 data at household level,
  - Zimbabwe 2015 Demographic and Health Survey (DHS) and Census 2022
    - Only needed for constructing population frame by estimating weights (raking)

Validation

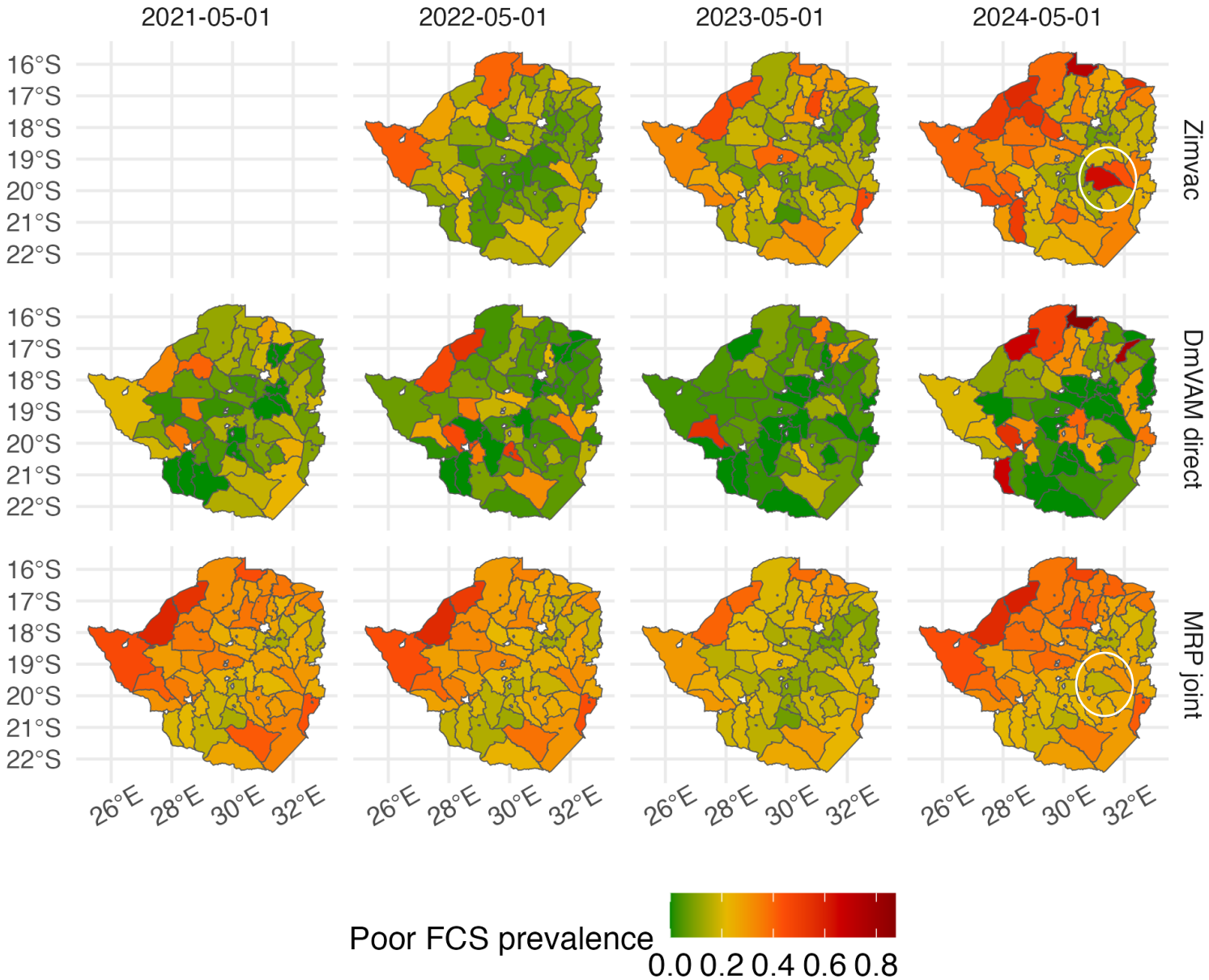
- Validating against ground truth – Zimvac (F2F) survey direct estimate
  - mVAM Direct Estimates (real-time mobile-phone survey)
  - mVAM survey smoothing model (no F2F survey data is included in model fitting)
    - MR estimates: Bayesian (multi-level) spatio-temporal smoothing + simple aggregation
    - MRP estimates: Bayesian (multi-level) spatio-temporal smoothing + post stratification
    - Poststratification weights are estimated using census2022 margins and dhs 2015
  - Joint model (Mobile-phone survey and 2023 F2F survey)
    - MR estimates
    - MRP estimates
    - Poststratification weights are computed from Zimvac 2023 survey
- In-sample error using the 2023 survey and out-of-sample error using the 2024 survey


