Enhancing MICS Surveys with Satellite Images and Ethnographic evidence

A case study for Vietnam

Rolando Gonzales Martínez
Marie Skłodowska-Curie Postdoctoral Fellow
Faculty of Spatial Sciences
University of Groningen (Netherlands)

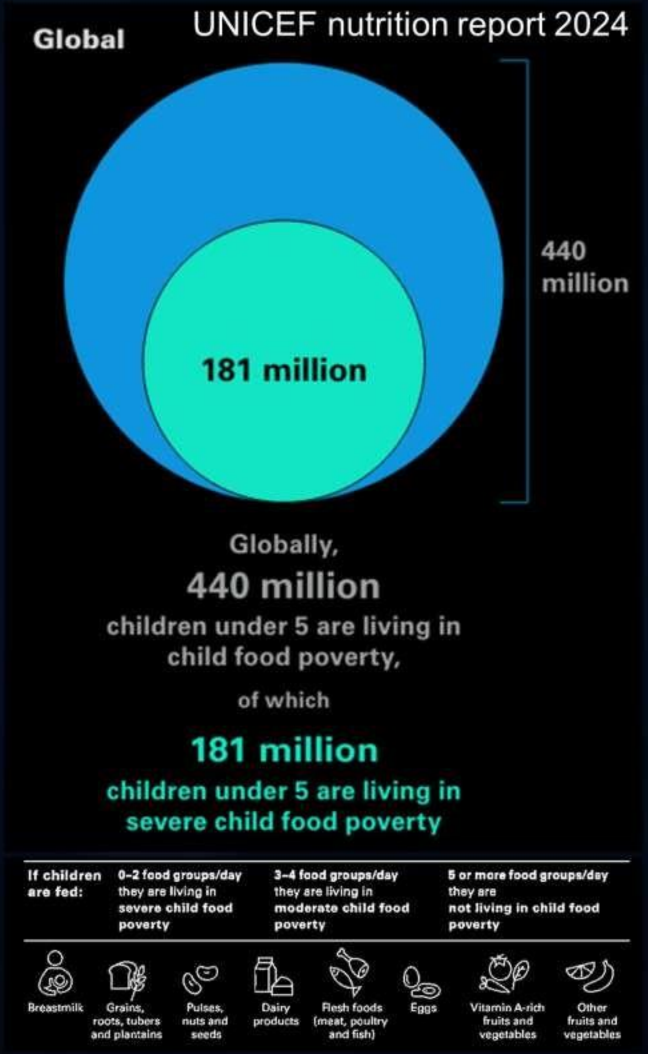
MSCA project:
Spatial machine learning and deep learning models for vulnerable children monitoring

MSCA (Horizon Europe's Marie Skłodowska-Curie Actions): European Commission funds that promote research and innovation projects.

Goal of the project: develop **tools** for monitoring progress towards the SDGs, crafting targeted interventions, and reducing the costs associated with development programs.

How?

- Expanding the information of survey data with satellite imagery
- Applying machine learning/deep learning to identify vulnerable children (with a special focus on nutrition)
- Collecting ethnographic evidence about vulnerability to natural disasters and other factors



Overview of MICS Survey for Vietnam (2020-2021)

	Children under 5		Health		Child development		WASH		Nutrition	
	Sample (HH)	Low functioning	Diarrhoea	Fever	Inadequate supervision	ECDI	Sewers	Sanitation	Diet diversity	Breast-feeding
National	100	1.2	4.8	17.4	6.5	78.2	5	87	55.2	23.5
Central Highlands	7	1.8	11.1	29.1	12.1	69.8	0	83	55.6	31.6
Norther Midlands and Mountainous	15	2.1	7	22.1	8.9	69.1	0	87	41.9	28
Mekong River Delta	15	1.6	3.5	15.4	1.7	79.2	1	77	52	14.6
North Central and Central Coastal	22	0.8	4.9	16.2	10.6	77.3	12	83	66.4	23.6
South East	16	0.6	4.1	16.5	3.8	77.2	3	95	46	25.2
Red River Delta	25	1.1	2.9	14	4.6	87.4	5	94	61.8	21.3
Ho Chi Minh City	8	0.3	3.2	11.8	2.6	78.5	5	94	42.3	34
Ha Noi	8	1.6	4	13.6	5	91.8	9	90	76.1	14.4

