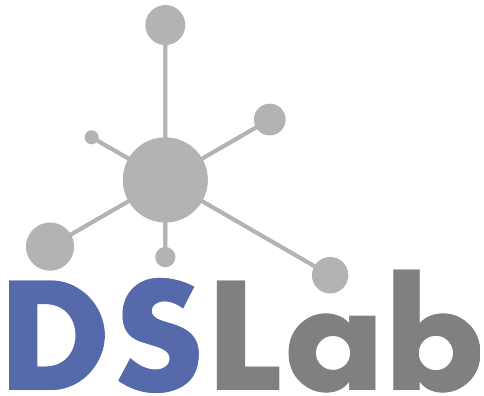
Predicting well-being based on features visible from space

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

QFRG and DSLab joint seminar
WNE UW, November 19th, 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.

DSLab

Aim of the project

- use features visible from space to proxy the level of well-being
- apply machine learning tools using indicators of well-being for administrative areas as outcomes
- apply models for non-administrative areas to predict economic well-being or economic potential
- work in progress

Economic development

- GDP *per capita* or Human Development Index (including also life expectancy at birth and scholarisation) are commonly used measures of **economic development** on national and regional level
- however they **cannot be measured on a local level**
- there are attempts to create **synthetic indices** based on several dimensions – for Poland see e.g. UNDP (2012), Sompolska-Rzechuła (2016), Pomianek (2016), Ciołek (2017)
- one might also try to spread GDP from regional to local level according to (working) population, potentially correcting for commuting to work

NTL intensity as a proxy of development

- **NTL intensity** is increasingly used as a proxy of economic activities on regional and local leves (3000 studies since 2000 according to Nordhaus and Chen, 2010)
- Doll et. al (2006) find a **strong positive relationship** between NTL and GDP across a range of spatial levels for 11 EU countries USA
- Henderson et al. (2009) use NTL to measure **true GDP** of 188 countries over 17 years
- Gennaioli et al. (2013) find that NTL are **related to human capital** similarly as regional income p.c. for a **large sample of regions from different countries**
- Chen and Nordhaus (2011) argue that **NTL has informational value** for countries with **poor quality of national income accounts**
- Bickenbach et al. (2015) show that the relationship between the growth of NTL intensity and regional GDP growth is **unstable on subnational level**

Economic and business applications of NTLI data

- can be easily **agregated for any territorial units**
- **uniformly measured** across the globe
- **independent** of politicians and response rates in surveys
- increasingly used as a **proxy of economic activities** at the regional and local level
- researchers find **strong positive relationship** between NTLI and GDP, and population at the **national level**
- **at the subnational level** the relationship is usually **weaker and unstable**
- 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. **specicic business regions**)

Night-time lights intensity (NTLI) data

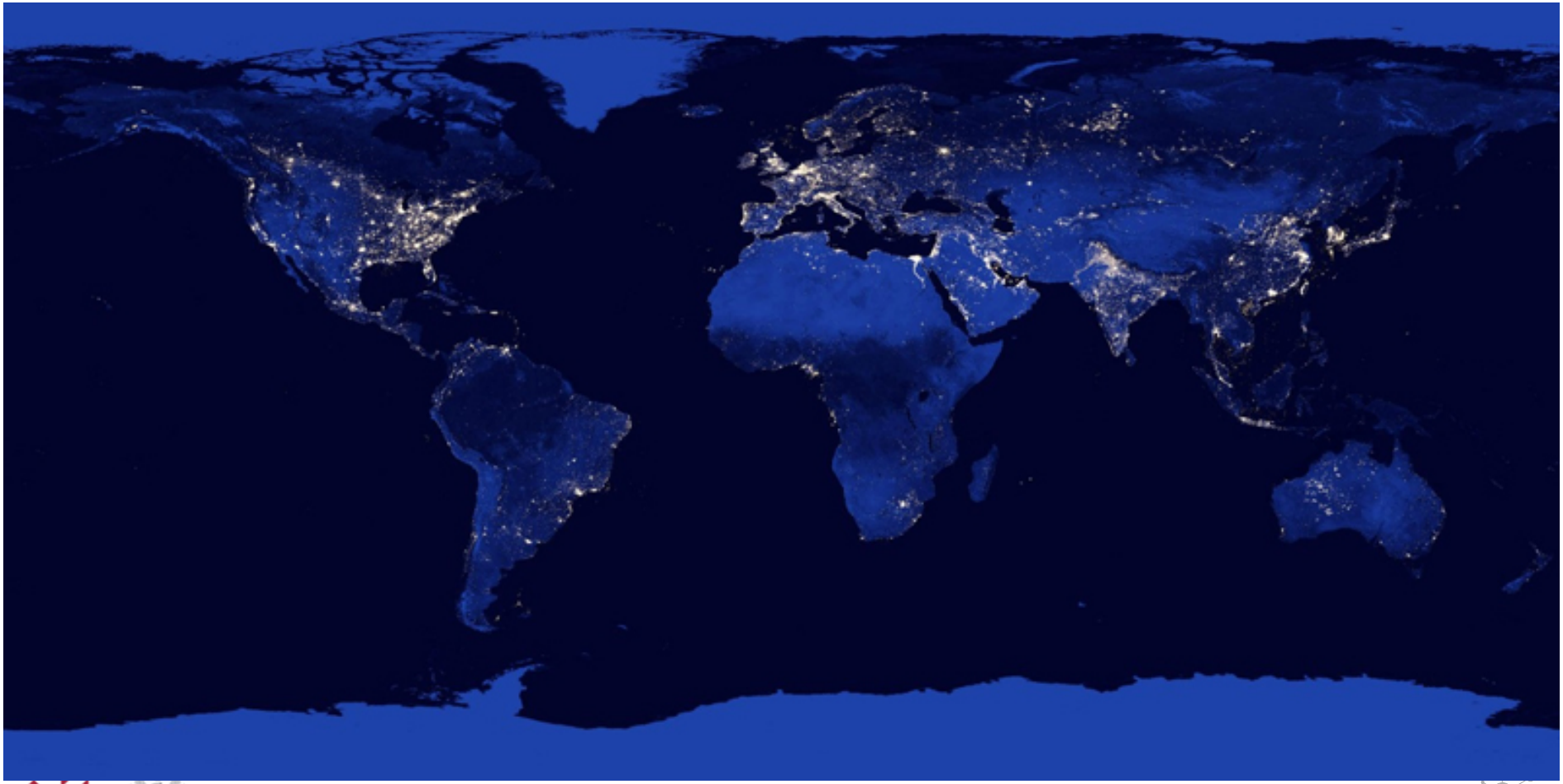
- NTLI data is based on **satellite images** collected and processed by the National Oceanic and Athmosferic Administration (NOAA)
- NOAA provides **two types** of NTLI data:
  - Version 4 **DMSP-OLS** – average annual data for the period 1992–2013
  - Version 1 **VIIRS** – monthly data since April 2012 and averaged annual (only 2015 and 2016)
- NTLI is measured for **pixels** with the size of 30×30 (DMSP-OLS) or 15x15 (VIIRS) arc seconds
- it relates to **less than 1 km2** on the equator (about 0.5 km2 in Europe or USA)
- for each pixel NTLI data is provided in **digital numbers** (DN) on the scale 0–63 (DMSP-OLS) or 0–16384 (VIIRS)

NTLI – two important limitations

- **first**, DN are often top-coded at 63 in the **centers of metropolitan areas** due to calibration of the sensors which allows to detect very low levels of illuminance
- therefore there is **no possibility to distinguish between bright urban centers and peripherial areas** – see Letu et al. (2012)
- **second**, permanent light sources of **low intensity** might be inappropriately **set to zero** by the processing steps involved in the preparation of the Stable Lights series
- this was indicated by Henderson et al. (2012), who observed that pixels with DN 1 and 2 are **underrepresented** in the data
- therefore there is a question if DN=0 should be perceived as complete absence of light emissions or rather very low light emissions that were filtered away

Sample analysis presented on UseR! 2019

- DMSP-OLS data for 2013 used in examples
- **full codes available on**: github.com/ptwojcik/UseR2019



