

Features of districts of Warsaw visible from space

Krystian Andruszek, Ewa Sobolewska, Piotr Wójcik
Data Science Lab, dslab.wne.uw.edu.pl

WhyR! 2019

Warsaw

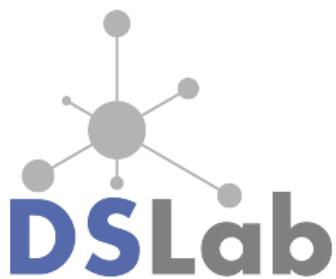
September 28-29, 2019



UNIVERSITY OF WARSAW

Faculty of Economic Sciences

About Data Science Lab



Data Science Lab is an organization created in 2019 at the Faculty of Economics of the University of Warsaw.

We are a group consisting of academic workers, business professionals and students who were gathered together by our passion for uncovering the unknown and discovering practical applications to new tools and methods in data analysis, Machine Learning and Artificial Intelligence.



Aim of the project

- use features visible from space to proxy the level of well-being
- apply models for non-administrative or poorly labelled areas to predict economic well-being or economic potential
- work in progress



Measuring inequalities from space

- night-time lights intensity used but imperfect (e.g. scale limitations)
- alternative: high-resolution daytime satellite pictures



Economic and business applications

- can be easily **aggregated for any territorial units**
- **independent** of politicians and response rates in surveys
- increasingly used as a **proxy of economic activities** at the regional and local level (e.g. Nature, 2017)
- it has informational value for **countries with poor quality of national income accounts**
- proxies for **economic well-being** or **market potential** can be calculated for **non-administrative areas** (e.g. **specific business regions**)

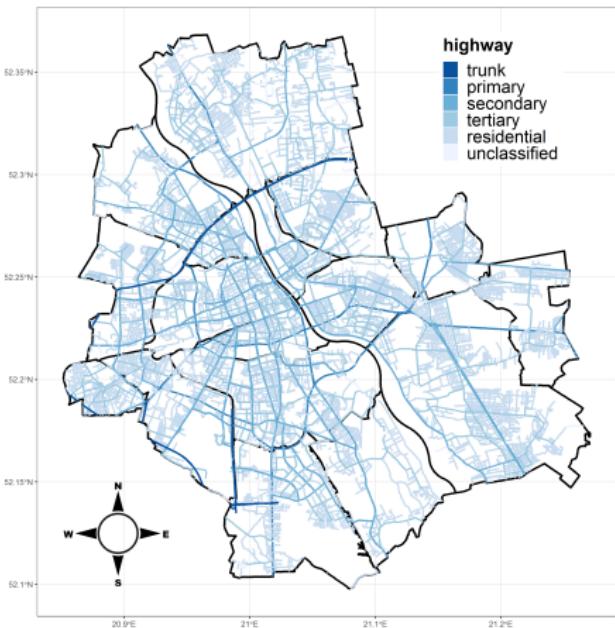
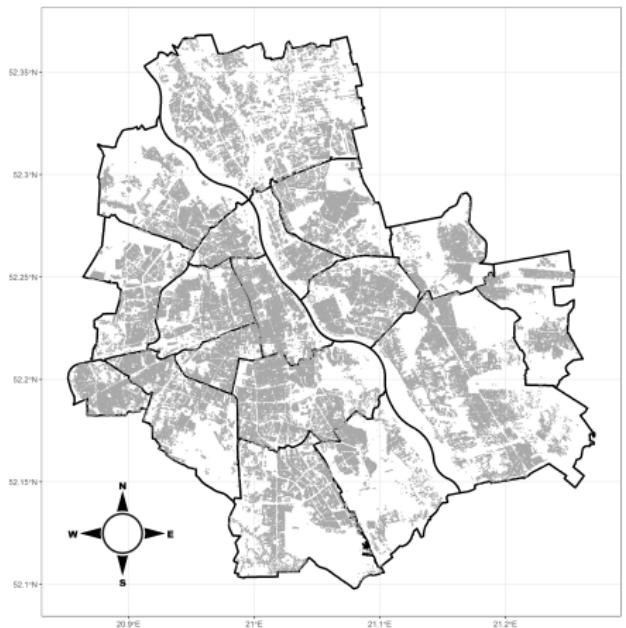