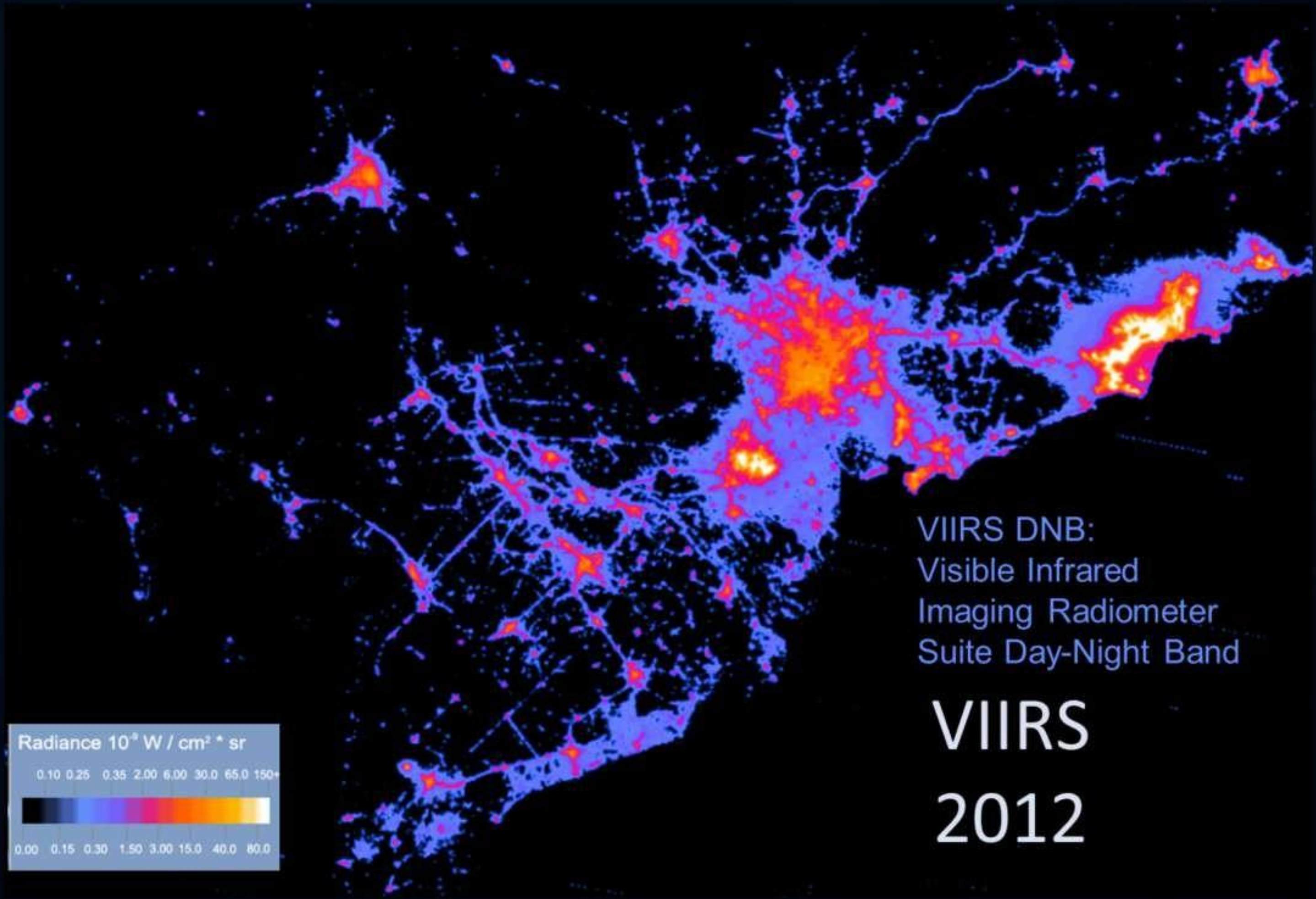
ECDI: Early Childhood Development Index

Spatio-temporal satellite images provide up-to-date geographic information that can enhance MICS surveys

Satellite images can capture:

- The economic situation of households outside (but around) those of the MICS sample through spatio-temporal microsimulations

- The expansion of obesogenic environment of micro-nutrient poor but ubiquitous energy-dense food choices: with urbanization, the consumption of oils, sweeteners, processed foods, and foods prepared away from home increases (Tzioumis and Adair, 2014)