mVAM direct
MR mVAM only
MRP mVAM only
MR joint
MRP joint
ZimVAC estimates
2023−05−01
2024−05−01
0.00
0.25
0.50
0.75
1.00
Province
Manicaland
Mashonaland Central
Mashonaland East
Mashonaland West
Masvingo
Matabeleland North
Matabeleland South
Midlands

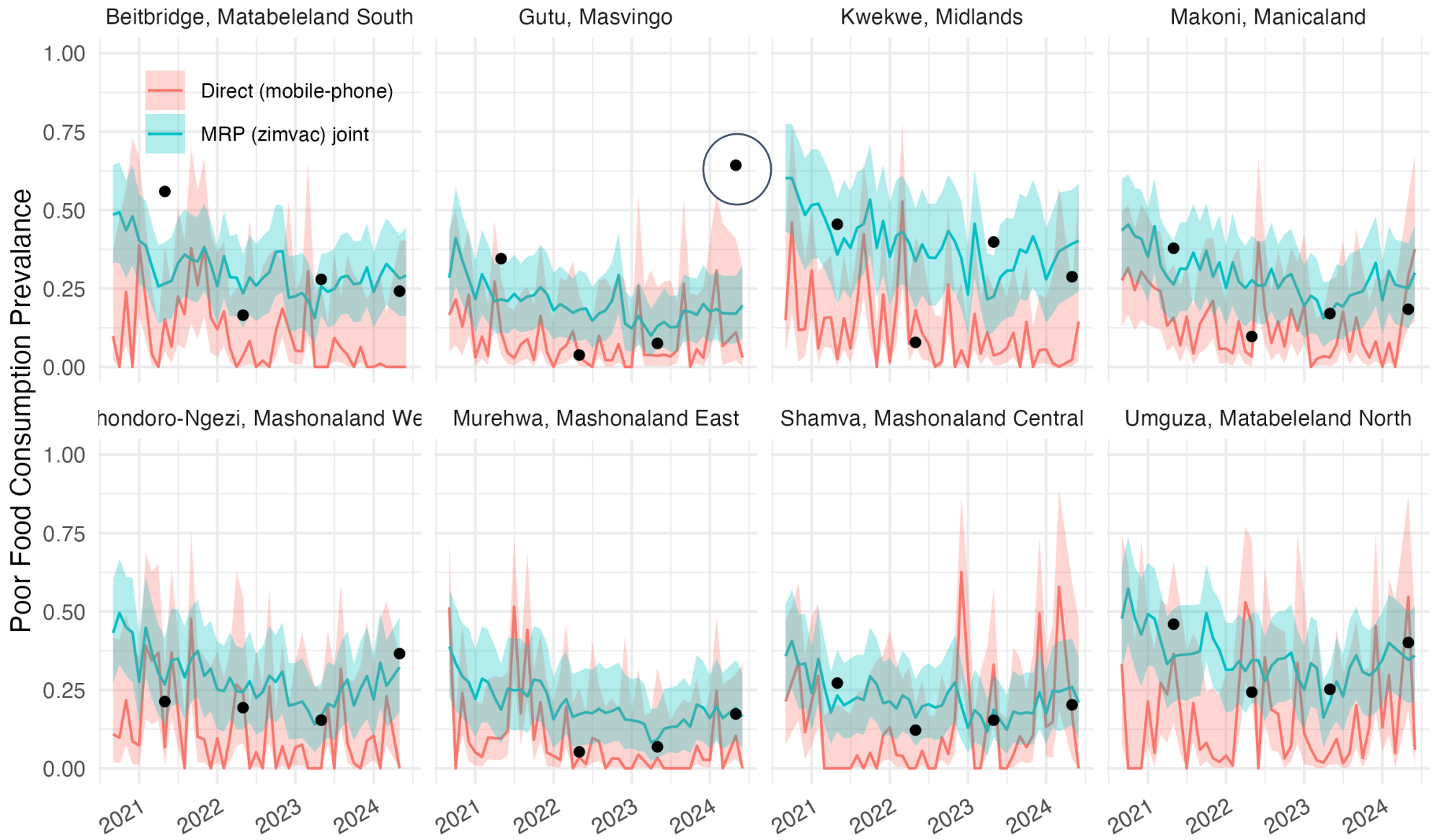
Validation

Year	Method	Pearson	CCC	MAE	Bias	Coverage90	CI_length90
2023	mVAM direct	0.051	0.028	0.159	-0.131	0.900	0.321
	MR mVAM only	0.480	0.110	0.127	-0.127	0.683	0.191
	MRP mVAM only	0.490	0.146	0.124	-0.123	0.583	0.142
	MR joint	0.827	0.203	0.128	-0.128	0.700	0.181
	MRP joint	**0.907**	**0.854**	**0.037**	**0.000**	0.983	0.195
2024	mVAM direct	0.333	0.248	0.210	-0.127	0.900	0.428
	MR mVAM only	**0.574**	0.212	0.162	-0.161	0.750	0.325
	MRP mVAM only	0.635	0.206	0.171	-0.171	0.467	0.184
	MR joint	0.519	0.215	0.163	-0.161	0.733	0.324
	MRP joint	0.525	0.484	**0.093**	**0.005**	0.933	0.279