Data sources

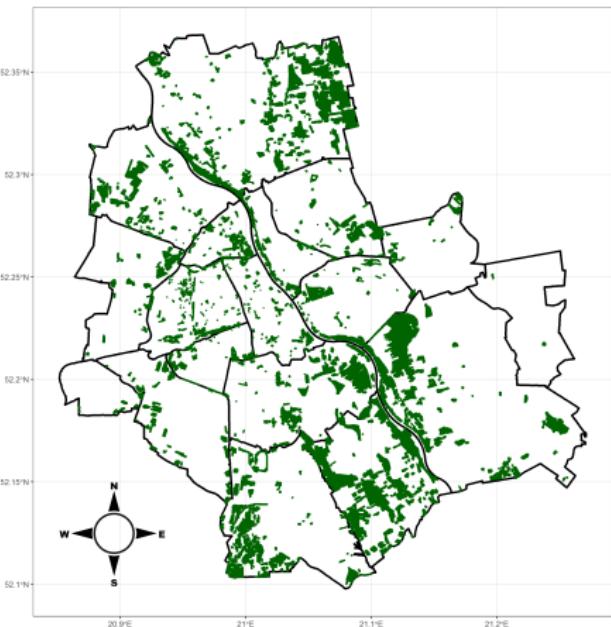
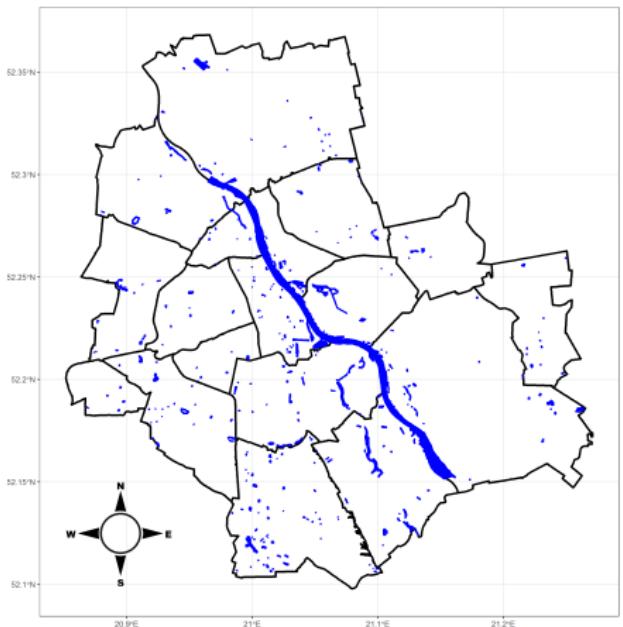
- Open Street Map – localisation of buildings, streets, rivers and lakes, green areas, public transport, bike rental, fuel stations, supermarkets and malls
- Google Maps: 6000+ high-resolution daytime images of Warsaw from space
- socio-economic indicators for Warsaw districts:
 - budget 2019 – real spendings
 - Panorama of Warsaw districts (2017) – total income, income share in PIT/CIT, population (total, men, women), vehicles (total, passenger)
 - report “Health condition of Warsaw residents” (2016) – life expectancy