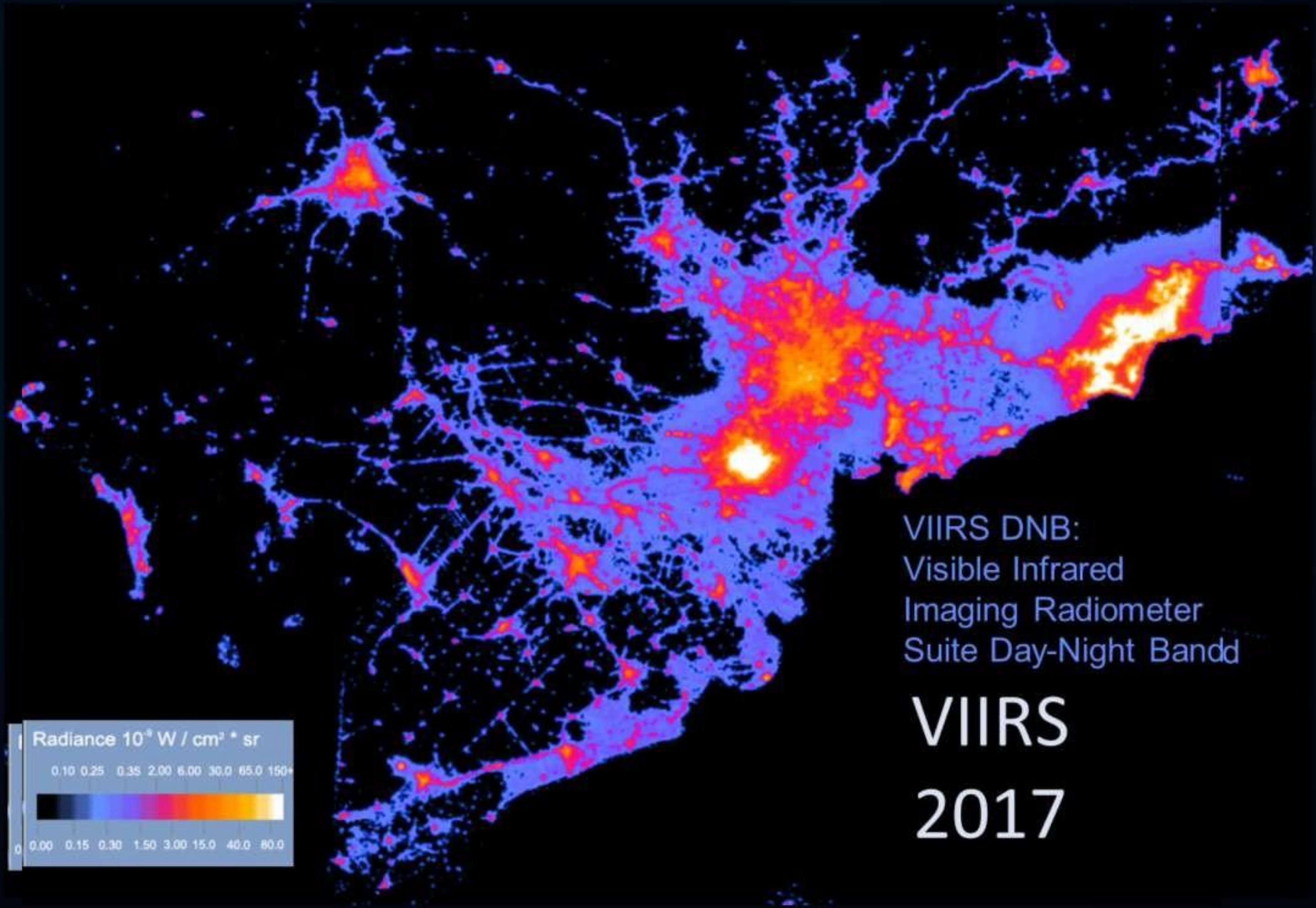


Spatio-temporal satellite images provide up-to-date geographic information that can enhance MICS surveys

Satellite images can capture:

- The economic situation of households outside (but around) those of the MICS sample through spatio-temporal microsimulations

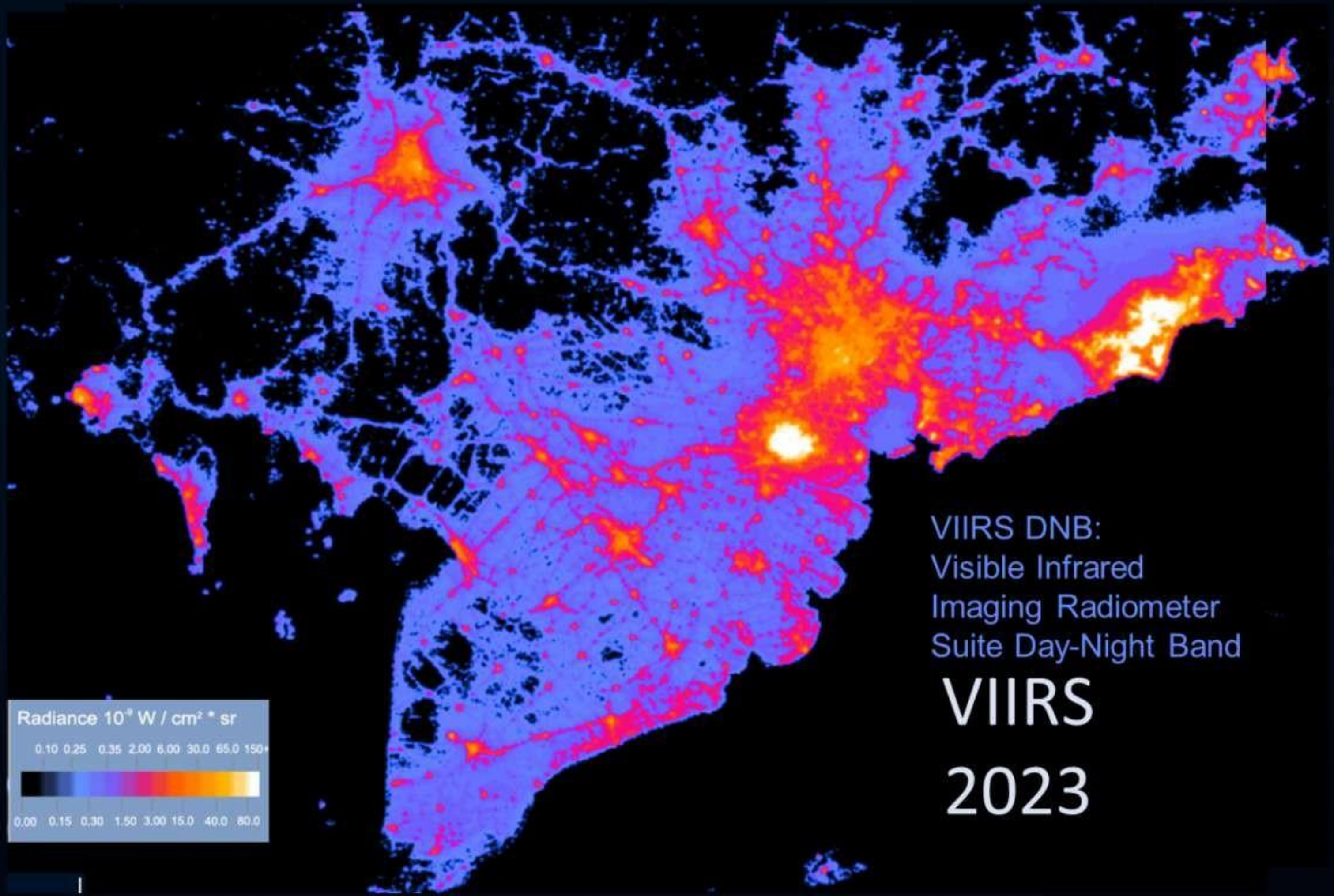
- The expansion of obesogenic environment of micro-nutrient poor but ubiquitous energy-dense food choices: with urbanization, the consumption of oils, sweeteners, processed foods, and foods prepared away from home increases (Tzioumis and Adair, 2014)



Spatio-temporal satellite images provide up-to-date geographic information that can enhance MICS surveys

Satellite images can capture:

- The economic situation of households outside (but around) those of the MICS sample through spatio-temporal microsimulations
- The expansion of obesogenic environment of micro-nutrient poor but ubiquitous energy-dense food choices: with urbanization, the consumption of oils, sweeteners, processed foods, and foods prepared away from home increases (Tzioumis and Adair, 2014)



Ethnographic studies can provide additional information to complement surveys

- Surveys usually focus on aggregates and central tendencies

- Ethnographic methods instead focus on tails and special cases, and thus can help to identify children and mothers that could be left behind during the development process towards the SDGs.

