



This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 870661



# Use of non-traditional data sources to nowcast migration trends through Artificial Intelligence technologies.

Diletta Goglia, Laura Pollacci, Alina Sîrbu  
Computer Science department, University of Pisa, Italy



AI4MIG

Diletta Goglia

Postgraduate Student in MSc in Artificial Intelligence  
Computer Science department, University of Pisa, Italy

✉ [d.goglia@studenti.unipi.it](mailto:d.goglia@studenti.unipi.it)

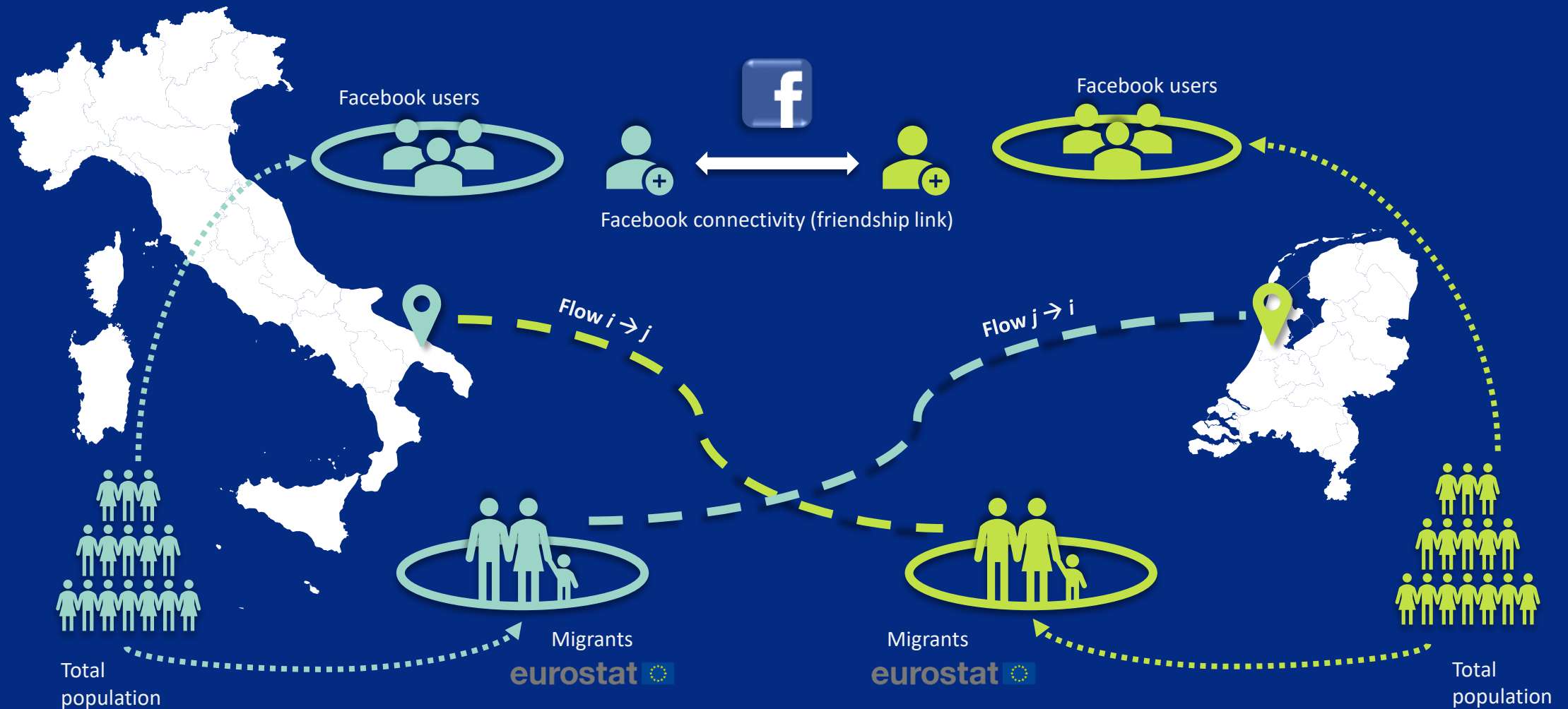
🌐 [www.dilettagoglia.netlify.app](http://www.dilettagoglia.netlify.app)

🐦 [@dilettagoglia](https://twitter.com/dilettagoglia)

