PREDICTING WELL BEING
BASED ON FEATURES VISIBLE FROM SPACE - THE CASE OF WARSAW

PIOTR WÓJCIK AND KRYSTIAN ANDRUSZEK
UNIVERSITY OF WARSAW, FACULTY OF ECONOMIC SCIENCES
DATA SCIENCE LAB

BACKGROUND

- data from space (night-time lights intensity or high-resolution day-time images) are increasingly used as a **proxy of economic activities at the regional and local level** (Jean et al., 2016)
- can be **easily aggregated for any territorial units**
- **independent** of politicians and response rates in surveys
- proxies for **economic well-being** or **market potential** can be calculated for **non-administrative areas**

OBJECTIVES

- use features visible from space to proxy the level of well-being
- extract features from day-time satellite images
- apply various machine learning tools to link extracted features with the measure of well-being for administrative units
- predict the level of economic well-being or economic potential for non-administrative or poorly labelled areas
- apply on sample data for Warsaw
- **work in progress**

CONCLUSIONS

- semantic segmentation on satellite images requires **large computing power** and **good training data**
  - bad or outdated labels, images taken at an angle cause problems
  - model predicts long buildings better
- green areas can be extracted better through cutting color channels
- districts are too big administrative areas for model training but they are the lowest level for which well-being/socio-economic data is available (model outcome)

FURTHER STEPS

- label images manually to obtain better training sample
- consider classification instead of regression problem
- add new features to the model, e.g. building height, roof type, building type (offices, houses, apartment blocks)
- obtain data over time to observe changes
- extend the analysis to other(rural) areas since metropolies are relativelly homogenous
- combine features extracted from daytime images and night-time light intensity

REFERENCES

Jean, N., Burke, M., Xie, M., Davis, W., Lobell, D. and Ermon, S. (2016), "Combining satellite imagery and machine learning to predict poverty", *Science* , Vol. 353, pp. 790–794.

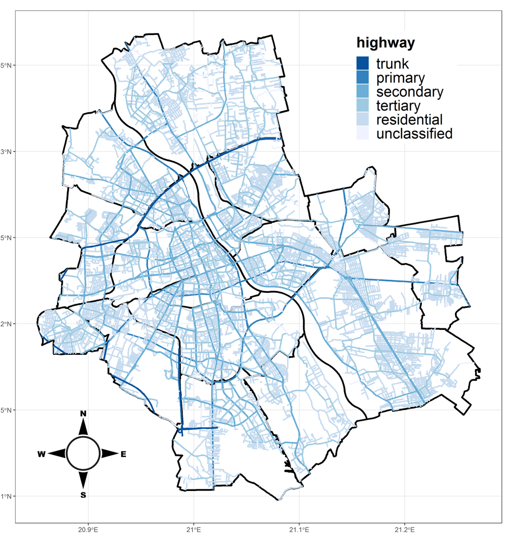
DATA AND METHODS

- **Open Street Map**: localisation of buildings, streets, rivers and lakes, green areas, public transport, bike rental, fuel stations, supermarkets and malls were used as: (1) model features; (2) inputs for labeling images
- **Google Maps**: 6000+ high-resolution daytime images of Warsaw (zoom 16)
- **socio-economic indicators** for Warsaw districts (budget spendings, income, share in PIT/CIT, population, no of vehicles, life expectancy)
- **Convolutional Neural Networks, transfer learning**: U-net and ResNet101 architectures pretrained on ImageNet used for semantic segmentation
- selected machine learning tools (**Random forest, LASSO**) applied to find relationships between features visible from space and economic well-being indicators