- Framed WHO-UNICEF dietary diversity score (DDS), what environmental, household and individual level factors (physical, psychological, social, economic, demographic, climate and geographic factors) can put children and women at risk of being left behind during the development process towards the SDGs?

We developed a Bayesian multi-level machine learning/deep learning (BML-pMDL) algorithm to idenfity which environmental, household and child level factors can affect malnutrition and multidimensional child growth

Algorithm 1 BML-pMDL: Find $\mathbf{x} \subseteq \mathbf{X}$

Require: z_i (HAZ), $\mathbf{X} \ni \{\mathbf{X}_e, \mathbf{X}_h, \mathbf{X}_c\}$, $i = 1, 2, \ldots, n$th-children

Ensure: $z_i \in \mathbb{R}$, $k > 0$ (covariates)

$s_i = 1$ if $z_i < -2$, $s_i \in \mathbb{Z}^{0,1}$ ▷ childhood stunting

procedure pMDL$(s_i, \mathbf{X}_e, \mathbf{X}_h, \mathbf{X}_c)$ ▷ Machine learning and deep learning (MDL) algorithms for predicting childhood stunting $\mathbf{X} \ni \{\mathbf{X}_e, \mathbf{X}_h, \mathbf{X}_c\} \equiv \{x_1, x_2, \ldots, x_k\}$, $j = 1, 2, \ldots, k$th-covariates

for $j = 1, 2, \ldots, k$ **do**

$\mathrm{PRAUC}_{j,m} = f_m(x_j)$

$\mathrm{FOR}_{j,m} = f_m(x_j)$

$\varphi_{\mathrm{PRAUC}_j} = m^{-1}\mathrm{PRAUC}_{j,m}$

$\varphi_{\mathrm{FOR}_j} = m^{-1}\mathrm{FOR}_{j,m}$

$\mathrm{PriorIP}_j = \frac{e^{\varphi_{\mathrm{PRAUC}}(1-\varphi_{\mathrm{FOR}})}}{1+e^{\varphi_{\mathrm{PRAUC}}(1-\varphi_{\mathrm{FOR}})}}$

end for

return $\mathrm{PriorIP}_j$ ▷ Prior probability of predicting stunting

end procedure

procedure BML$(z_i, \mathbf{X}_e, \mathbf{X}_h, \mathbf{X}_c)$ ▷ BML modeling of child growth

$\mathbb{P}(\theta_j) \leftarrow \mathrm{PriorIP}_j$

for $\mathbf{X}_e \subseteq \mathbf{X}$ **do**

$\mathbb{P}(\theta_e|\mathbf{X}_e) \propto \mathbb{P}(\theta_e)\mathcal{L}(\mathbf{X}_e|\theta_e)$

$\beta_{ej} := \mathbb{E}[\beta_e|\mathbf{z}, \mathbf{X}_e] = (g+1)^{-1}\left(\bar{\beta}_e + g\hat{\beta}_e\right)$

$PostIP_{ej} = \sum_{j=1}^{2^k} I(x_{e,k} \in \mathcal{M}_{ej})\mathbb{P}(\mathcal{M}_{ej}|\mathbf{z})$

end for

return β_{ej}, $PostIP_{ej}$, $\bar{\mathbf{x}}_e \subseteq \mathbf{X}_e$ ▷ β_{ej} estimates, posterior probability (PostIP) of external environmental factors being associated with child growth, and $\bar{\mathbf{x}}_e$ subset of relevant external environmental factors

for $\mathbf{X}_h \subseteq \mathbf{X}$ **do**

$\mathbb{P}(\theta_h|\mathbf{X}_h, \bar{\mathbf{x}}_e) \propto \mathbb{P}(\theta_h)\mathcal{L}(\mathbf{X}_h, \bar{\mathbf{x}}_e|\theta_h)$

$\beta_{hj} := \mathbb{E}[\beta_h|\mathbf{z}, \mathbf{X}_h, \bar{\mathbf{x}}_e] = (g+1)^{-1}\left(\bar{\beta}_h + g\hat{\beta}_h\right)$

$PostIP_{hj} = \sum_{j=1}^{2^k} I(x_{h,k} \in \mathcal{M}_{hj})\mathbb{P}(\mathcal{M}_{hj}|\mathbf{z})$

end for

return β_{hj}, $PostIP_{hj}$, $\bar{\mathbf{x}}_h \subseteq \mathbf{X}_h$ ▷ β_{hj} estimates, posterior probability (PostIP) of household factors being associated with child growth, and $\bar{\mathbf{x}}_h$ subset of relevant household factors for child growth

for $\mathbf{X}_c \subseteq \mathbf{X}$ **do**

$\mathbb{P}(\theta_c|\mathbf{X}_c, \bar{\mathbf{x}}_e, \bar{\mathbf{x}}_h) \propto \mathbb{P}(\theta_c)\mathcal{L}(\mathbf{X}_c, \bar{\mathbf{x}}_e, \bar{\mathbf{x}}_h|\theta_c)$