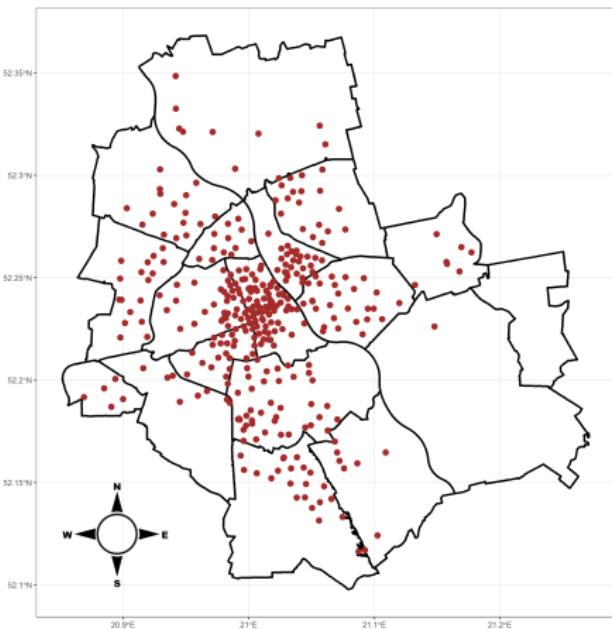
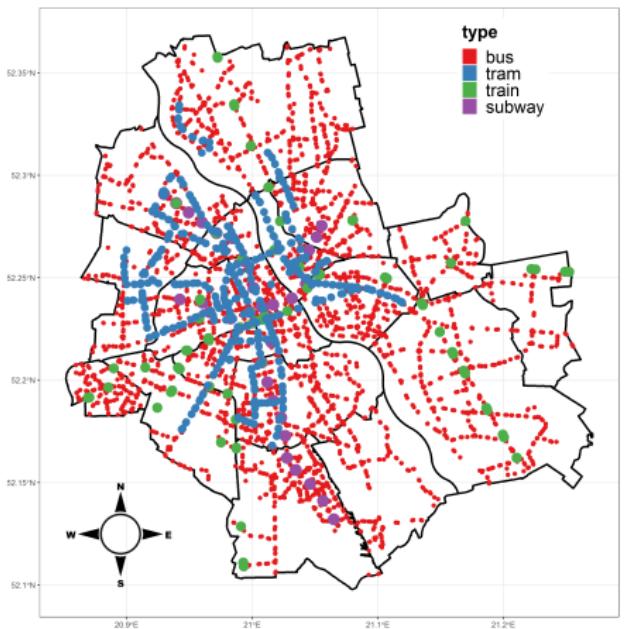
OSM buildings and streets by type



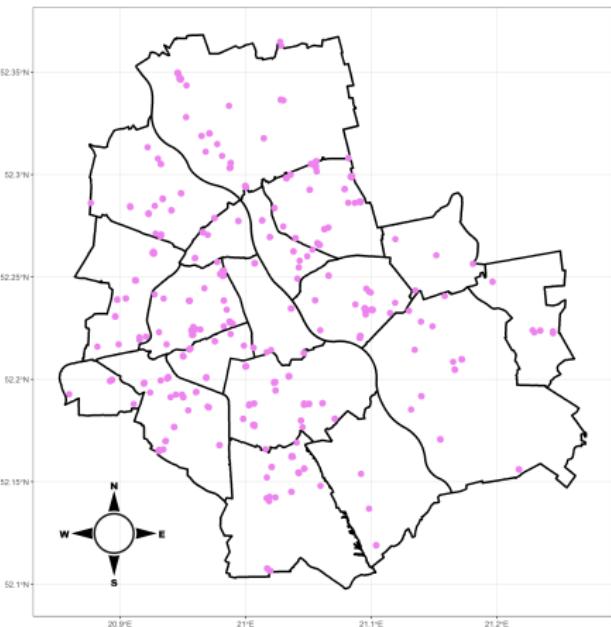
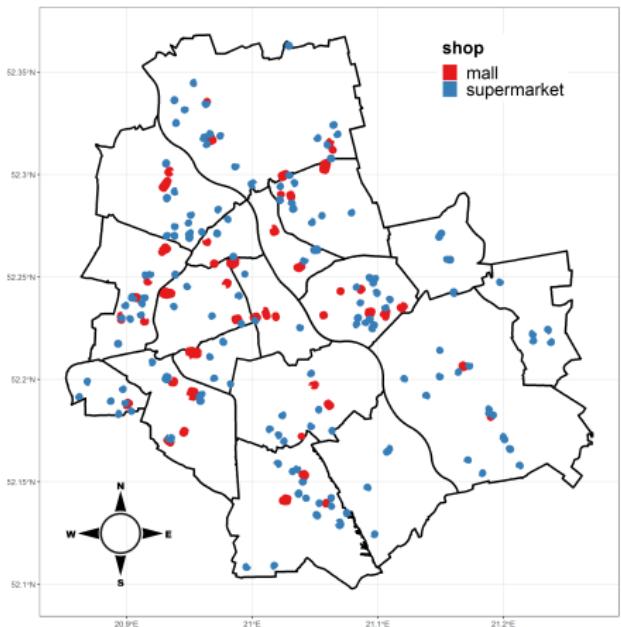
OSM water and green areas



OSM public transport (stops by type) and bike rental (Veturilo stations)