# Nowcasting flows with Facebook Social Connectedness

Country  $i$

Country  $j$





# Starting point: MIMI dataset

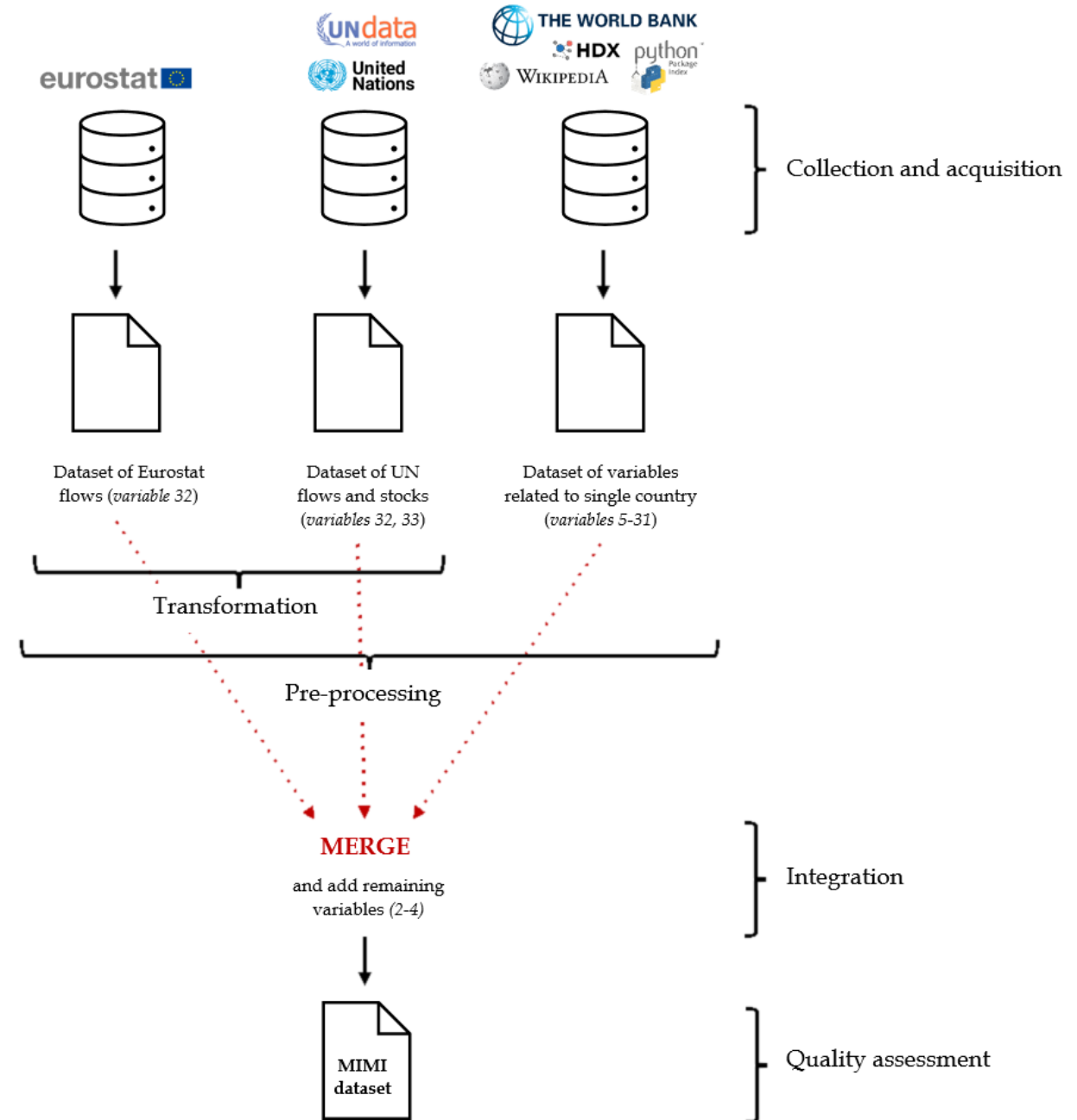
Multi-aspect Integrated Migration Indicators dataset (CC BY 4.0). Publicly available at [10.5281/zenodo.6493325](https://doi.org/10.5281/zenodo.6493325)



- Non-trivial integration of scattered and heterogeneous data.
- Multidisciplinary variables and original indicators (economic, demographic, cultural and geographic), including Social Network Data.
- Demographers, sociologists, economists and researchers could explore and investigate trends.
- Possibility to develop complex models based on MIMI data.

## Related works

- Goglia, D. (2022) "Multi-aspect Integrated Migration Indicators (MIMI) dataset v2.0", Zenodo, [10.5281/zenodo.6493325](https://doi.org/10.5281/zenodo.6493325).
- Goglia, D., Pollacci, L., Sirbu, A. (2022) "Dataset of Multi-aspect Integrated Migration Indicators", ArXiv pre-print, <https://arxiv.org/abs/2204.14223>





# Facebook connectivity & cross-border human mobility

Hypothesis: Facebook strength of connectivity between two countries is related to the bilateral contribution of migration flows.  
We built a new indicator (BMI) to nowcast migration flows based on Facebook SCI.

## Facebook Social Connectedness Index (SCI)

$$SCI_{i,j} = \frac{FB\_Connections_{i,j}}{FB\_Users_i * FB\_Users_j}$$

- **Non-traditional** variable within the context of migration studies, included in MIMI for 2020 and 2021 connectivity.
- Provided by “Data for Good at Meta (previously Facebook)” on Humanitarian Data Exchange.
- Measures the **relative probability** that two individuals across two locations  $i$  and  $j$  are friends with each other on Facebook.
- Could be employed as a proxy of social connections across borders, to be studied as a **possible driver of migration**.
- **Symmetric** structure by definition of “friendship” concept.

## Bidirectional Migration Index (BMI)

$$BMI_{i,j} = \frac{Flow_{i \rightarrow j}(t) + Flow_{j \rightarrow i}(t)}{Pop_i(t) * Pop_j(t)}$$

- $i$  and  $j$  represent countries (NUT0 level).
- $t$  is the **reference year**.
- Sum of **direct** and **inverse** contributions → **symmetric** function considering the sharing of both flows.
- Strong **correlation** with SCI for flows provided by Eurostat statistics.
- Can be computed also by sex and by age group.