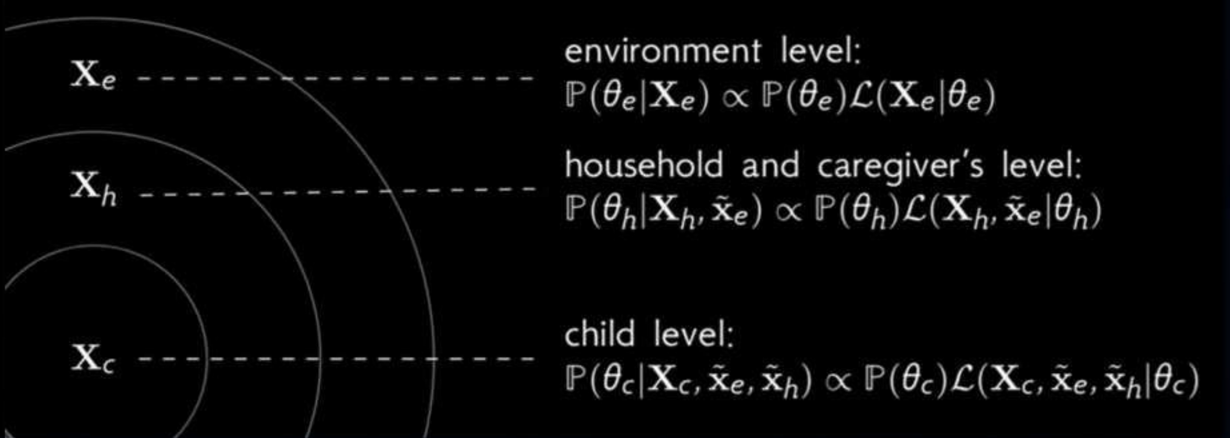
$\beta_{cj} := \mathbb{E}[\beta_c|\mathbf{z}, \mathbf{X}_c, \bar{\mathbf{x}}_e, \bar{\mathbf{x}}_h] = (g+1)^{-1}\left(\bar{\beta}_c + g\hat{\beta}_c\right)$

$PostIP_{cj} = \sum_{j=1}^{2^k} I(x_{c,k} \in \mathcal{M}_{cj})\mathbb{P}(\mathcal{M}_{cj}|\mathbf{z})$

end for

return β_{cj}, $PostIP_{cj}$, $\bar{\mathbf{x}}_c \subseteq \mathbf{X}_c$ ▷ β_{cj} estimates, posterior probability (PostIP) of child level factors being associated with child growth, and the $\bar{\mathbf{x}}_c$ subset of relevant child level factors

end procedure



BML-pMDL algorithm (1/2)

Prior probability of predicting stunting based on machine learning and deep learning: focus on prevention policies: PR-AUC and FOR of six machine learning models (logistic regression, random forests, multilayered perceptrons, stochastic gradient classifiers, XGBoost, support vector machines) and a deep learning model with RELU activations functions and L1-L2 regularizers

Epistemological and ontological approach: Anarchistic Theory of Knowledge (Feyerabend) and Bayesian methods

Evidence from surveys and satellite images

$$\mathbb{p}(\theta|\mathbf{X}) \propto \mathbb{p}(\theta)\mathcal{L}(\mathbf{X}|\theta)$$

Bayesian model averaging to identify the most important determinant of child growth (HAZ z-scores)

Vulnerable children that could be left behind

$$\mathbb{p}(\theta|\mathbf{D},\mathbf{X}) \propto \mathbb{p}(\theta)\mathbb{p}(\theta|\mathbf{X})\mathbb{p}(\theta|D)$$

Ethnographic evidence based on interpretivism subjective experiences and meanings of individuals in the cultural, historical, and social contexts in which they occur

Ethnographic evidence based on semi-structured interviews

BML-pMDL algorithm (2/2)

$$s_i = \begin{cases} 1 & \text{if } z_{ha} < -2 \\ 0 & \text{else} \end{cases}$$

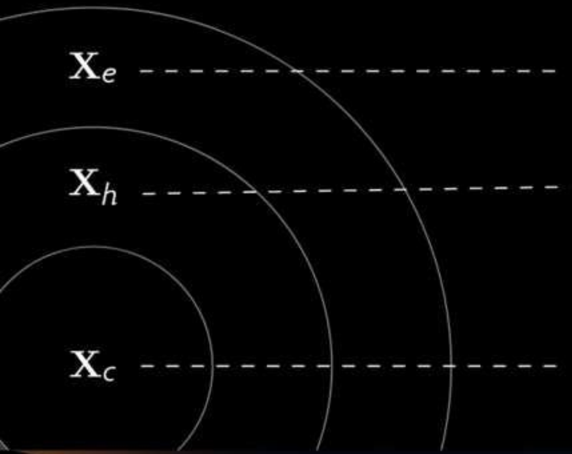
$$= \frac{e^{\bar{\varphi}_{\text{PRAUC}}(1-\bar{\varphi}_{\text{FOR}})}}{1 + e^{\bar{\varphi}_{\text{PRAUC}}(1-\bar{\varphi}_{\text{FOR}})}}$$

$$\mathbb{E}\left[\beta|\mathbf{z},\mathbf{X}\right] = \frac{1}{g+1}\left(\tilde{\beta} + g\hat{\beta}\right),$$

$$\sum_{k=0}^{m}\binom{m}{k} := \sum_{k=0}^{m}\frac{m!}{k!(m-k)!} = 2^k$$

$$\mathbb{E}\left[\sigma^2|\mathbf{z},\mathbf{X}\right] = \frac{\hat{\sigma}^2 + \left(\tilde{\beta}-\hat{\beta}\right)^T\mathbf{X}^T\mathbf{X}\left(\tilde{\beta}-\hat{\beta}\right)/(g+1)}{n-2}$$

$$PIP(x_k) = \sum_{j=1}^{2^k} I(x_k \in \mathcal{M}_j)\mathbb{P}(\mathcal{M}_j|\mathbf{z})$$



environment level:
$\mathbb{P}(\theta_e|\mathbf{X}_e) \propto \mathbb{P}(\theta_e)\mathcal{L}(\mathbf{X}_e|\theta_e)$

household and caregiver's level:
$\mathbb{P}(\theta_h|\mathbf{X}_h, \tilde{\mathbf{x}}_e) \propto \mathbb{P}(\theta_h)\mathcal{L}(\mathbf{X}_h, \tilde{\mathbf{x}}_e|\theta_h)$