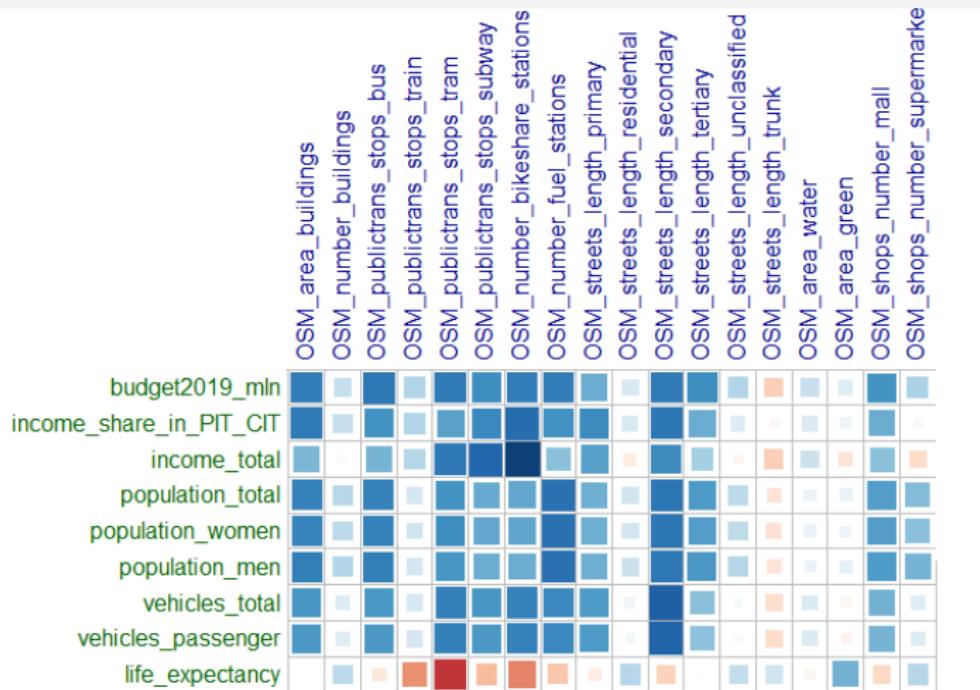
OSM supermarkets and fuel stations



Identifying objects on satellite images



Correlation between OSM extracted features and indicators



-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1



OSM feature importance based on LASSO

	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

OSM feature importance based on random forest

variable		budget2019_mln	income_share_in_PIT_CIT	income_total	population_total	population_women	population_men	vehicles_total	vehicles_passenger	life_expectancy
RF	OSM_number_bikeshare_stations	1	3	2	1	1	1	2	1	1
	OSM_number_fuel_stations	5	10	9	5	4	6	5	5	11
	OSM_publictrans_stops_bus	4	9	8	4	5	4	6	6	13
	OSM_publictrans_stops_tram	6	2	3	10	10	12	4	4	4
	OSM_shops_number_mall	3	4	6	3	3	3	8	7	17
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings	7	7	11	7	7	7	9	9	14
	OSM_area_green	12	16	17	15	15	15	15	16	10
	MAE	128.39	18053.00	81590.29	34313.99	18808.38	15505.85	35024.39	29581.98	1.70
	MAPE	0.390	0.605	1.352	0.518	0.532	0.501	0.678	0.711	0.022



Feature extraction from satellite images

- Features extracted from OSM were used to label objects on satellite images.
- Buildings were recognized in pictures using deep neural networks.
- Transfer learning was performed based on models pre-trained on Imagenet.
- Dataset was split proportionally on district level.
- Better estimations of green areas were calculated.