For couples of countries with inverse flow not available:  
**Unidirectional Migration Index** (asymmetric – directed – version of BMI).

$$UMI_{i,j} = \frac{Flow_{i,j}}{Pop_i * Pop_j}$$





# Linear Regression

## Analysis of migration drivers

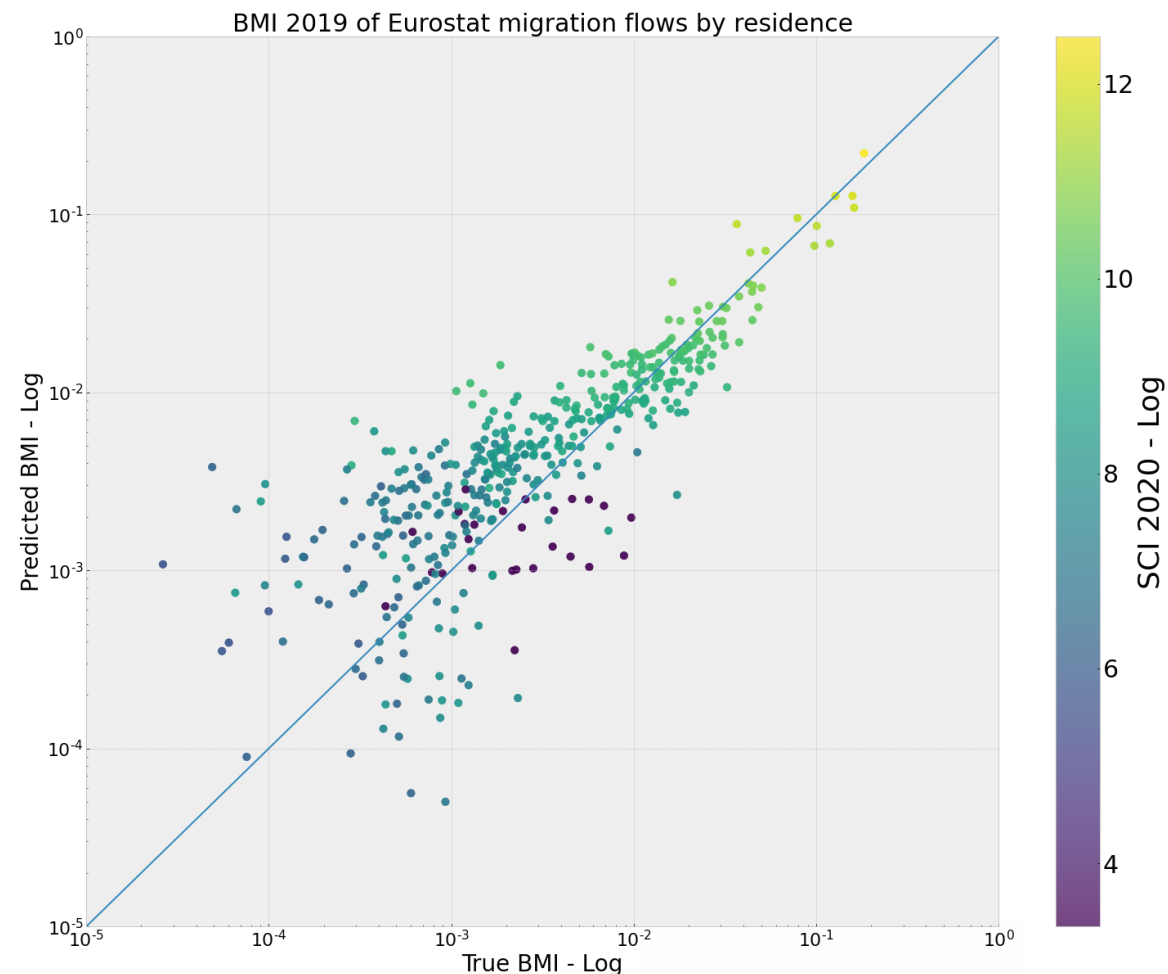
- **Model:** Linear regression (Ordinary Least Squares model).
- **Task:** simple fit to determine which variables are relevant and to prove the importance of SCI on BMI prediction.
- **Aim:** feature selection for further neural models.

*BMI 2019 of EUROSTAT migration flows by residence: true value VS linear regression predictions (best fit, model with  $R^2 = 0.88$ ). Each data point in the plot represents a couple of countries.*



## Preliminary results

- Facebook strength of connectivity between two countries is strongly positively related to the amount of migration flows they share.
- Higher connectivity, more accurate migration prediction.