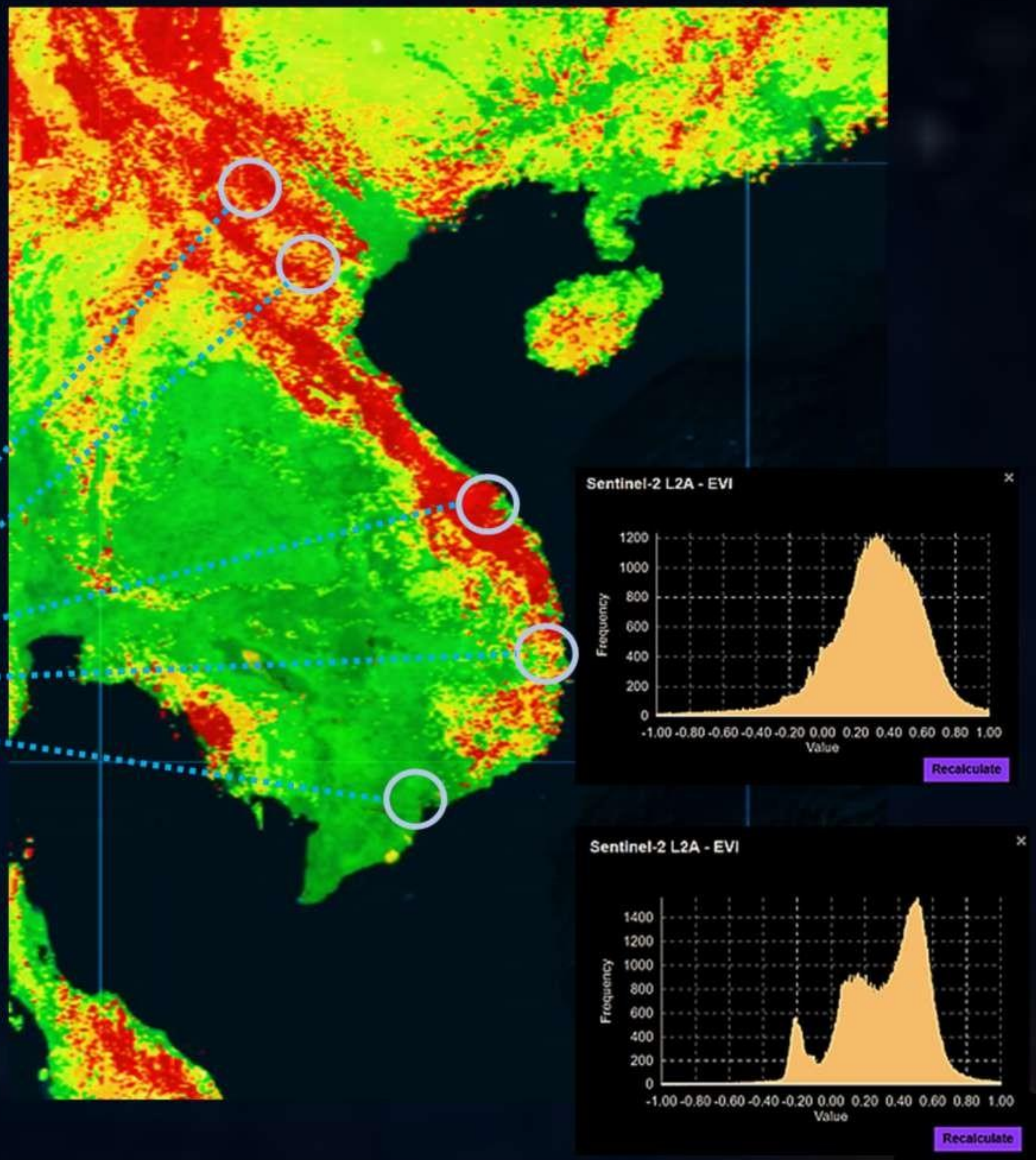
child level:
$\mathbb{P}(\theta_c|\mathbf{X}_c, \tilde{\mathbf{x}}_e, \tilde{\mathbf{x}}_h) \propto \mathbb{P}(\theta_c)\mathcal{L}(\mathbf{X}_c, \tilde{\mathbf{x}}_e, \tilde{\mathbf{x}}_h|\theta_c)$

Bayesian multilevel modelling based on an ecobiological approach: Dahlgren-Whitehead model, MICG-IUNS approach, WHO-Lancet comission/UNICEF: physical and non-physical determinants