Types of problems in computer vision

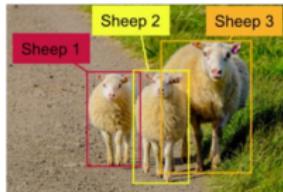
Image classification



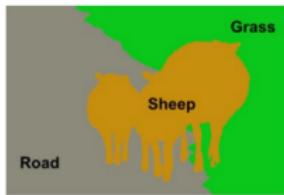
Classification with Localization



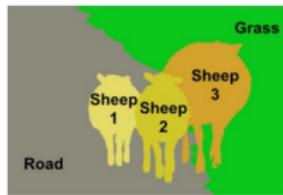
Object detection



Semantic segmentation



Instance segmentation



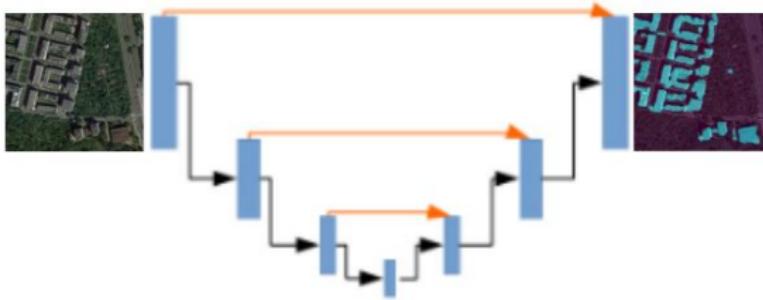
Transfer learning

- Problem: Training a neural network requires time and powerful computational resources.
- Possible solution: Transfer learning
 - The model is not trained from scratch, but uses a model pre-trained on a large benchmark dataset to solve a similar problem.

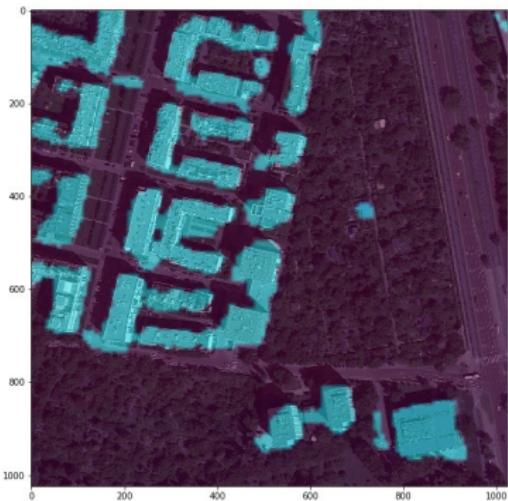
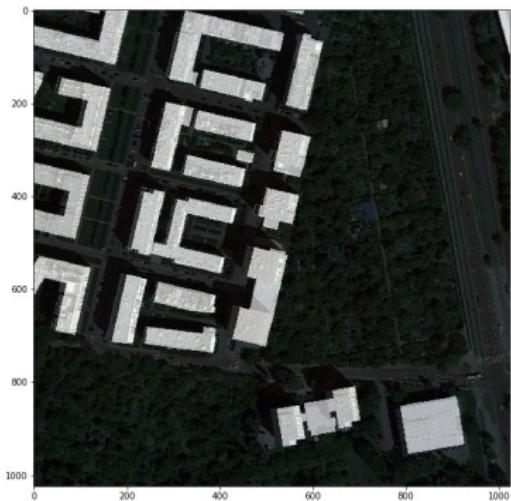


Unet and ResNet101

- ResNet is one of the known, state-of-the-art architectures with an open source pretrained model.
 - It is used to classify images, resulting in a single value.
- By removing the top dense layers of the network, model outputs a vector of values.
- In a U-Net, an image is converted into a vector and then the same mapping is used to convert it again to an image.
 - It is used for classification per pixel, not per image.



True and predicted images



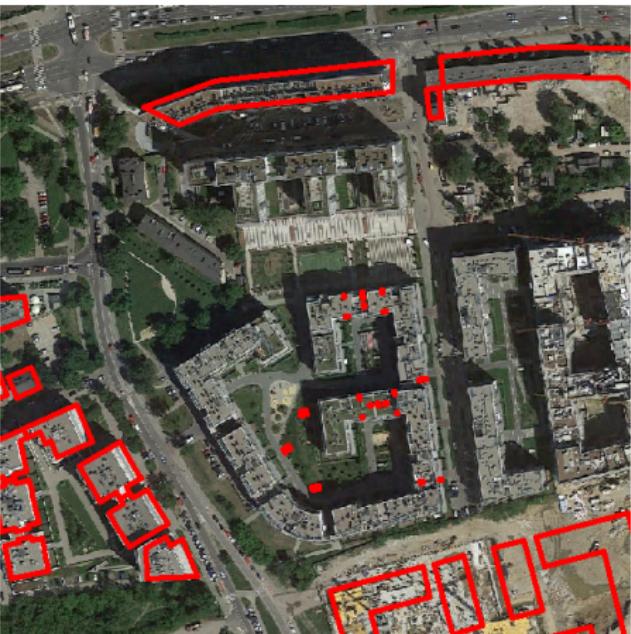
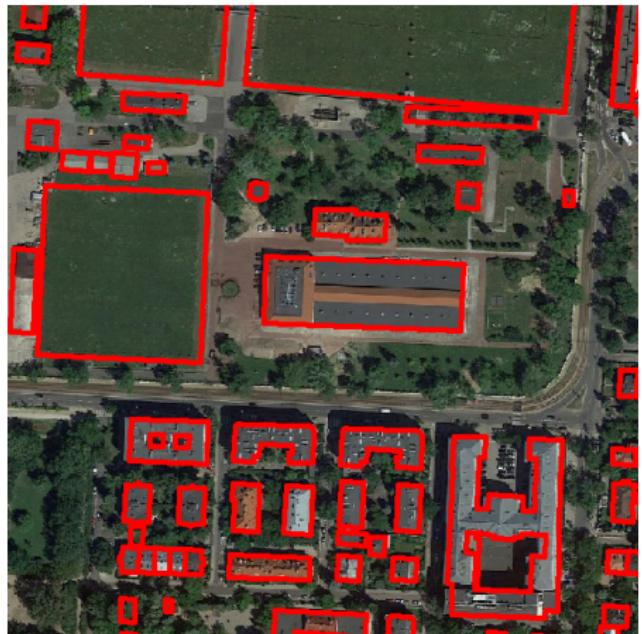
Preliminary results

District	Corr	District	Corr
Żoliborz	0.85	Targówek	0.69
Śródmieście	0.85	Ochota	0.68
Rembertów	0.83	Mokotów	0.68
Praga-Południe	0.75	Wilanów	0.64
Bielany	0.75	Ursynów	0.64
Białołęka	0.74	Wawer	0.59
Wesoła	0.74	Praga-Północ	0.59
Włochy	0.74	Wola	0.55
Bemowo	0.73	Ursus	0.19

- Predicted area of buildings was calculated.
- Correlations between true and predicted area percentage values look promising.
- It could look better but we encountered some obstacles...

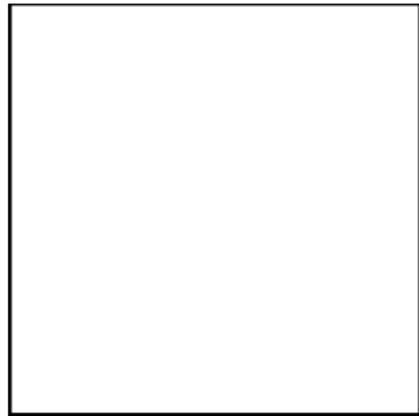


Examples of problematic images



Improving green area labels

- Green areas in OSM are poorly labelled.
- It can be done better by extracting only pixels in chosen color channel ranges.



OSM feature importance based on LASSO + predicted variables

	variable	budget2019_min	income_share_in_PT_Ctr	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



OSM feature importance based on random forest + predicted variables

variable	budget2019_mln	income_share_in_PIT_CIT	income_total	population_total	population_women	population_men	vehicles_total	vehicles_passenger	life_expectancy
Random Forest	OSM_number_bikeshare_stations	1	3	2	1	1	1	2	1
	OSM_number_fuel_stations	5	10	9	5	4	6	5	5
	OSM_publictrans_stops_bus	4	9	8	4	5	4	6	13
	OSM_publictrans_stops_tram	6	2	3	10	10	12	4	4
	OSM_shops_number_mall	3	4	6	3	3	3	8	17
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings	7	7	11	7	7	7	9	9
	OSM_area_green	12	16	17	15	15	15	15	10
	MAE	128.39	18053.00	81590.29	34313.99	18808.38	15505.85	35024.39	29581.98
	MAPE	0.390	0.605	1.352	0.518	0.532	0.501	0.678	0.711
RF { w/ pred variables }	OSM_number_bikeshare_stations	1	2	2	1	1	1	2	2
	OSM_number_fuel_stations	5	7	10	3	3	4	4	12
	OSM_publictrans_stops_bus	4	9	7	5	5	5	7	7
	OSM_publictrans_stops_tram	7	3	3	11	9	11	5	5
	OSM_shops_number_mall	3	5	9	4	4	3	8	17
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings_pred	13	16	16	14	14	14	12	14
	OSM_area_green_pred	15	14	6	16	16	16	16	2
	MAE	129.64	18594.46	77569.34	35042.93	19080.28	15920.14	34132.85	28925.08
	MAPE	0.400	0.651	1.267	0.524	0.537	0.510	0.663	0.694