District-level estimate

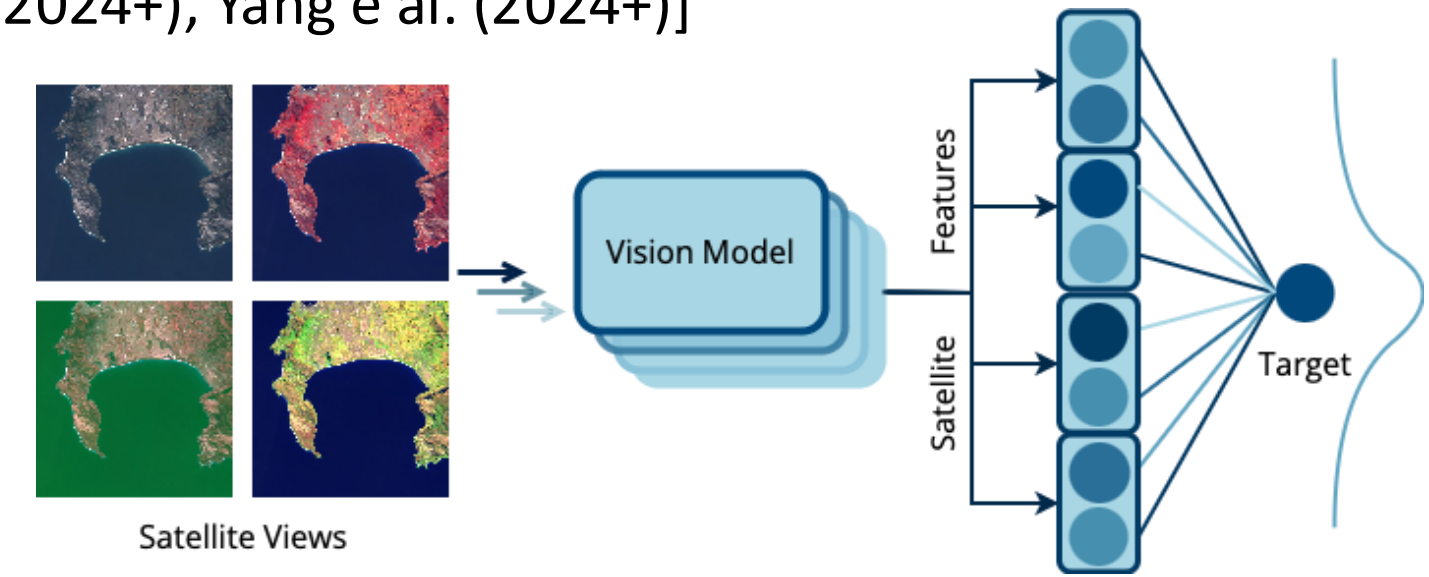


Now-casting


Beitbridge, Matabeleland South
Gutu, Masvingo
Kwekwe, Midlands
Makoni, Manicaland
hondoro-Ngezi, Mashonaland We
Murehwa, Mashonaland East
Shamva, Mashonaland Central
Umguza, Matabeleland North
Direct (mobile-phone)
MRP (zimvac) joint
Poor Food Consumption Prevalance
1.00
0.75
0.50
0.25
0.00
2021
2022
2023
2024

Future & ongoing work

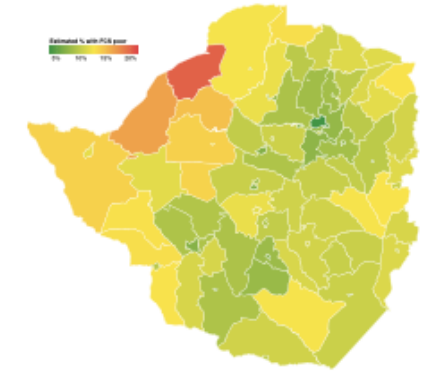
- Forecasting, higher time resolution (from monthly to weekly)
- Climate variables and economic, socio, and political variables
  - Readily available / processed data
  - Extracting richer representations from satellite imagery
    - Pre-trained computer vision models such as DINOv2
    - Applied in the context of estimating child poverty [Sharma et al.(2024+), Yang e al. (2024+)]

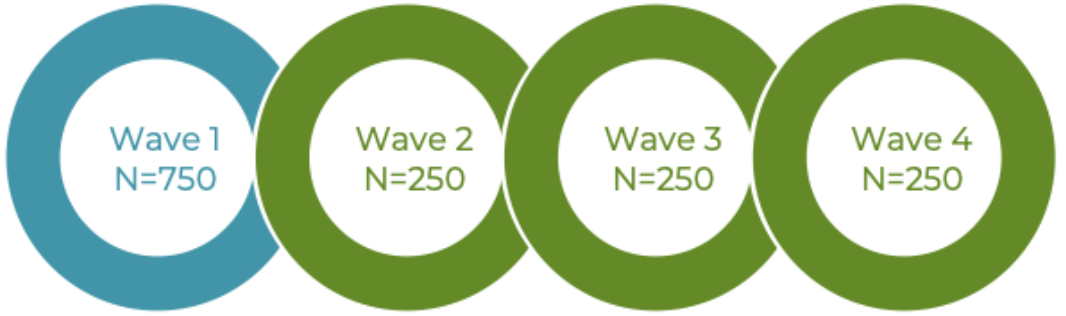


Future & ongoing work

- Adaptive learning survey design (A-LDS): new survey design using active learning technique to increase effective sample size of the real-time monitoring survey and better representation of marginalized population

ADAPTIVE LEARNING SAMPLING DESIGN (A-LSD) IN ZIMBABWE





1. Oct 18-Nov 14 2023: stratified random sample: province x water source x toilet type
2. Nov 28-Dec 4: active learning to prioritize random sample: stratify by urban/rural and gender
3. Dec 8-14: active learning to prioritize random sample: stratify by urban/rural and gender
4. Dec 18-24 2023: active learning to prioritize random sample: stratify by urban/rural and gender

Active learning objective: maximize representativeness of sample

Reference

Yang, F., Ishida, S., Zhang, M., Jenson, D., Mishra, S., Navott, J., & Flaxman, S. (2024). Uncertainty-Aware Regression for Socio-Economic Estimation via Multi-View Remote Sensing. arXiv preprint arXiv:2411.14119.

Sharma, M., Yang, F., Vo, D. N., Suel, E., Mishra, S., Bhatt, S., ... & Flaxman, S. (2024). KidSat: satellite imagery to map childhood poverty dataset and benchmark. arXiv preprint arXiv:2407.05986.): 95-115.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. Political Analysis, 12(4), 375-385.

Zimbabwe