Example of application: Combining information of satellite images with surveys

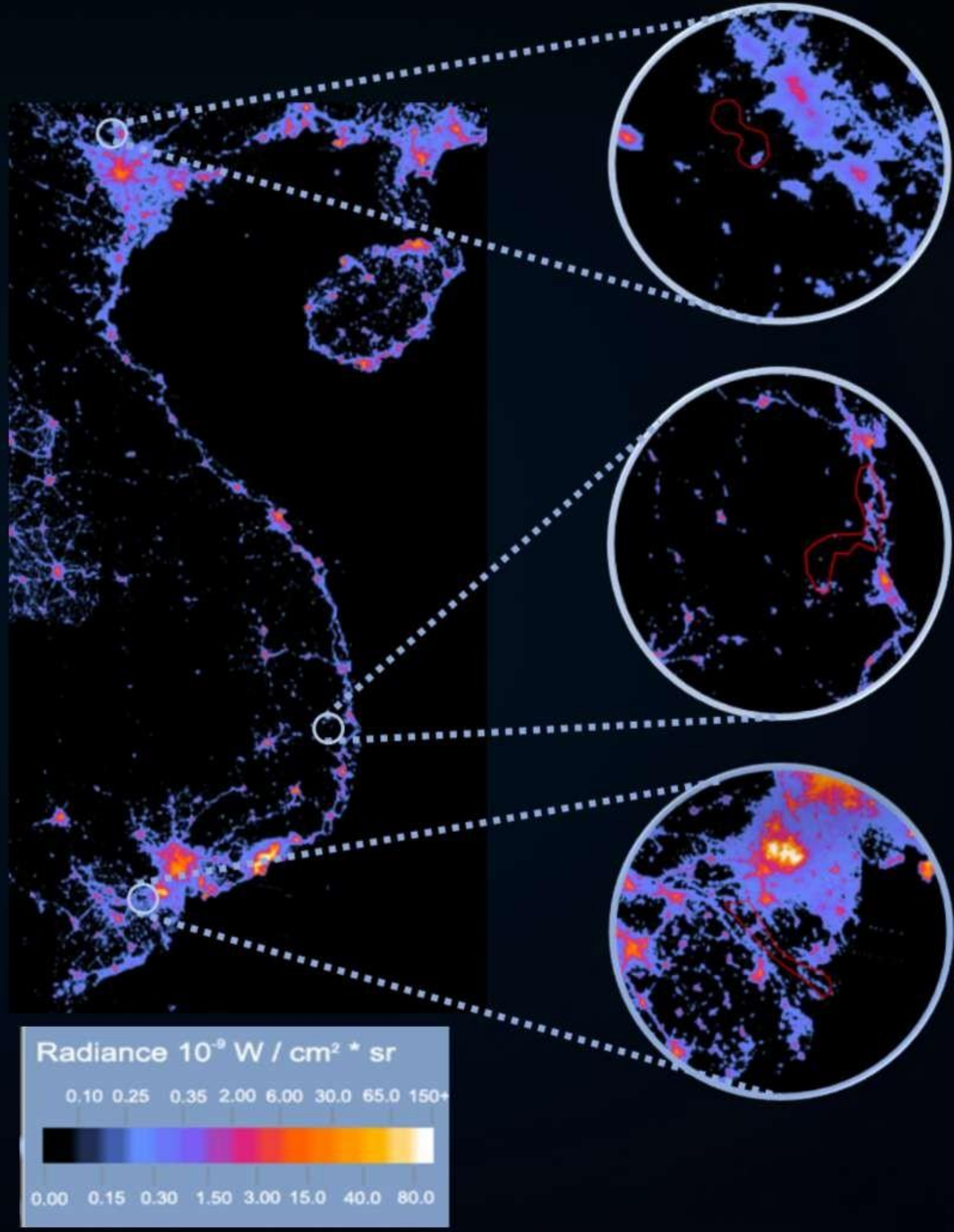
k = 215 potential explanatory variables of stunting/child growth:

- 207 from Young Lives Survey (YLS) 12,000 children in four countries, (Ethiopia, India, Peru, and Vietnam). Round 2 (5yo, 2006) in Vietnam collected from sentinel sites of Northern Uplands, the Red River Delta, the Central Coastal region, the Mekong River Delta and Da Nang.
- 7 environmental variables: Vegetation index, luminosity, landslide risk
- 1 geographical variable: fish production



Information of satellite images

- Vegetation index (EVI, Copernicus Sentinel-2 mission)
- Luminosity (categorized:low, middle, high): VIIRS DNB: Visible Infrared Imaging Radiometer Suite Day-Night Band (NASA/ NOAA missions and Suomi-NOAA satellites)
- Landslide risk (categorized: low, middle, high).



Results

By using stunting as a marker of multidimensional child growth in Bayesian ML/DL with 215 variables:

- Rural regions (e.g. Ben Tre)
- Natural disasters
- Low fish production
- Low luminosity
- Low mother's BMI
- Mother's ethnic group
- Children in the household (parity)
- Health of children compared to peers
- Low verbal development
- No attendance to pre-school

Results

Feature (potential explanatory variables of stunting and child growth)	Machine learning-deep learning*			Bayesian estimates**		
	PR-AUC	FOR	Prior IP	Mean	Std. dev	Post IP
rural region (rural=1)	0.762	0.311	0.628	-0.855	0.074	1.000
people in the street look down on their familiy (yes=1)	0.700	0.568	0.575	-0.166	0.124	0.728
safety in the street (no=1)	0.746	0.378	0.614	0.070	0.076	0.532
low fish production by province	0.755	0.307	0.628	-0.432	0.469	0.493
high landsalide risk (from satellite images)	0.707	0.515	0.585	-0.311	0.383	0.460
medium luminosity (from satellite images)	0.675	0.383	0.603	0.103	0.314	0.433
not enough money to buy food (yes=1)	0.659	0.547	0.574	-0.053	0.074	0.388
medium fish production by province	0.748	0.414	0.608	0.261	0.396	0.342
low luminosity (from satellite images)	0.673	0.549	0.575	-0.075	0.254	0.279
drought in community (yes=1)	0.660	0.592	0.567	-0.042	0.097	0.198
low landsalide risk (from satellite images)	0.719	0.486	0.591	0.015	0.053	0.102
medium landsalide risk (from satellite images)	0.684	0.553	0.576	-0.012	0.049	0.086
enhanced vegetation index (from satellite images)	0.684	0.585	0.571	-0.034	0.136	0.073
natural disaster in community (=1)	0.761	0.374	0.617	-0.003	0.019	0.044
any natural disaster in community in last 4 years (yes=1)	0.682	0.533	0.579	-0.003	0.020	0.038
flooding in community (yes=1)	0.722	0.478	0.593	-0.003	0.022	0.033
time to health facilities in community	0.646	0.475	0.584	-0.001	0.007	0.029
community received investment for medical station (yes=1)	0.699	0.751	0.543	0.001	0.010	0.011
access to education (1=no)	0.708	0.534	0.582	-0.001	0.016	0.011