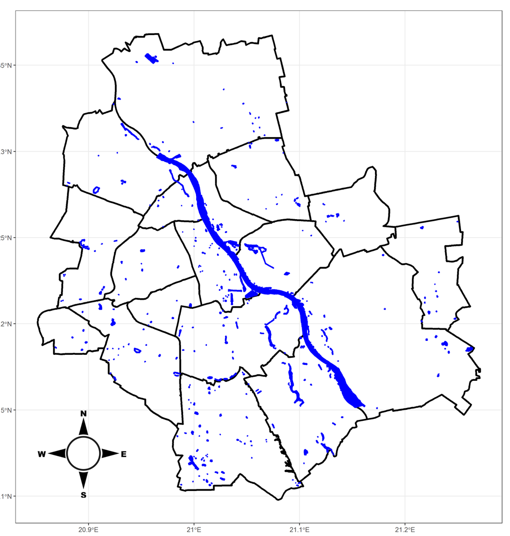
RESULTS

OSM DATA FOR LABELING

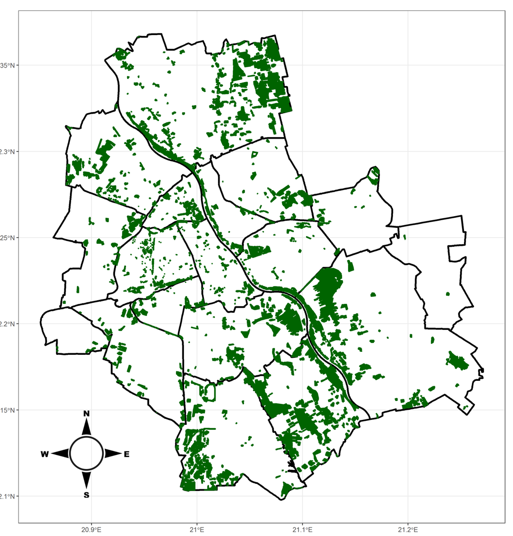
buildings



streets by type



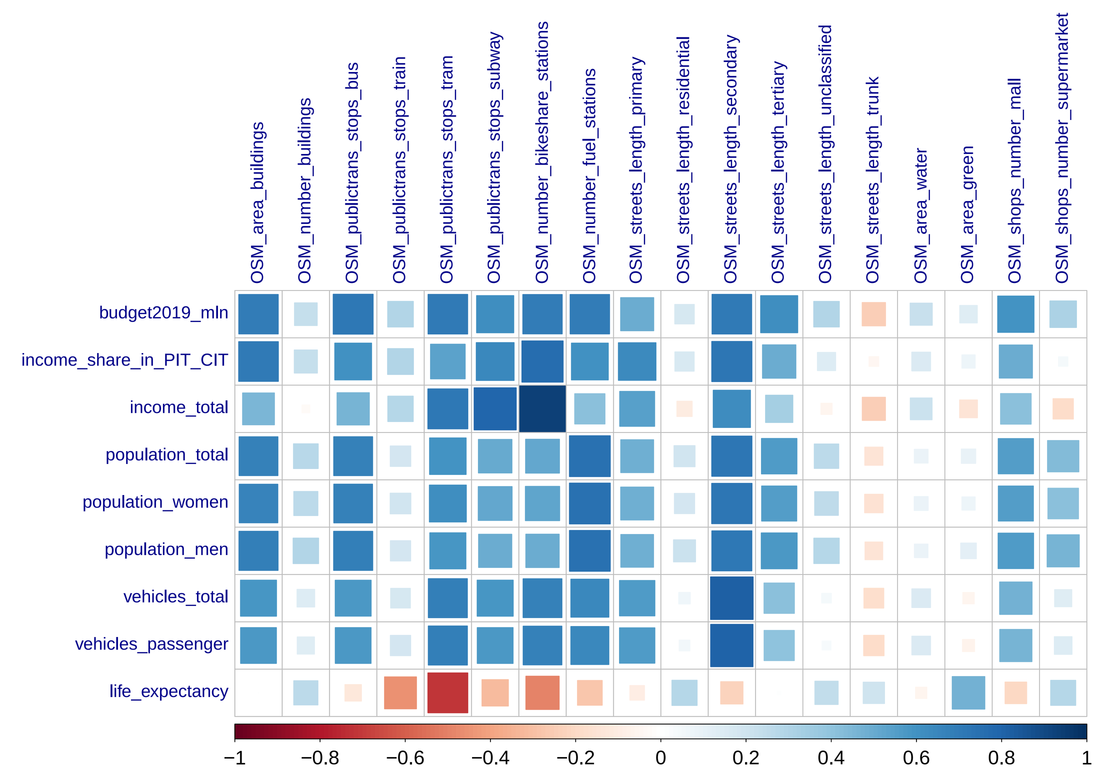
water



green areas

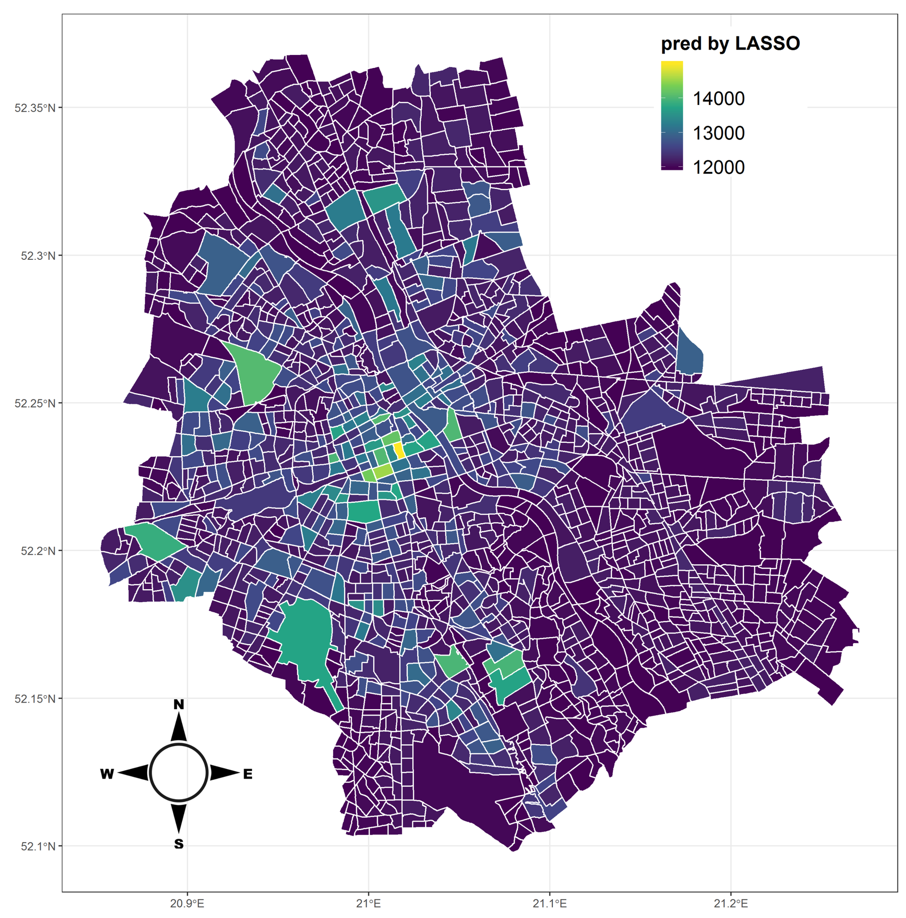
OSM BASED LABELING APPLIED

CORRELATION OF EXTRACTED FEATURES AND WELL-BEING INDICATORS



	variable	budget2019_mln	income_share_in_PIT_CIT	income_total	population_total	population_women	population_men	vehicles_total	vehicles_passenger	life_expectancy
LASSO	OSM_number_bikeshare_stations		1	2	6	4		2	2	5
	OSM_number_fuel_stations			4	5	1	1	4	4	8
	OSM_publictrans_stops_bus				7	3	2			7
	OSM_publictrans_stops_tram				3	2				4
	OSM_shops_number_mall			6	2			1	1	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings		3	13	14					14
	OSM_area_green				13			8	8	16
	MAE	197.25	19270.77	52750.28	47211.16	23124.53	19630.11	33544.59	29918.29	4.62
	MAPE	0.649	0.600	0.846	0.539	0.632	0.636	0.550	0.588	0.060
LASSO (w/ pred variables)	OSM_number_bikeshare_stations		2	2	4	4		1	2	6
	OSM_number_fuel_stations		6	4	3	1	1	4	4	8
	OSM_publictrans_stops_bus				6	3	2			7
	OSM_publictrans_stops_tram		7		5	2		5	5	4
	OSM_shops_number_mall		1					2	1	5
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings_pred		15							15
	OSM_area_green_pred							9	9	16
	MAE	197.25	32043.48	56837.23	40367.19	23124.53	19553.72	34638.79	30788.64	4.28
	MAPE	0.649	1.026	0.846	0.526	0.632	0.633	0.570	0.598	0.056

Feature importance in models



Predictions of LASSO for districts

CONTACT INFORMATION

web wne.uw.edu.pl/pwojcik , dslab.wne.uw.edu.pl
email pwojcik@wne.uw.edu.pl

Research financed by Polish National Science Center, project no. 2016/21/B/HS4/00670