Alternative measuring inequalities from space

Science Contents News Careers Journals

SHARE

RESEARCH ARTICLE

Combining satellite imagery and machine learning to predict poverty

Neal Jean[1,2,*], Marshall Burke[3,4,5,*,†], Michael Xie[1], W. Matthew Davis[4], David B. Lobell[3,4], Stefano Ermon[1]
+ See all authors and affiliations

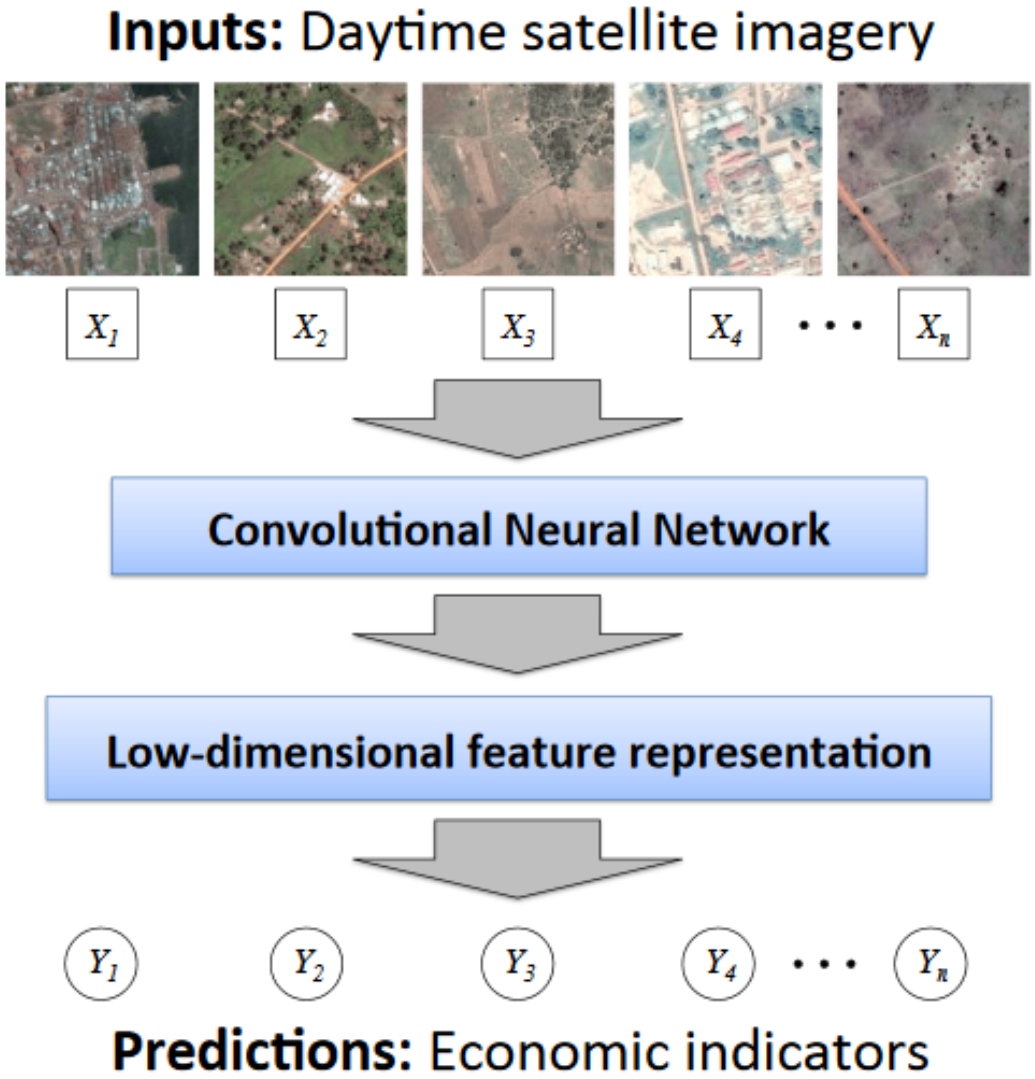
Science 19 Aug 2016:
Vol. 353, Issue 6301, pp. 790-794
DOI: 10.1126/science.aaf7894

Article Figures & Data Info & Metrics eLetters PDF

Measuring consumption and wealth remotely

Nighttime lighting is a rough proxy for economic wealth, and nighttime maps of the world show that many developing countries are sparsely illuminated. Jean *et al.* combined nighttime maps with high-resolution daytime satellite images (see the Perspective by Blumenstock). With a bit of machine-learning wizardry, the combined images can be converted into accurate estimates of household consumption and assets, both of which are hard to measure in poorer countries.