Feature (potential explanatory variables of stunting and child growth)	Machine learning-deep learning*			Bayesian estimates**		
	PR-AUC	FOR	Prior IP	Mean	Std. dev	Post IP
rural region (rural=1)	0.762	0.311	0.628	-0.855	0.074	1.000
people in the street look down on their familiy (yes=1)	0.700	0.568	0.575	-0.166	0.124	0.728
safety in the street (no=1)	0.746	0.378	0.614	0.070	0.076	0.532
low fish production by province	0.755	0.307	0.628	-0.432	0.469	0.493
high landsalide risk (from satellite images)	0.707	0.515	0.585	-0.311	0.383	0.460
medium luminosity (from satellite images)	0.675	0.383	0.603	0.103	0.314	0.433
not enough money to buy food (yes=1)	0.659	0.547	0.574	-0.053	0.074	0.388
medium fish production by province	0.748	0.414	0.608	0.261	0.396	0.342
low luminosity (from satellite images)	0.673	0.549	0.575	-0.075	0.254	0.279
drought in community (yes=1)	0.660	0.592	0.567	-0.042	0.097	0.198
low landsalide risk (from satellite images)	0.719	0.486	0.591	0.015	0.053	0.102
medium landsalide risk (from satellite images)	0.684	0.553	0.576	-0.012	0.049	0.086
enhanced vegetation index (from satellite images)	0.684	0.585	0.571	-0.034	0.136	0.073
natural disaster in community (=1)	0.761	0.374	0.617	-0.003	0.019	0.044
any natural disaster in community in last 4 years (yes=1)	0.682	0.533	0.579	-0.003	0.020	0.038
flooding in community (yes=1)	0.722	0.478	0.593	-0.003	0.022	0.033
time to health facilities in community	0.646	0.475	0.584	-0.001	0.007	0.029
community received investment for medical station (yes=1)	0.699	0.751	0.543	0.001	0.010	0.011
access to education (1=no)	0.708	0.534	0.582	-0.001	0.016	0.011

Feature (potential explanatory variables of stunting and child growth)	Machine learning-deep learning*			Bayesian estimates**		
	PR-AUC	FOR	Prior IP	Mean	Std. dev	Post IP
health compared to other children (1=worse)	0.601	0.508	0.573	-0.468	0.049	1.000
score of verbal development (standardized)	0.634	0.449	0.586	0.094	0.027	0.984
attendance to pre-school (=1)	0.608	0.507	0.574	0.172	0.126	0.719
child's height shorter than other children (yes=1)	0.691	0.510	0.584	-0.025	0.049	0.238
children received hib immunization (yes=1)	0.654	0.669	0.554	0.025	0.052	0.221
eat sugar/honey in last 24 hours (yes=1)	0.700	0.513	0.584	0.003	0.017	0.034
medical expenses in child in last 12 months (yes=1)	0.673	0.560	0.573	0.002	0.015	0.033
ethnicity (1=not kinh)	0.685	0.430	0.596	-0.005	0.039	0.027
hours sleep time of children (standardized)	0.613	0.650	0.553	-0.001	0.007	0.024
any type of physical disability affecting child (yes=1)	0.613	0.759	0.537	-0.004	0.031	0.020

environment level:
$\mathbb{P}(\theta_e|\mathbf{X}_e) \propto \mathbb{P}(\theta_e)\mathcal{L}(\mathbf{X}_e|\theta_e)$

household and caregiver's level:
$\mathbb{P}(\theta_h|\mathbf{X}_h, \tilde{\mathbf{x}}_e) \propto \mathbb{P}(\theta_h)\mathcal{L}(\mathbf{X}_h, \tilde{\mathbf{x}}_e|\theta_h)$

child level:
$\mathbb{P}(\theta_c|\mathbf{X}_c, \tilde{\mathbf{x}}_e, \tilde{\mathbf{x}}_h) \propto \mathbb{P}(\theta_c)\mathcal{L}(\mathbf{X}_c, \tilde{\mathbf{x}}_e, \tilde{\mathbf{x}}_h|\theta_c)$

Conclusion and future work

- By integrating satellite images with surveys, researchers and policy makers gain an up-to-date and geogrpahically comprehensive understanding of socio-environmental determinants impacting child well-being

- Ethnographic studies can further improve contextual understanding and offer a holistic view of children's lives that helps to identify mother's and children that could be left behind during the development progress towards the SDGs

- Bayesian spatial machine learning/deep learning methods help to identify which environmental, household and child level variables are relation to multidimensional child growth and can prevent malnutrition

Future work:
- Ethnographic ieldwork in Vietnam: childcare volunteer program in Ho Chi Minh (participatory observation and semi-structured interviews).
- Integration of spatial contiguity matrices in the BML-pMDL algorithm and additional spatial information and satellite information (ethnicity, natural disasters)
- Microsimulations based on MICS surveys and satellite images

Enhancing MICS Surveys with Satellite Images and Ethnographic evidence

A case study for Vietnam

Rolando Gonzales Martínez

Marie Skłodowska-Curie Postdoctoral Fellow
Faculty of Spatial Sciences
University of Groningen (Netherlands)