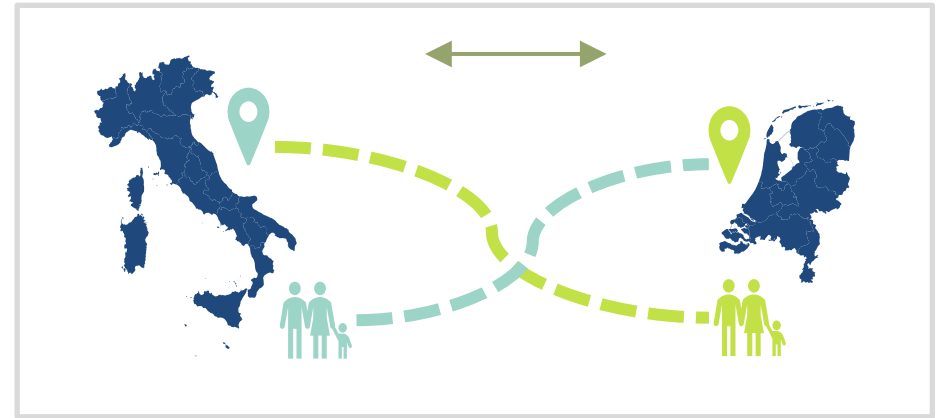


# Results of OLS Backward Elimination Stepwise Linear Regression

## Bidirectional Migration Index

	By citizenship (148 couples, 14 countries)				By residence (1114 couples, 63 countries)			
Y	ESTAT BMI 2019 cit, without sci 2020		ESTAT BMI 2019 cit, with sci 2020		ESTAT BMI 2019 res, without sci 2020		ESTAT BMI 2019 res, with sci 2020	
Model n.	1	11	1	6	1	10	1	11
Feature	coef, P> t	coef, P> t	coef, P> t	coef, P> t	coef, P> t	coef, P> t	coef, P> t	coef, P> t
Intercept	-0.0005		0.0409 ***	0.0412 ***	0.0187 ***	0.0210 ***	0.1079 ***	0.1075 ***
sci_2020			0.0438 ***	0.0438 ***			0.1090 ***	0.1092 ***
geodesic_distance_km	-0.0161 **	-0.0114 ***	-0.0023	-0.0021	-0.0025		9.253e-05	
gdp_diff_2018	0.0074 **	0.0064 **	0.0032 ***	0.0033 ***	0.0040 ***	0.0039 ***	0.0018 ***	0.0016 ***
gdp_mean_2018	-0.0091 *	-0.0074 **	-0.0117 ***	-0.0116 ***	0.0058 ***	0.0055 ***	0.0005	
neighbours	0.0075 **	0.0081 ***	-0.0021 *	-0.0020 **	0.0138 ***	0.0140 ***	-0.0011	-0.0012 *
share_rel	0.0078 ***	0.0070 ***	0.0014	0.0013	0.0006		-0.0005 *	-0.0005 *
share_lang	-0.0024		-0.0013 *	-0.0014 **	0.0020 ***	0.0020 ***	0.0004	0.0004
PDI_diff	-0.0038	-0.0034	0.0020	0.0020	-0.0025 *	-0.0026 *	-0.0006	
IDV_diff	-0.0067 **	-0.0065 **	-0.0021 *	-0.0021 *	-0.0012		-0.0017 ***	-0.0017 ***
UAI_diff	0.0014		-0.0033 *	-0.0033 *	0.0003		0.0006	
MAS_diff	-0.0103 ***	-0.0088 ***	-0.0021 *	-0.0021 *	-0.0051 ***	-0.0050 ***	-0.0017 ***	-0.0018 ***
fb_users_diff	-0.0018		-0.0023 *	-0.0024 ***	0.0066	0.0068	-0.0007	
fb_users_perc_diff	0.0025		0.0043 ***	0.0044 ***	-0.0006		-8.071e-05	
fb_users_perc_mean	0.0028	0.0049	0.0025	0.0026 *	0.0054 ***	0.0055 ***	0.0044 ***	0.0047 ***
fb_users_mean	-0.0134 **	-0.0145 ***	-0.0005		-0.0114 *	-0.0112 *	-0.0010	-0.0016 **
area_diff	-0.0019		0.0002		-0.0029		0.0012	
area_mean	0.0036		0.0004		0.0034		-0.0011	
share_cont	0.0005				0.0047 ***	0.0063 ***	0.0010	0.0011 ***
R2 (centered)	0.574	0.560	0.949	<b>0.949</b>	0.363	0.361	0.880	<b>0.880</b>
AIC	-846.3	-855.7	-1158.	<b>-1164.</b>	-6321.	-6332.	-8179.	<b>-8191.</b>
BIC	-795.4	-825.7	-1104.	<b>-1119.</b>	-6231.	-6276.	-8084.	<b>-8136.</b>

Previous plot



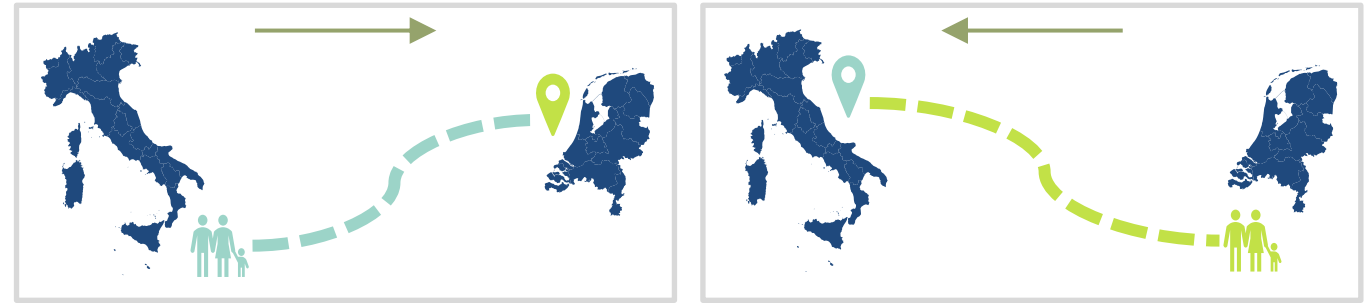
- Dependent variable (y) is BMI → Symmetric indicator → Independent variables are undirected.
- Four different models: flows by citizenship / residence, including / not including SCI.
- Substantial improvement in models with SCI.
- SCI positively and strongly relevant feature.