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 oblong buildings better
- green areas can be extracted better through cutting color channels
- districts are too big administrative areas but they are the lowest level of financial data publicly available



Further steps

- obtain well-being data on lower level than district, e.g. postal code
- change prediction to classification instead of regression
- add new variables to model, e.g. building height, roof type, building type (offices, houses, apartment blocks)
- obtain data over time to observe changes
- broaden analysis to other areas since metropolies are homogenous
- use model on areas that have even worse labelled data



Thank you

THANK YOU FOR YOUR ATTENTION!



OSM feature importance based on LASSO

variable									
	budget2019_mln	income_share_in_PIT_CIT	income_total	population_total	population_women	population_men	vehicles_total	vehicles_passenger	life_expectancy
OSM_area_buildings	3	13	14						14
OSM_area_green			13				8	8	16
OSM_area_water		12	15						15
OSM_number_bikeshare_stations	1	2	6	4			2	2	5
OSM_number_buildings									9
OSM_number_fuel_stations		4	5	1	1	4	4	4	8
OSM_publictrans_stops_bus			7	3	2				7
OSM_publictrans_stops_subway		3							6
OSM_publictrans_stops_train		5	4			3	3	3	2
OSM_publictrans_stops_tram			3	2					4
OSM_shops_number_mall		6	2			1	1	1	1
OSM_shops_number_supermarket		1	1						3
OSM_streets_length_primary		7	12			7	7	7	10
OSM_streets_length_residential				11					
OSM_streets_length_secondary	2	11	10	5	3	5	5	5	12
OSM_streets_length_tertiary		10							
OSM_streets_length_trunk		9	9			6	6	6	13
OSM_streets_length_unclassified		8	8						11
lambda	162.38	10000.00	753.69	364.25	10000.00	10000.00	4111.83	3498.32	0.01
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



OSM feature importance based on random forest

variable	budget2019_mln	income_share_in_PIT_CIT	income_total	population_total	population_women	population_men	vehicles_total	vehicles_passenger	life_expectancy
OSM_area_buildings	7	7	11	7	7	7	9	9	14
OSM_area_green	12	16	17	15	15	15	15	16	10
OSM_area_water	9	12	10	13	12	13	11	11	3
OSM_number_bikeshare_stations	1	3	2	1	1	1	2	1	1
OSM_number_buildings	11	11	12	6	8	5	10	10	5
OSM_number_fuel_stations	5	10	9	5	4	6	5	5	11
OSM_publictrans_stops_bus	4	9	8	4	5	4	6	6	13
OSM_publictrans_stops_subway	14	5	4	16	16	16	12	12	16
OSM_publictrans_stops_train	17	13	13	17	17	17	16	15	2
OSM_publictrans_stops_tram	6	2	3	10	10	12	4	4	4
OSM_shops_number_mall	3	4	6	3	3	3	8	7	17
OSM_shops_number_supermarket	15	17	14	11	11	10	17	17	12
OSM_streets_length_primary	8	1	1	9	9	9	3	3	8
OSM_streets_length_residential	16	14	16	14	14	14	14	13	6
OSM_streets_length_secondary	10	6	5	8	6	8	1	2	9
OSM_streets_length_tertiary	2	8	7	2	2	2	7	8	15
OSM_streets_length_trunk									
OSM_streets_length_unclassified	13	15	15	12	13	11	13	14	7
mtry	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MAE	128.39	18053.00	81590.29	34313.99	18808.38	15505.85	35024.39	29581.98	1.70
MAPE	0.390	0.605	1.352	0.518	0.532	0.501	0.678	0.711	0.022



LASSO prediction on map (postal code level)