Alternative measuring inequalities from space

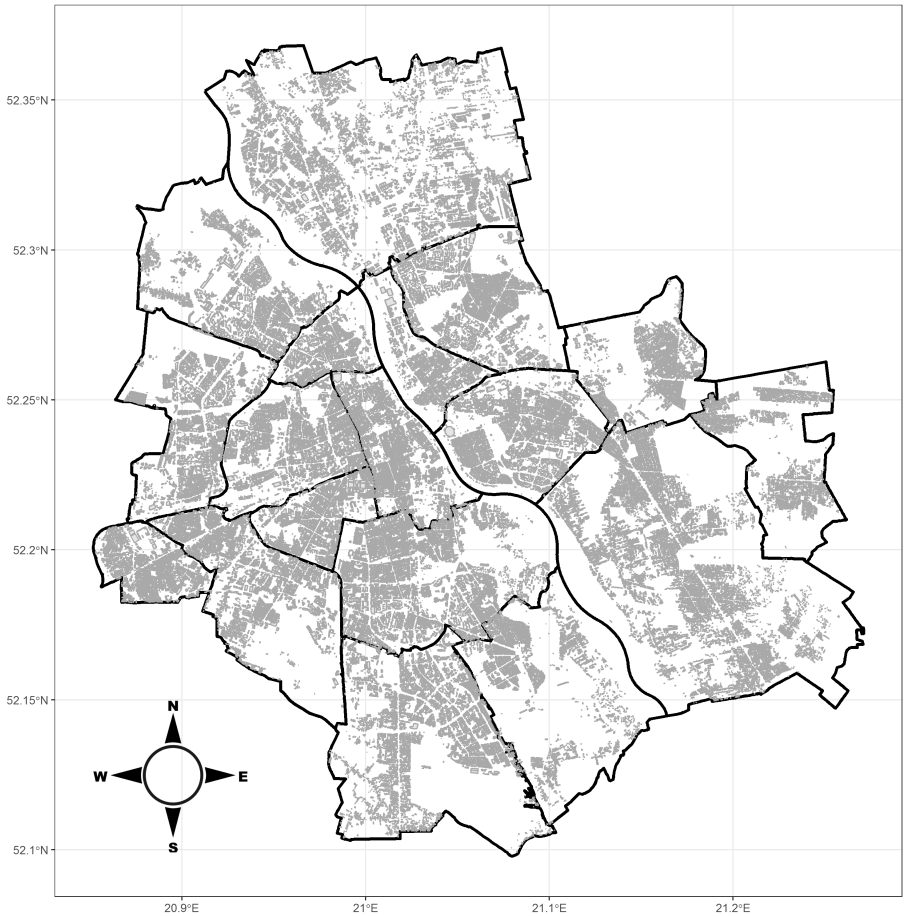


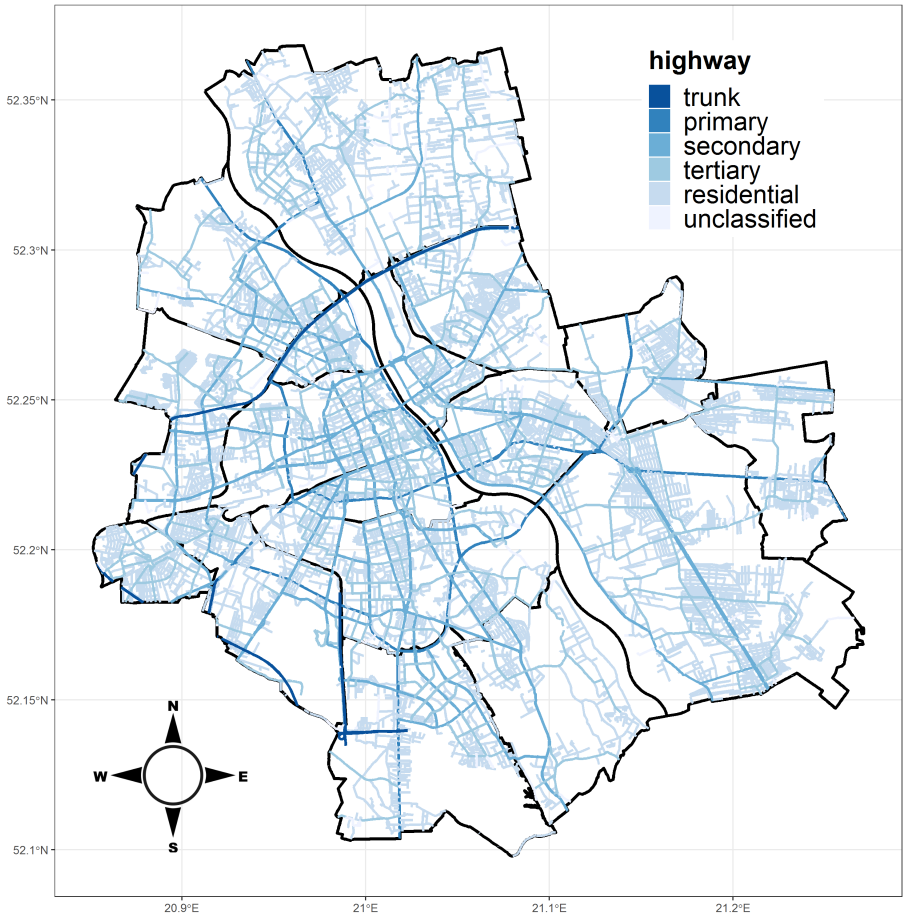
Data sources used in this project

- Open Street Map – localisation of buildings, streets, rivers and lakes, green areas, public transport, bike rental, fuel stations, supermarkets and malls (see: https://wiki.openstreetmap.org/wiki/Map_Features)
- 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