\* p-value < 0.05  
 \*\* p-value < 0.01  
 \*\*\* p-value < 0.001

# Results of OLS Backward Elimination Stepwise Linear Regression (II)

## Unidirectional Migration Index

	By citizenship (748 couples, 200+ countries)			
y	ESTAT UMI 2019 cit, without sci 2020		ESTAT UMI 2019 cit, with sci 2020	
Model n.	1	20	1	18
Feature	coef, P> t	coef, P> t	coef, P> t	coef, P> t
intercept	-0.0027		0.2080 ***	0.2072 ***
sci_2020			0.2024 ***	0.2024 ***
origin_PDI	-0.0056		-0.0009	
destination_PDI	0.0012		0.0036*	
origin_IDV	-0.0053		-0.0011	
destination_IDV	-0.0021		-0.0051	
origin_UAI	-0.0033		-0.0017	
destination_UAI	-0.0033		-0.0049 *	-0.0023 *
origin_MAS	-0.0035	-0.0053	-0.0005	
destination_MAS	0.0022		-0.0002	
origin_area	0.0002		0.0004	
destination_area	-0.0017		-0.0040 *	-0.0020 *
origin_fb_users	-0.0012		-0.0012	
destination_fb_users	-0.0038	-0.0071 ***	0.0036	
origin_fb_users_perc	0.0054	0.0043	0.0039 *	0.0025
destination_fb_users_perc	0.0012		0.0010	
geodesic_distance_km	-0.0114 *	-0.0111 ***	-0.0025	-0.0038 **
origin_gdp_2018	0.0007		-0.0015	
destination_gdp_2018	0.0026		0.0026	
gdp_diff_2018	-0.0002		-0.0018	
neighbours	-0.0049		0.0042 **	0.0040 **
share_cont	0.0010		0.0017	
share_rel	0.0031	0.0033 **	0.0009	0.0013 *
share_lang	0.0084 ***	0.0077 ***	0.0010	0.0012
R2 (centered)	0.115	0.103	0.780	<b>0.777</b>
AIC	-3043.	-3065.	-4081.	<b>-4100.</b>
BIC	-2941.	-3037.	-3975.	<b>-4059.</b>



- Dependent variable (y) is UMI → Asymmetric indicator → Direction preserved in independent variables (distinction kept between origin and destination).
- Strong improvement in the model with SCI.
- SCI is again a very relevant feature, with positive coefficient.
- Other relevant features are different w.r.t. the previous model.
- GDP does not appear but remains significant since approximated by the presence of «country area» and «FB users» variables.

\* p-value < 0.05  
 \*\* p-value < 0.01  
 \*\*\* p-value < 0.001



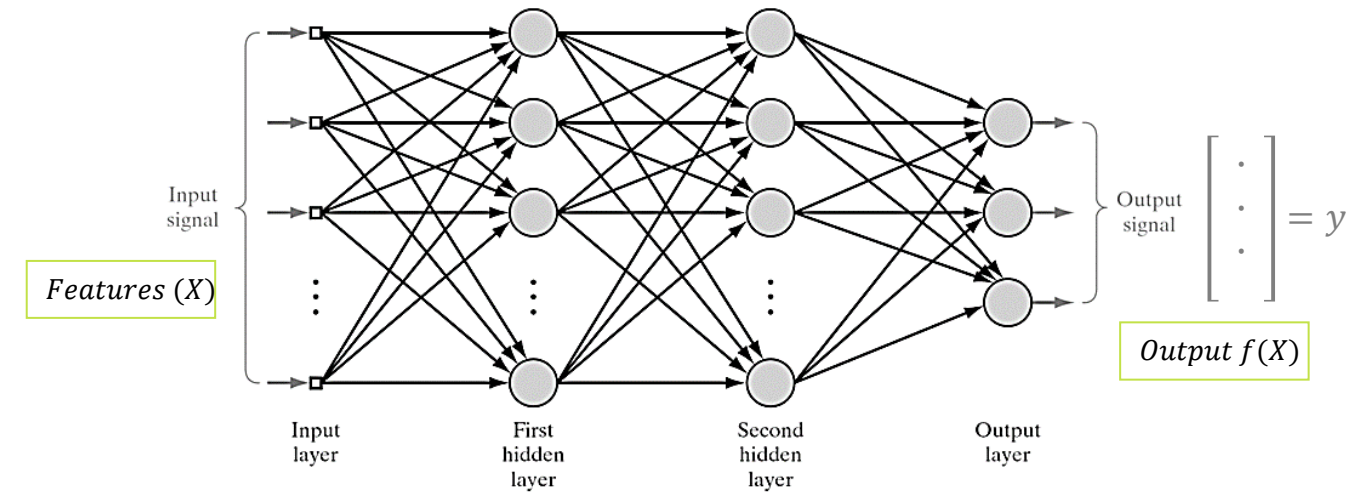
# Predictive models

New task: nowcast and predict migration flows with Machine Learning models.

- Feed Forward Artificial Neural Network (Multi-layer Perceptron)
- Random Forest



- **Supervised Learning environment (regression task)**
- **Dataset split: 80% Training, 20% Test**
- **Hyperparameter tuning**
- **Nested K-Fold Cross Validation**
- **Randomized Grid Search**



*Standard structure of MLP*



# Dataset split

Index Couple of countries	Independent variables (X matrix) Features selected by OLS model										Dependent variable (Y vector) Indicator to predict
from_to	gdp_diff_2018	sci_2020	neighbors	share_rel	share_lang	IDV_diff	MAS_diff	fb_users_perc_mean	fb_users_mean	share_cont	ESTAT_flow_BMI_2019_T_T_res
AE-AT	11616.024097	4959.0	0	0	0	17.0	27.0	63.734935	6218500.0	0	0.001545
AE-BE	16047.996984	4705.0	0	0	0	37.0	2.0	71.183916	7618500.0	0	0.007550
AE-CH	3084.091317	7485.0	0	0	0	30.0	18.0	64.416517	6218500.0	0	0.015846
AE-DK	11093.153123	4904.0	0	0	0	36.0	36.0	75.036119	6218500.0	0	0.012044
AE-FI	18995.510327	3516.0	0	0	0	25.0	26.0	67.633226	5718500.0	0	0.005339
...	...	...	...	...	...	...	...	...	...	...	...
ZM-FI	45974.507563	2855.0	0	1	0	36.0	15.0	30.088360	2476500.0	0	0.001890
ZM-IT	39430.146145	884.0	0	1	0	49.0	29.0	30.676745	16126500.0	0	0.000035
ZM-NL	54220.530866	1373.0	0	1	1	53.0	27.0	34.255180	6026500.0	0	0.000354
ZM-NO	66202.231779	2056.0	0	1	0	42.0	33.0	36.631080	2826500.0	0	0.000424
ZM-SE	49915.533011	2064.0	0	1	1	44.0	36.0	36.610776	4276500.0	0	0.000557

80% TR set<sup>1</sup>

20% TS set<sup>2</sup>

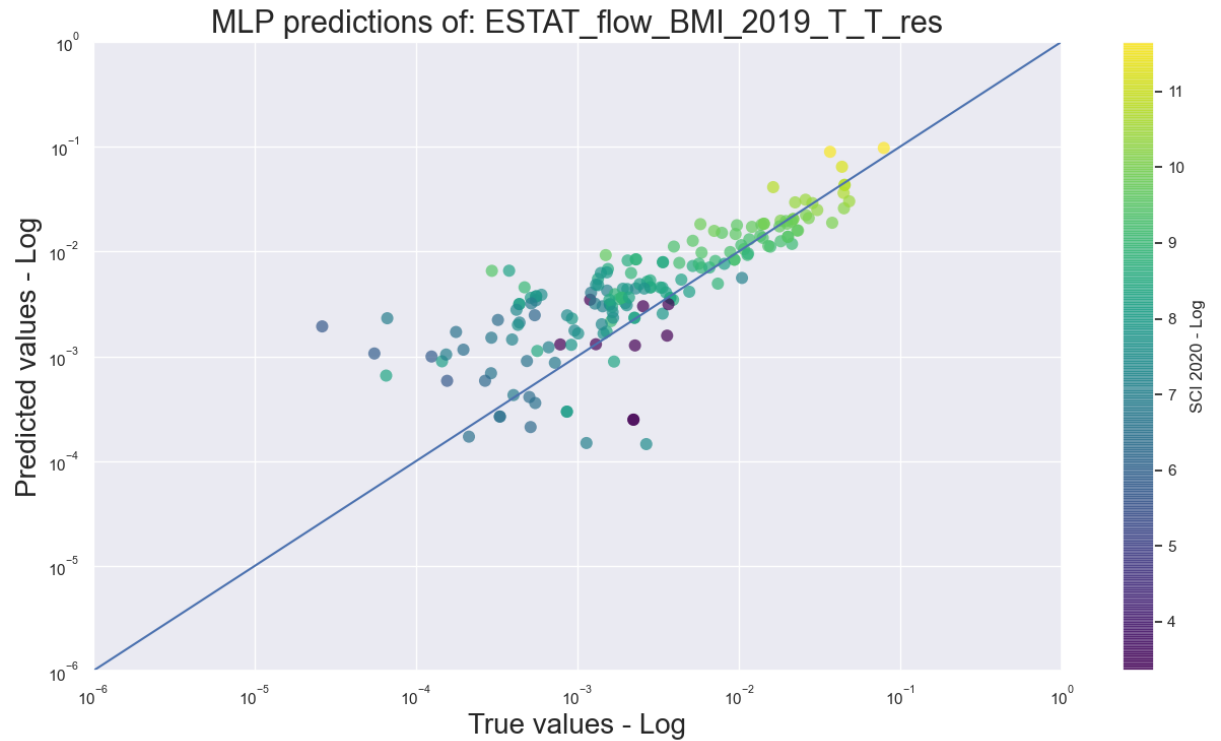
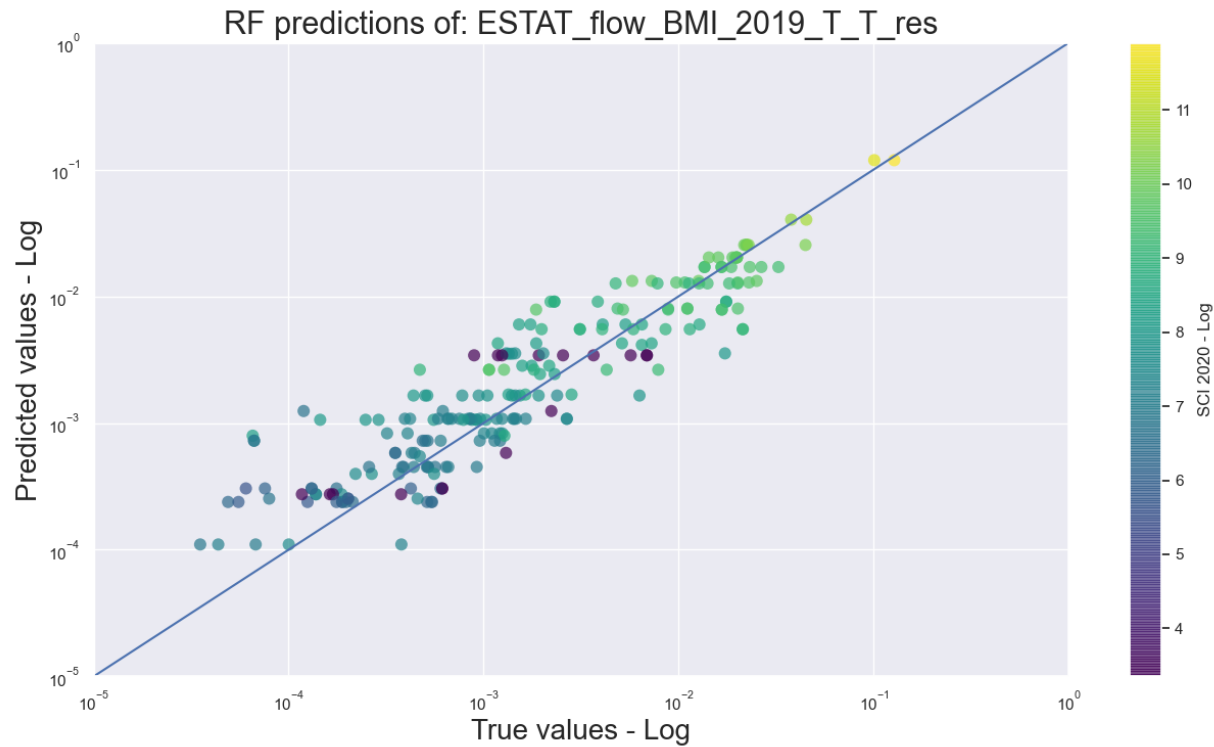
1. 891 records
2. 223 records