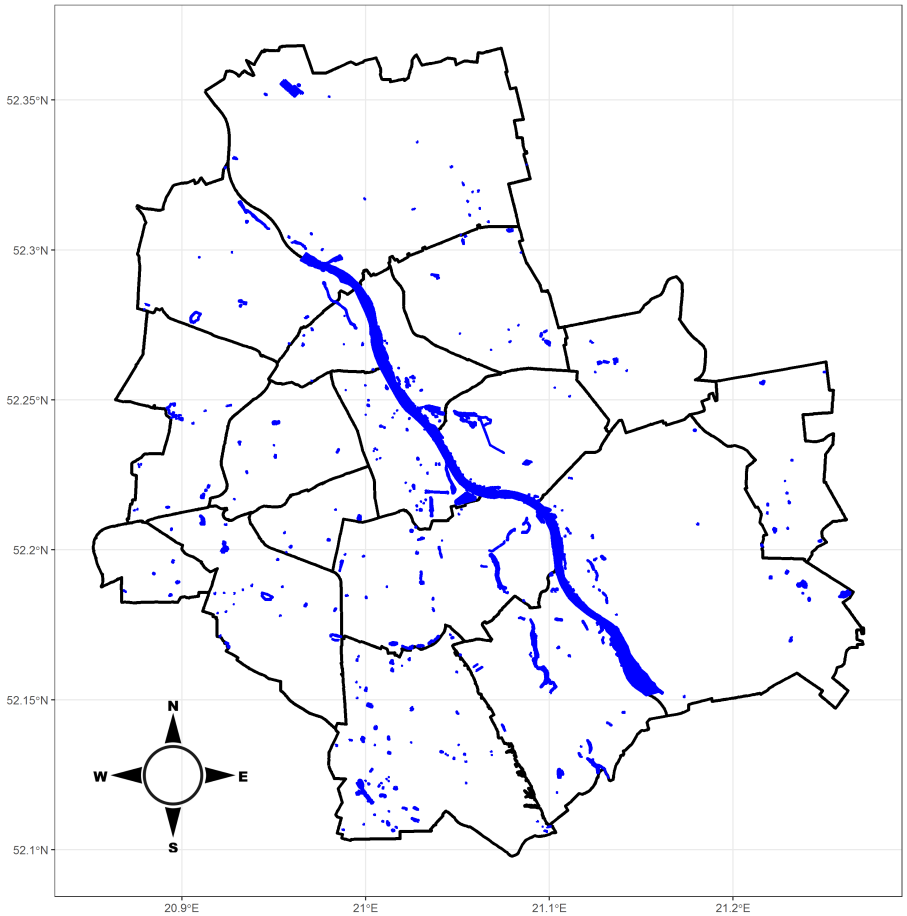
Identifying objects on satellite images

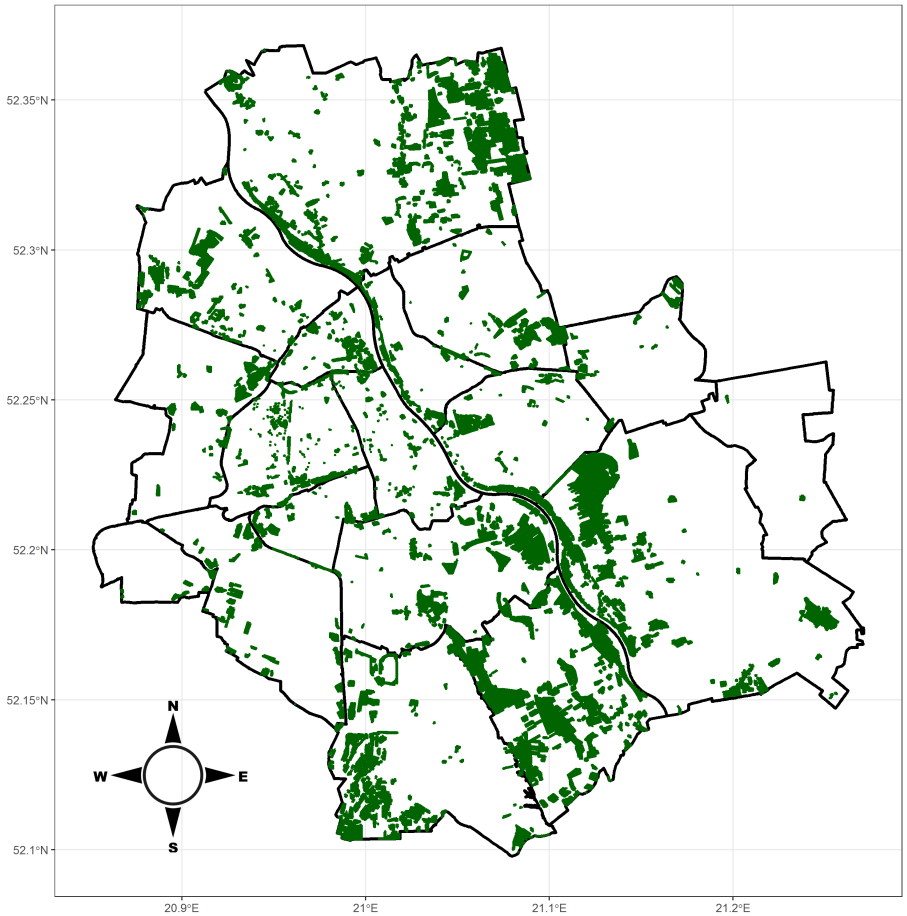
OSM buildings and streets by type



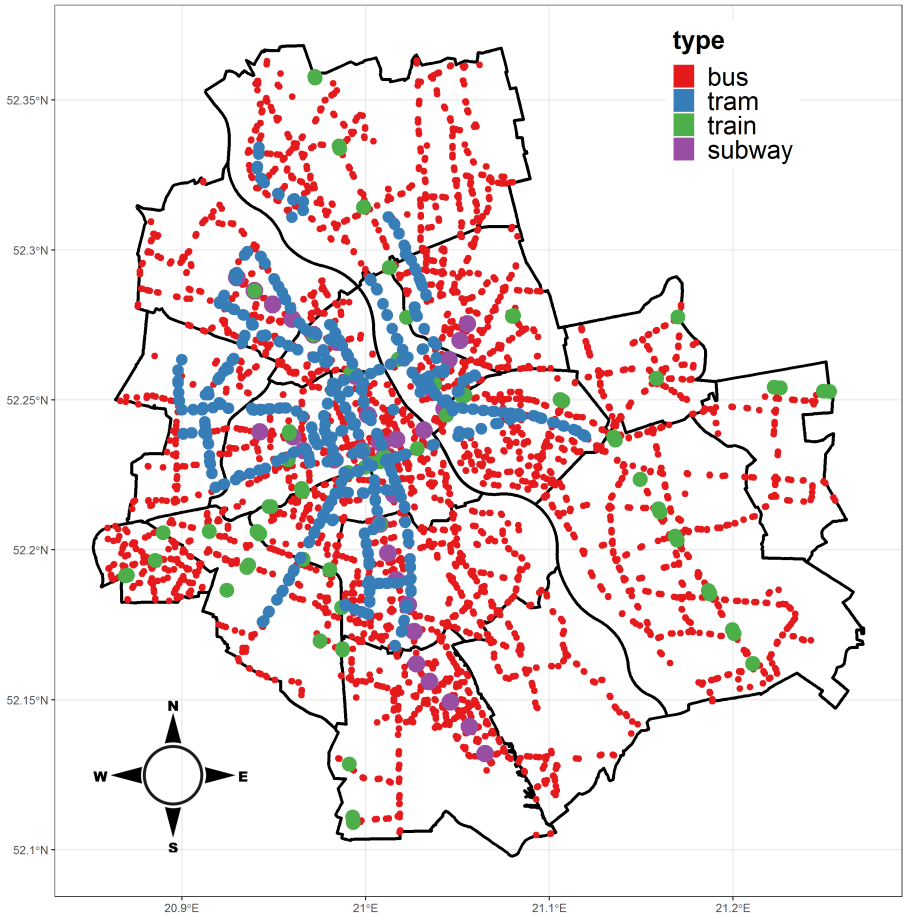


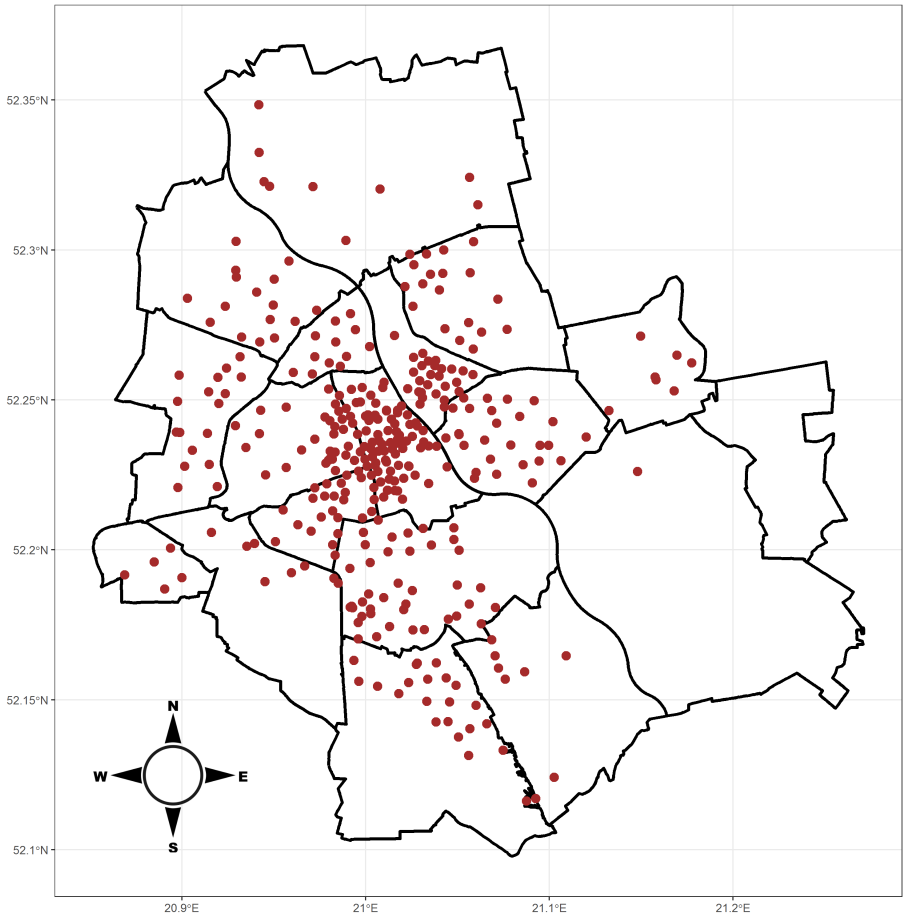
OSM water and green areas



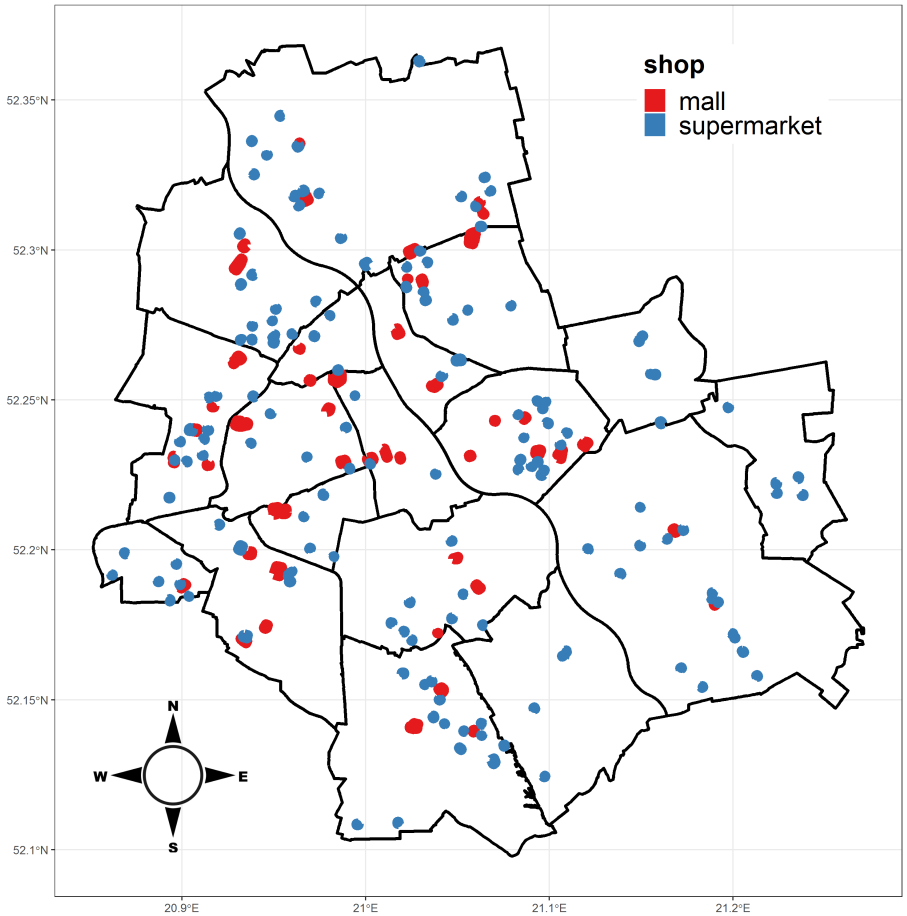


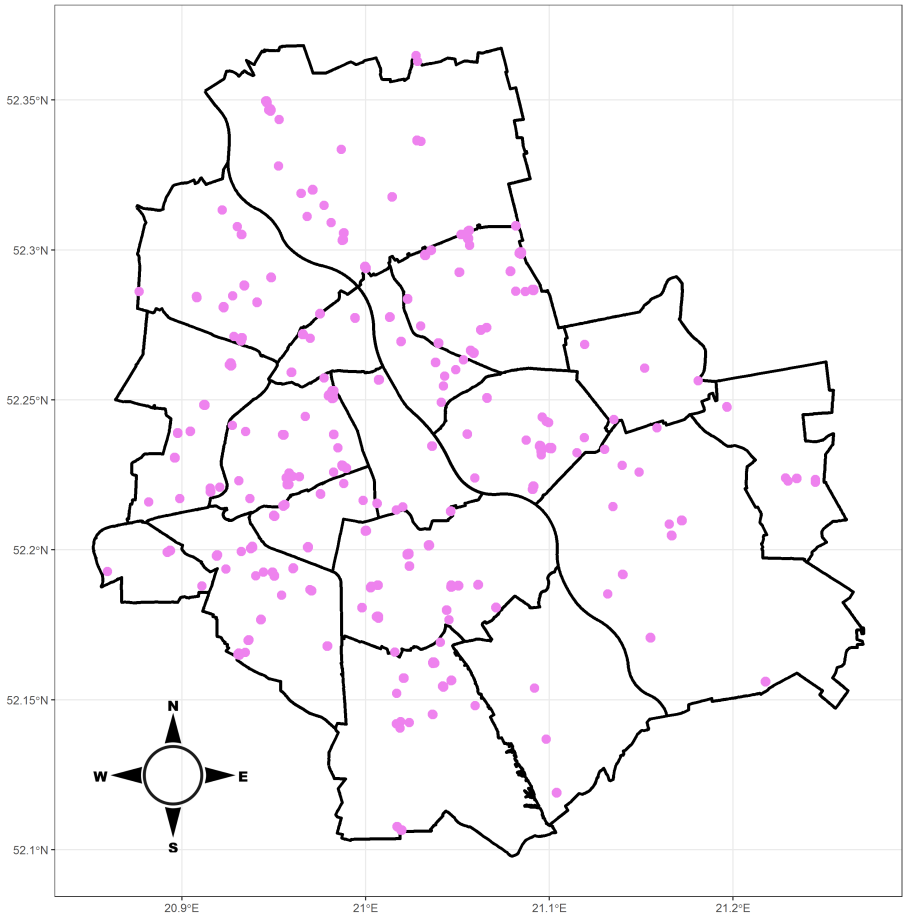
OSM public transport (stops by type) and bike rental (Veturilo stations)



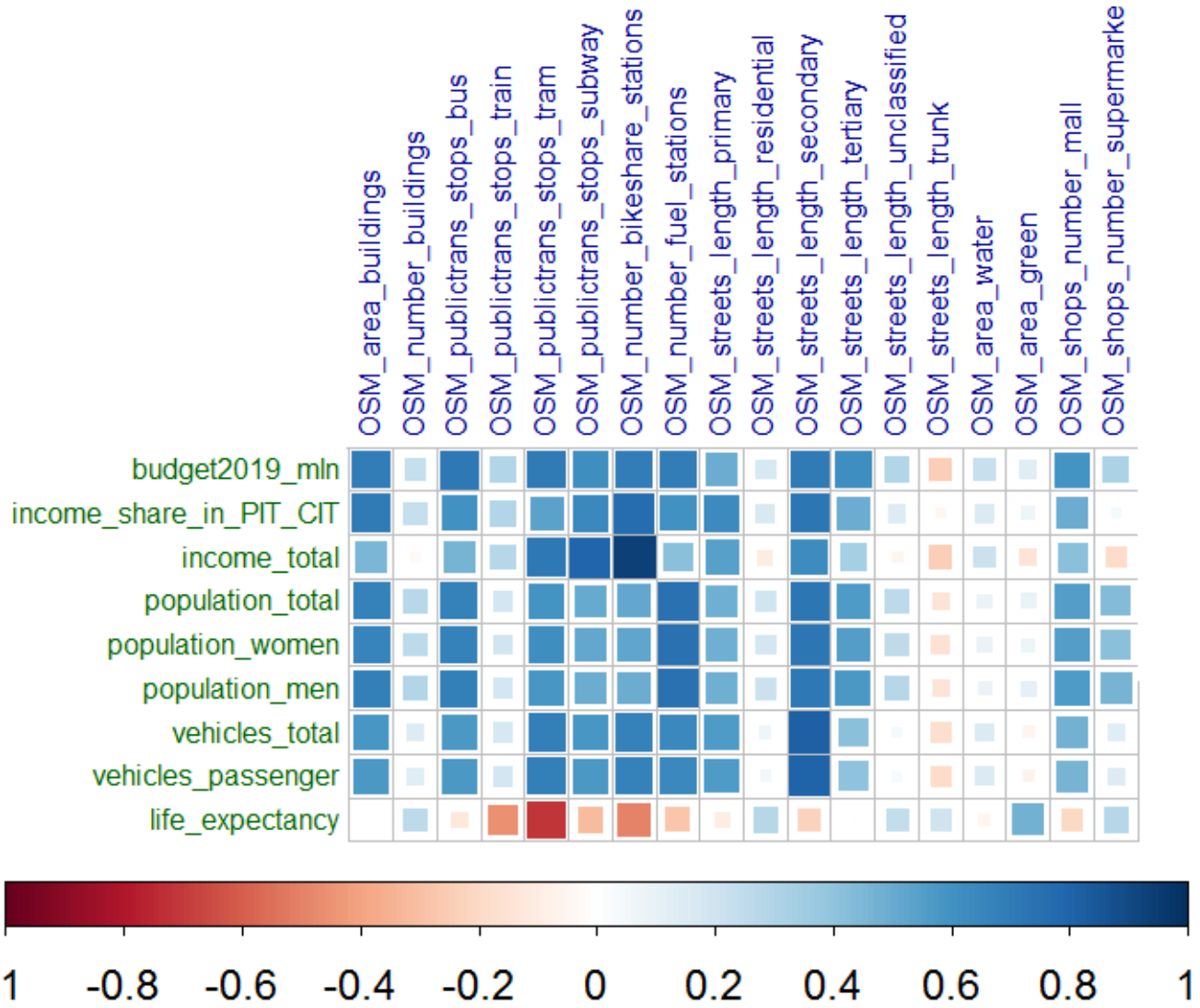


OSM supermarkets and fuel stations





Correlation between OSM extracted features and indicators



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

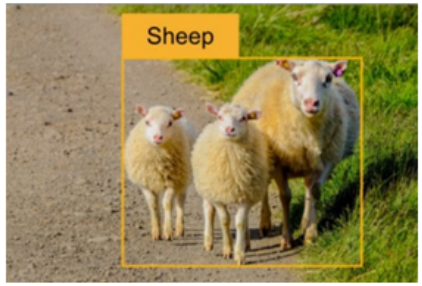
Feature extraction from satellite images

- 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.

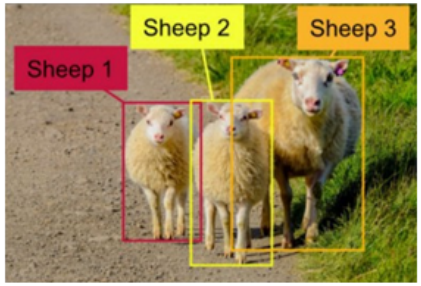
Types of problems in computer vision

Image classification

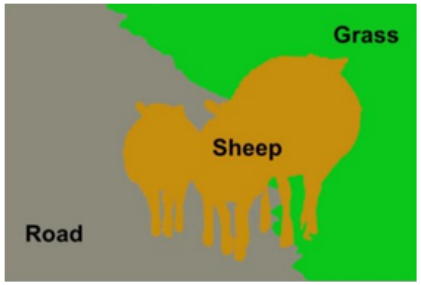
Classification with Localization



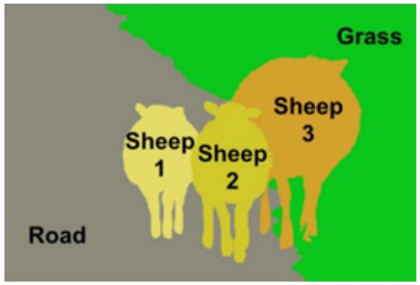
Object detection



Semantic segmentation

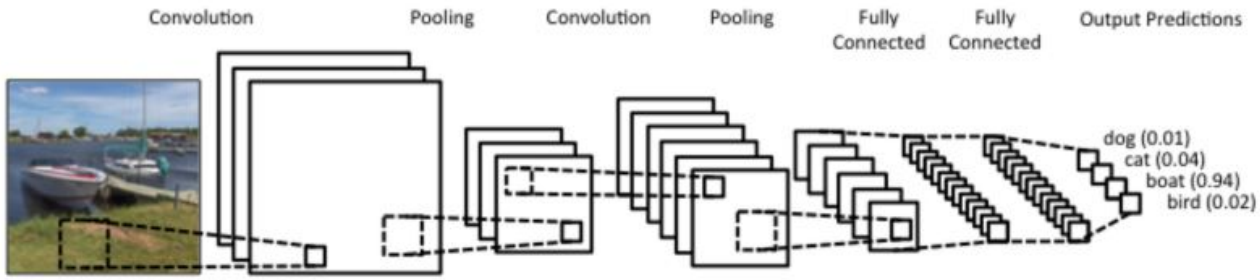


Instance segmentation



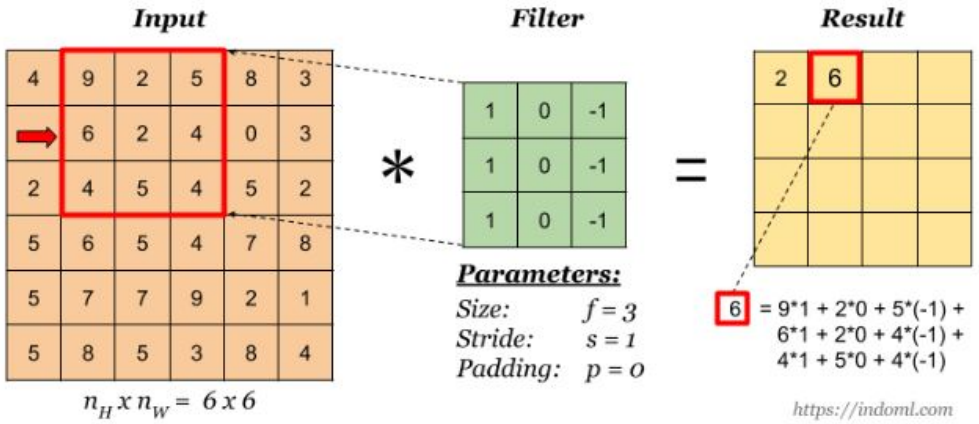
CNN

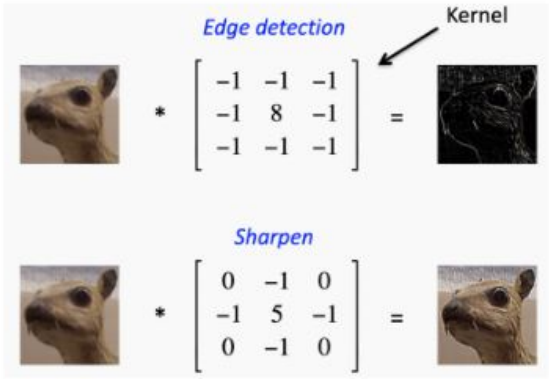
- Based on Feed-forward deep networks.
- Commonly used in image and video recognition.



Convolutional layer

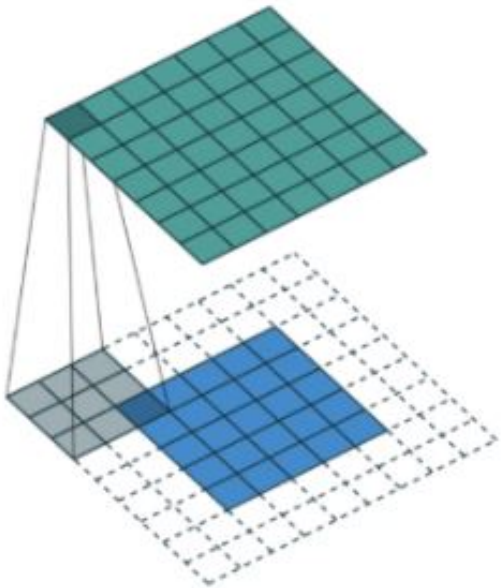
- Extracts features.



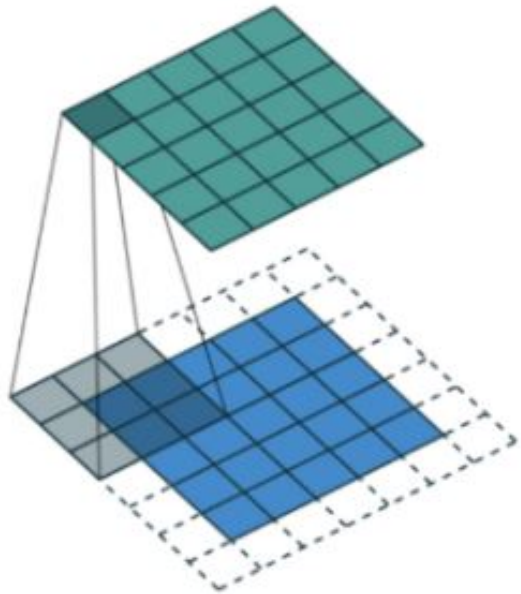


Padding

- Prevents image shrinking.
- Increase contribution of the pixels at the border of image.



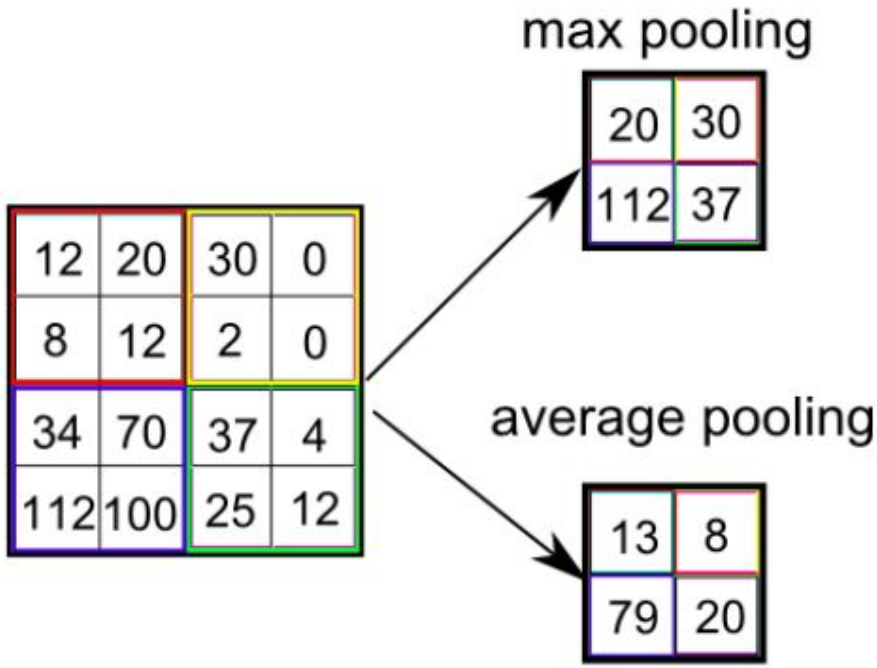
Full padding



Same padding

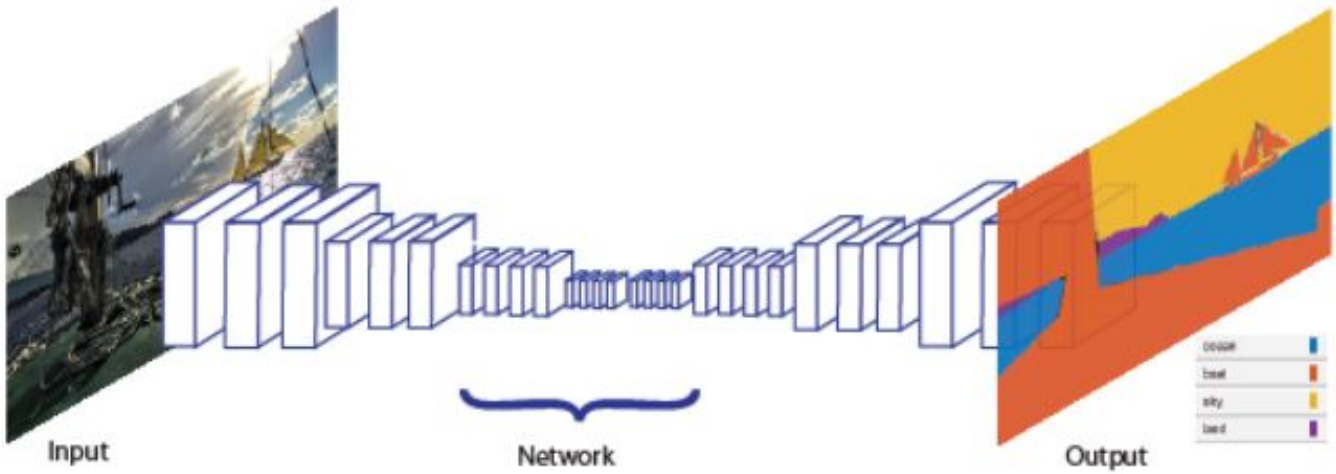
Pooling layer

- Reduce dimensionality.
- Extract maximum or average of region.



Semantic Segmentation

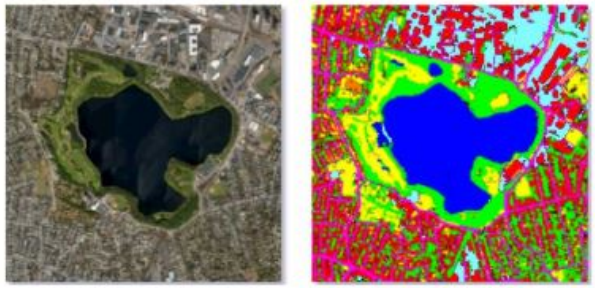
- Semantic segmentation is the process of classifying each pixel of an image to a class label.



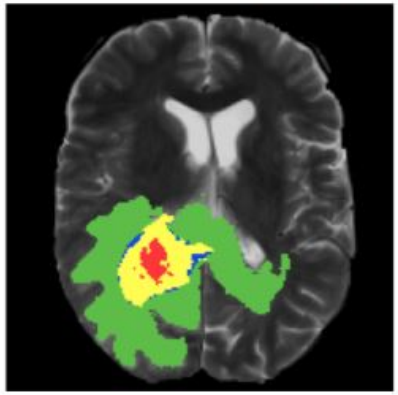
Applications for semantic segmentation

Autonomous driving

Satellite image analysis

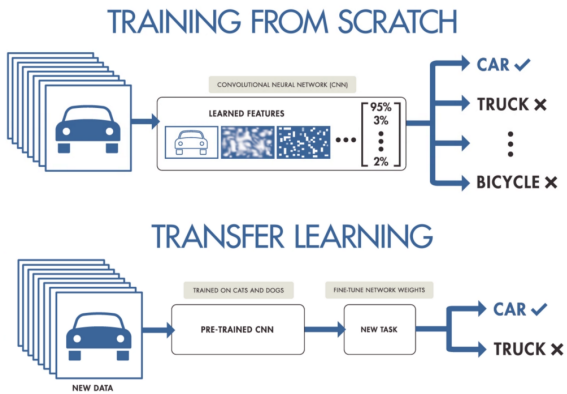


Medical imaging analysis



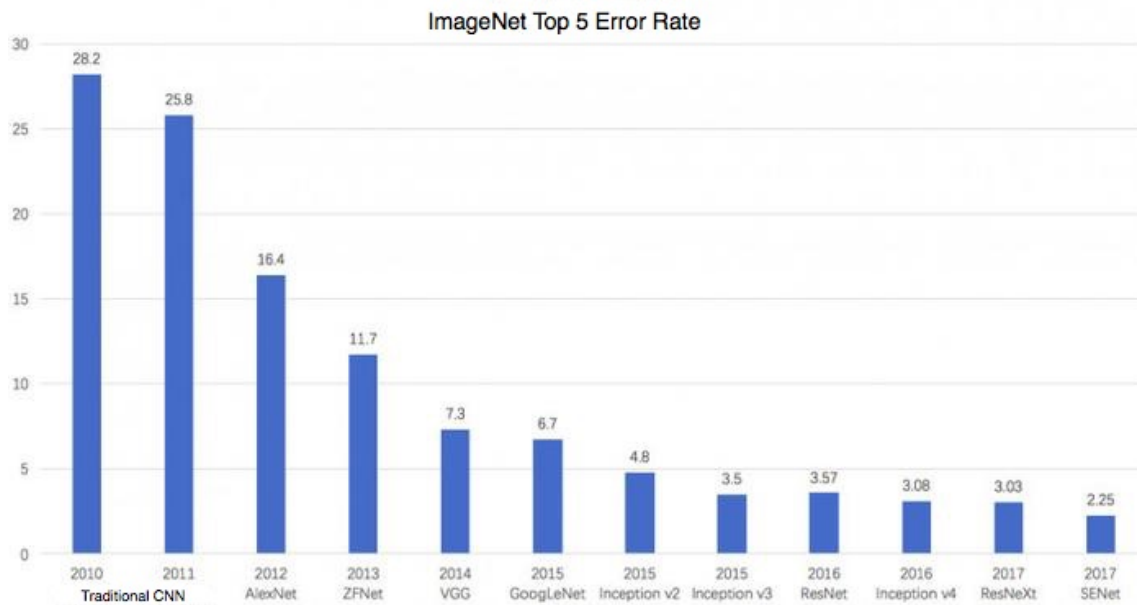
Transfer learning

- 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.



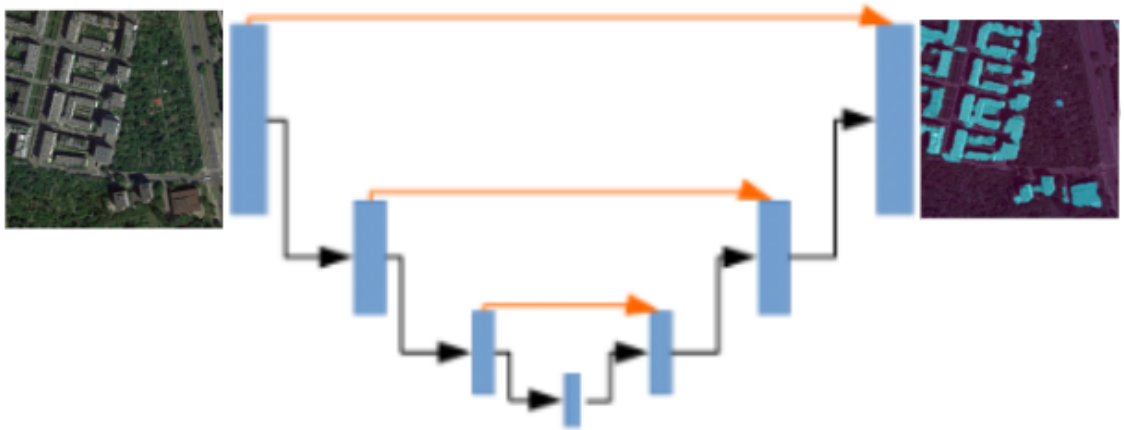
ImageNet

- The ImageNet project is a large visual database designed for use in visual object recognition software research.
- More than 14 million images have been hand-annotated by the project to indicate what objects are pictured.
- ImageNet contains more than 20,000 categories

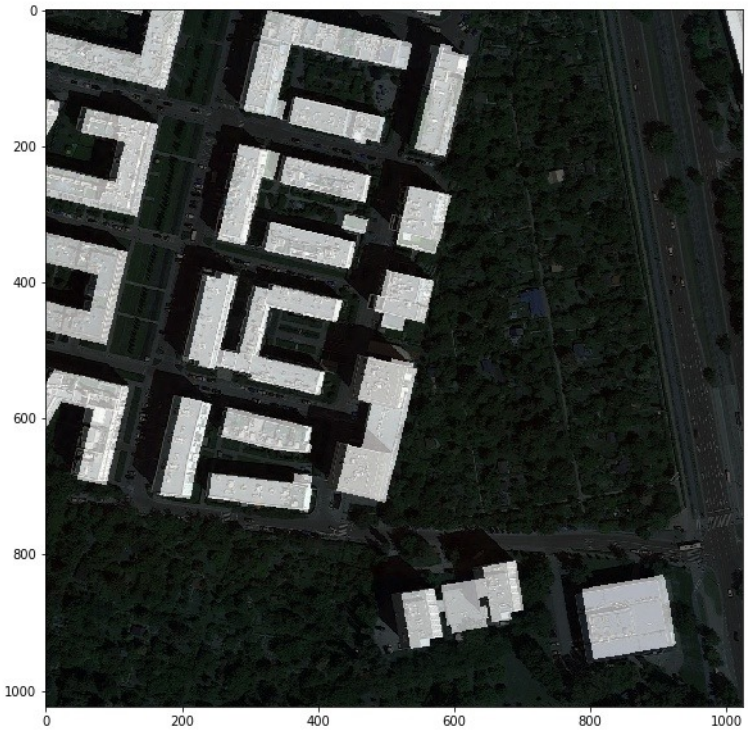


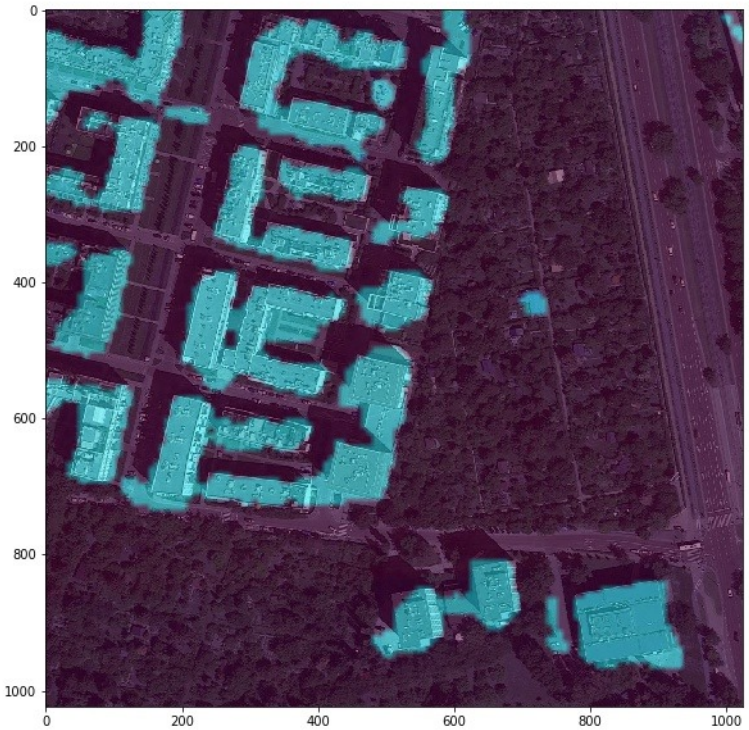
Unet and ResNet101

- 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



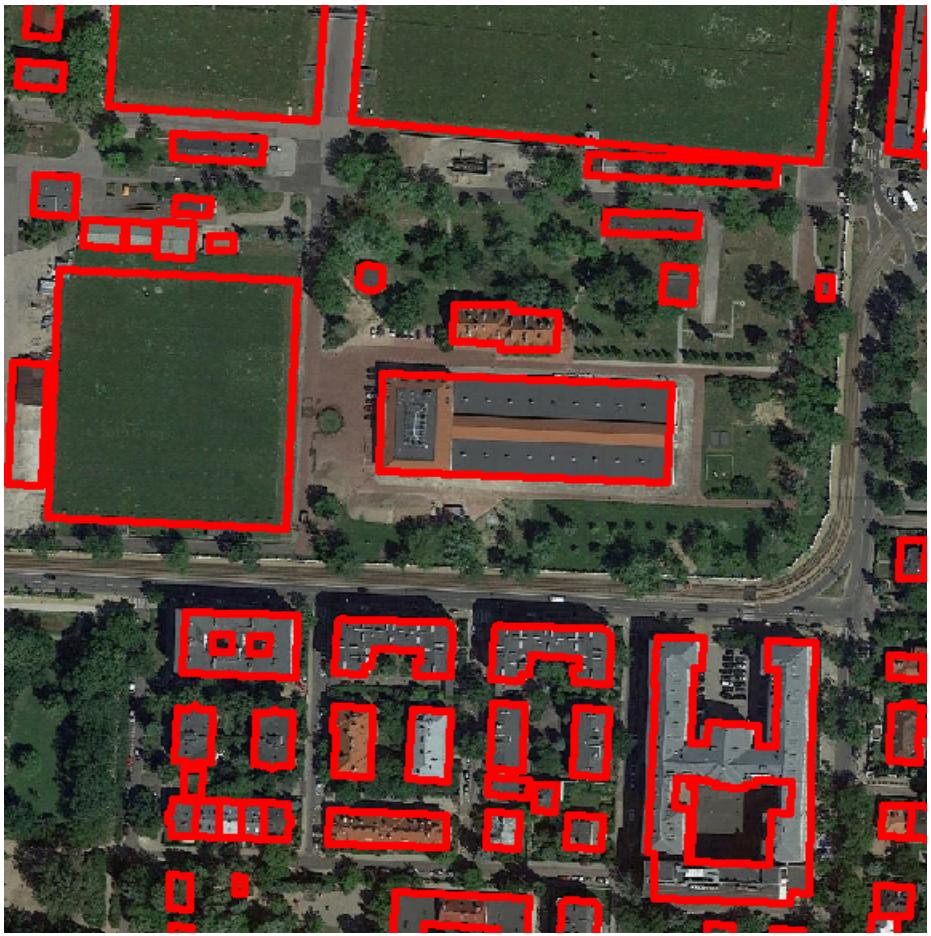


Preliminary results

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

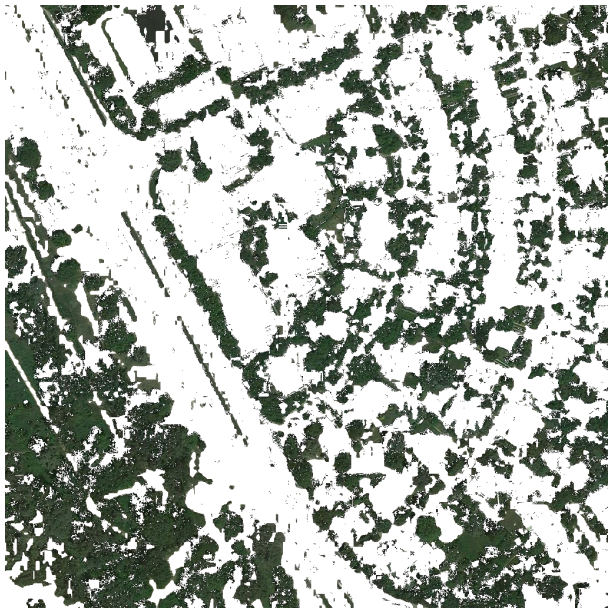
- Predicted area of buildings was calculated.
- Correlations between true and predicted area percentage values look promising.
- It could look better but we encountered some obstacles...

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_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

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
LASSO	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
LASSO (w/ pred variables)	OSM_number_bikeshare_stations	1	2	2	1	1	1	2	2	1
	OSM_number_fuel_stations	5	7	10	3	3	4	4	3	12
	OSM_publictrans_stops_bus	4	9	7	5	5	5	7	7	13
	OSM_publictrans_stops_tram	7	3	3	11	9	11	5	4	5
	OSM_shops_number_mall	3	5	9	4	4	3	8	8	17
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	OSM_area_buildings_pred	13	16	16	14	14	14	12	14	16
	OSM_area_green_pred	15	14	6	16	16	16	16	16	2
	MAE	129.64	18594.46	77569.34	35042.93	19080.28	15920.14	34132.85	28925.08	1.68
	MAPE	0.400	0.651	1.267	0.524	0.537	0.510	0.663	0.694	0.022

Conclusions

- object detection on satellite images requires large computing power and good training data
- well-being prediction requires many variables, not only buildings and green areas
- 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 vs houses)
- obtain data over time to observe changes
- broaden analysis to other areas since metropolies are heterogenous
- use model on areas that have even worse labelled data

Thank you

THANK YOU FOR YOUR ATTENTION!