## Model performance outcome on nowcasting

BMI 2019 of EUROSTAT migration flows by residence: true values VS predicted model predictions.

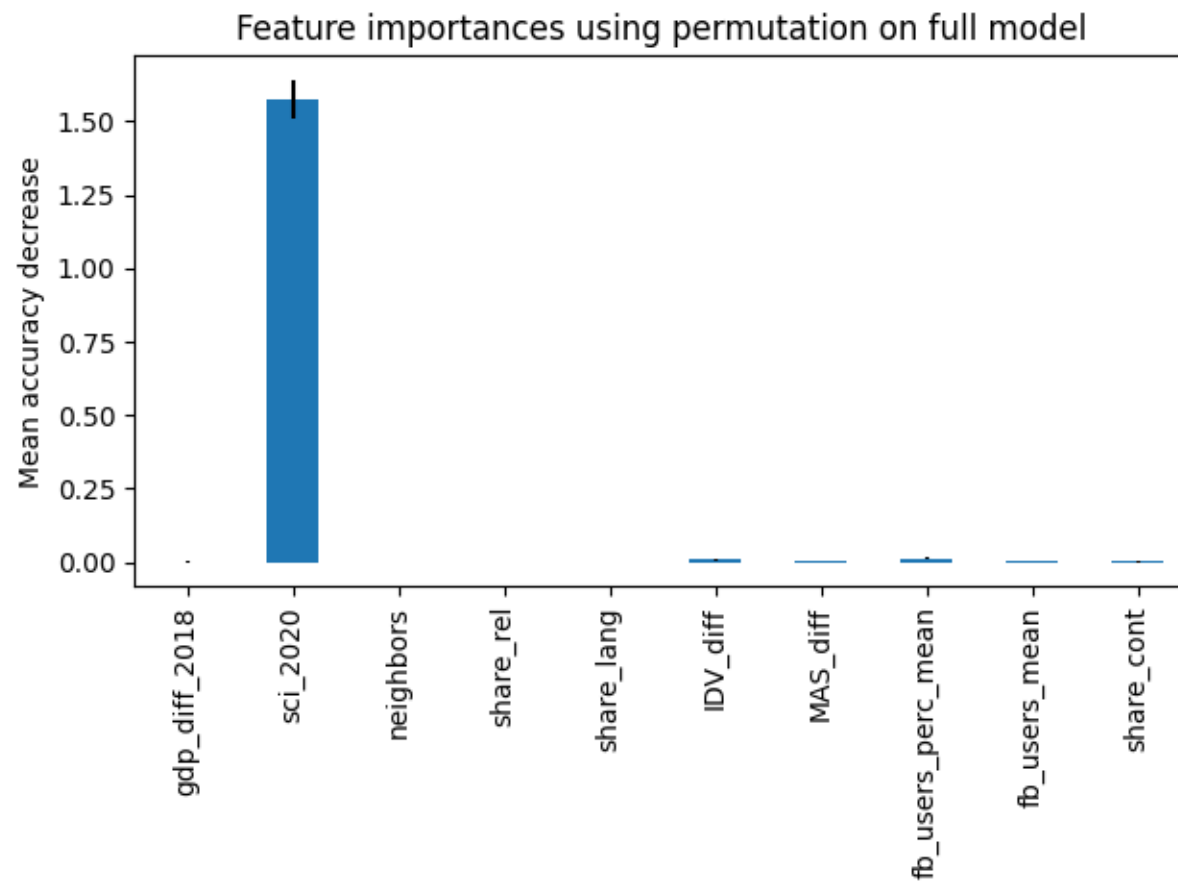
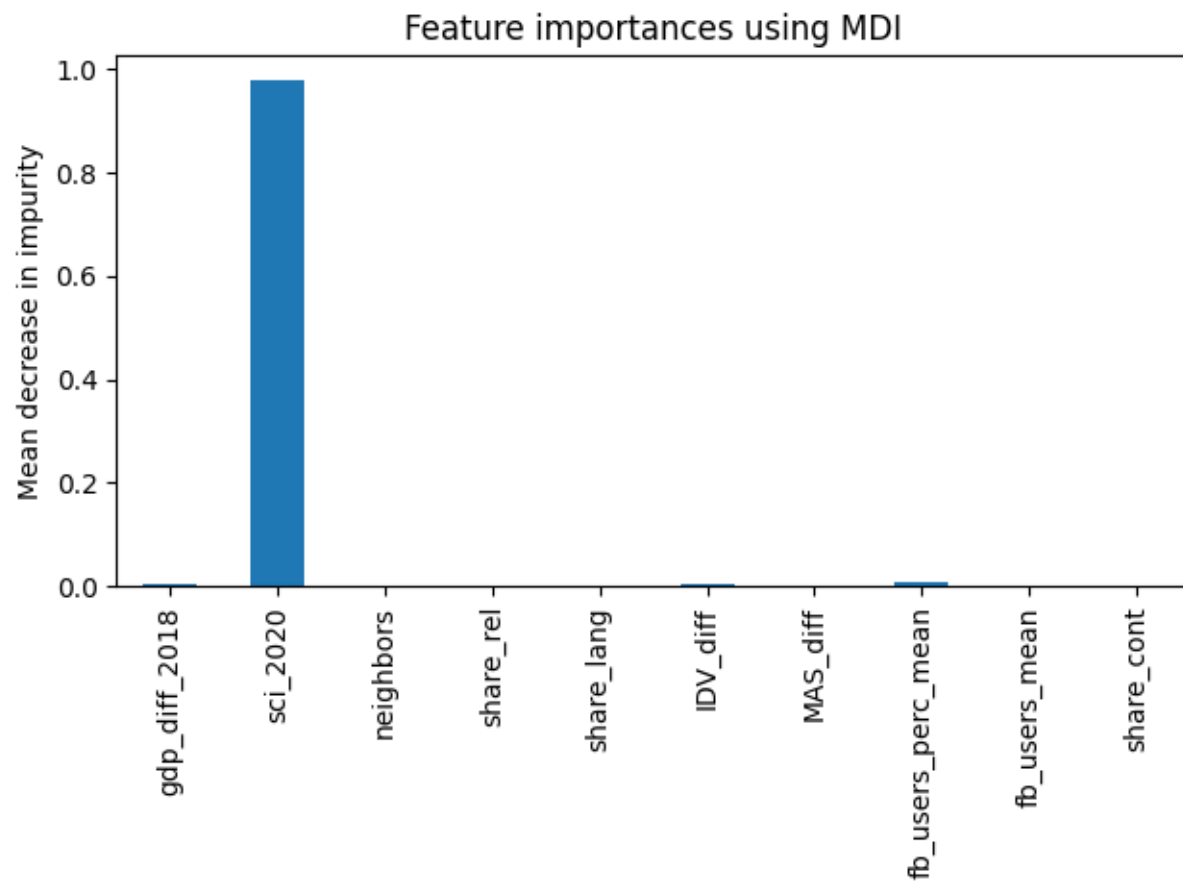


# Model comparison

	Metric	Learning (regression) model	
		Random Forest	Multi-layer Perceptron
Validation	MSE	0.00005	0.00003
	RMSE	0.007	0.006
	MAE	0.003	0.003
	R <sup>2</sup> (average on 5 folds)	0.85	0.84
Test	MSE	0.00003	0.00005
	RMSE	0.006	0.006
	MAE	0.002	0.003
	R <sup>2</sup> (best model)	0.88	0.73

# Feature importance

**Random Forest** model with best hyperparameter configuration.





# Conclusion

## Results

- Facebook strength of connectivity proved to be relevant for nowcasting country to country migration flows

## Limitations

- Concrete applications in filling the gaps of traditional migration data are still far away, due to the limitations of non-traditional sources (e.g., selection bias)

## Scheduled work

- Exploit a time window of past flows (e.g. 5 years of migration timeseries data) to predict future migration
- Nowcasting Unidirectional Migration Index (UMI)
- Further models: Support Vector Machin, XGBoost
- Restrict countries to FB penetration over a certain threshold

## Future improvements

- Nowcast / predict migration for a specific age group or by gender
- Include other non-traditional sources from Meta



## AI4MIG



Want more? Scan me!

## Contact

Diletta Goglia

Postgraduate Student in MSc in Artificial Intelligence

Computer Science department, University of Pisa, Italy

✉ [d.goglia@studenti.unipi.it](mailto:d.goglia@studenti.unipi.it)

🌐 [www.dilettagoglia.netlify.app](http://www.dilettagoglia.netlify.app)

🐦 [@dilettagoglia](https://twitter.com/dilettagoglia)

## COORDINATOR

HIVA - Research Institute for Work and Society, KU Leuven (BE)

## PARTNERS

University of Utrecht, Department of Information and Computing Sciences (NL) • University of Salamanca, Faculty of Social Sciences (ES) • Otto-Friedrich-University of Bamberg, Geographic Migration and Transition Studies (DE) • CESSDA ERIC (NO) • EKKE - National Centre for Social Research (GR) • IEN - Institute of Economic Sciences, Data Center Serbia for Social Sciences (RS) • Malmö University, Malmö Institute for Studies of Migration, Diversity and Welfare (SE) • Turkcell Technology (TR) • GMV NSL Limited (UK) • MPG - Migration Policy Group (BE) • ENoMW - European Network of Migrant Women (BE) • White Research (BE) • University of Pisa, Department of Computer Science (IT) • CNR - National Research Council (IT) • VUB - Vrije Universiteit Brussel, Interface Demography, Department of Sociology (BE) • University of Zurich, Department of Social Anthropology and Cultural Studies (CH) • CILD - Italian Coalition for Civil Liberties and Rights (IT) • Caritas International (BE)



UNIVERSITÀ DI PISA



Horizon2020  
European Union Funding  
for Research & Innovation

VISIT: <http://www.hummingbird-H2020.eu>

CONTACT US: [HumMingBird@kuleuven.be](mailto:HumMingBird@kuleuven.be)

FOLLOW US:

