Prognozowanie poziomu dobrostanu lokalnego na podstawie cech widocznych z kosmosu na przykładzie Warszawy

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

seminarium EUROREG
6.05.2021 r.

UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Cel badania

- **celem badania** jest odpowiedź na pytanie, czy cechy obszarów geograficznych widoczne z kosmosu mogą posłużyć do przewidywania dobrostanu społeczno-ekonomicznego na poziomie lokalnym
- badanie wykonane jest na przykładzie **dzielnic Warszawy**
- konstruujemy **indeks dobrostanu lokalnego** oparty na trzech wymiarach: zdrowie, edukacja i dochody
- wśród potencjalnych jego **czynników** rozpatrujemy wskaźniki związane ze stopniem urbanizacji, dostępem do walorów przyrodniczych, systemem transportowym i dostępem do transportu publicznego
- cechy poszczególnych obszarów identyfikowane są na podstawie **dziennych zdjęć satelitarnych** oraz **danych OSM** (POI)

Motywacja badania

- w badaniach empirycznych **wskaźnikiem dobrostanu** jest zwykle PKB *per capita*
- dobrostan jest jednak **zjawiskiem złożonym**, a jego pomiar wymaga uwzględnienia **większej liczby elementów**
- jednym z przykładów jest Human Development Index (HDI), koncentrujący się na trzech wymiarach: **zdrowiu** i długowieczności, wiedzy i **edukacji** oraz **dobrobycie materialnym**
- konstruowanie lokalnych wskaźników dobrostanu jest trudniejsze (brak porównywalnych danych na poziomie niższym niż krajowy)
- próby stworzenia syntetycznych wskaźników dobrostanu na poziomie lokalnym są zwykle **działaniami jednorazowymi** – UNDP (2012), Sompolska-Rzechuła (2016), Pomianek (2016), Ciołek (2017)
- ponadto podejścia wykorzystujące dane z oficjalnych rejestrów statystycznych można stosować **tylko na poziomie jednostek administracyjnych**, o ile dostępne są wiarygodne dane statystyczne



Motywacja badania – c.d.

- w ostatnich latach często wykorzystuje się **dane satelitarne** dla przybliżenia natężenia aktywności gospodarczej na poziomie regionalnym i lokalnym
- znalezienie powiązania między wskaźnikami dobrostanu a danymi satelitarnymi pozwoliłoby przewidzieć wartości tych wskaźników dla **dowolnego obszaru**, nawet nie administracyjnego
- związki między **intensywnością świateł nocnych** a wskaźnikami dobrostanu na poziomie niższym niż krajowy są często **słabe**
- w większości badań stosuje się **jedynie liniowe** lub log-liniowe modele statystyczne
- bardziej obiecująca jest analiza **dziennych zdjęć satelitarnych** o wysokiej rozdzielczości

Motywacja badania – c.d.

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

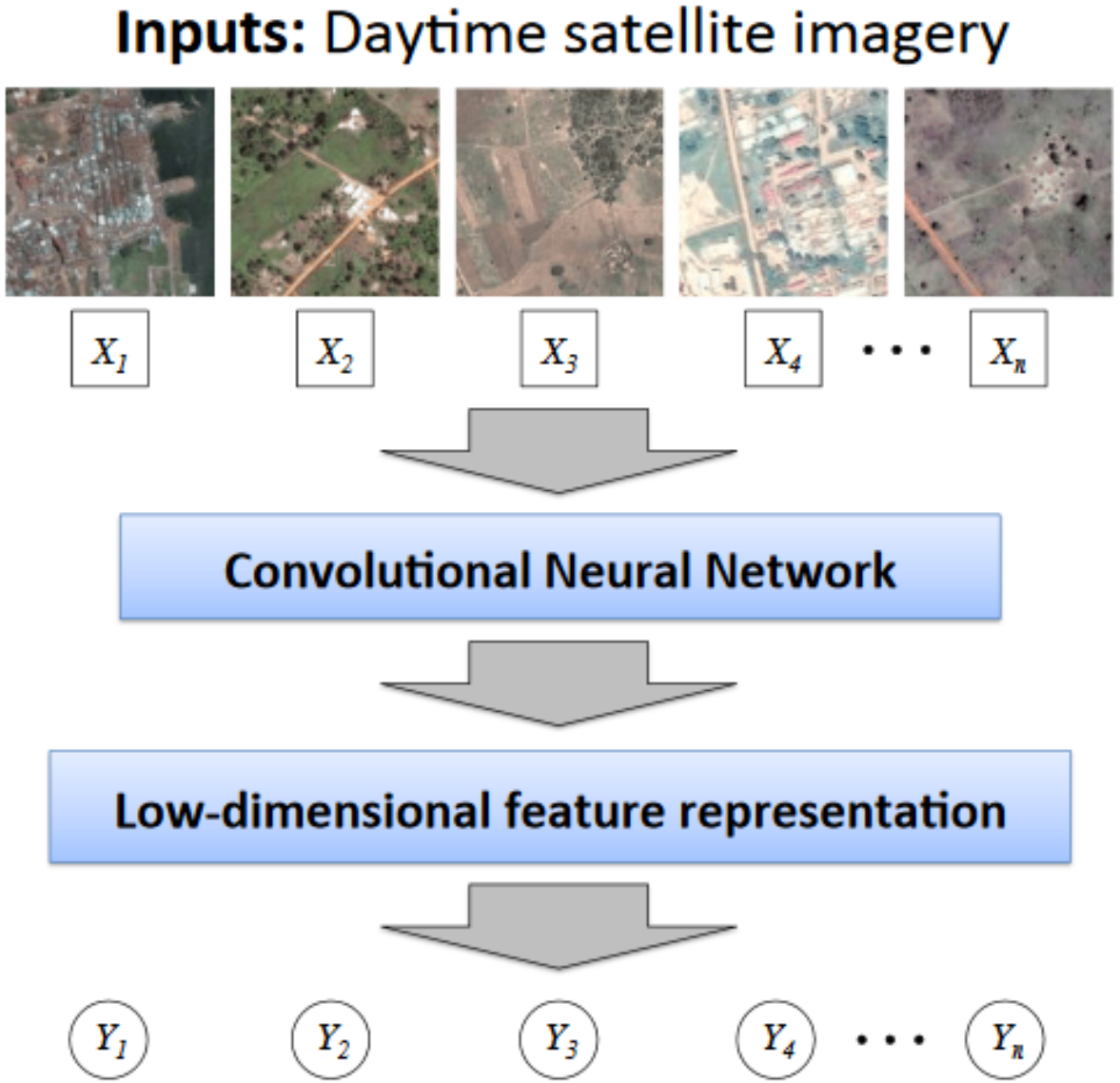
Science 19 Aug 2016:
Vol. 353, Issue 6301, pp. 790-794
DOI: 10.1126/science.aaf7894

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.

Motywacja badania – c.d.



Motywacja badania – c.d.

- jednak pozyskanie danych (cech obszarów) z wielu zdjęć wymaga **wyspecjalizowanych narzędzi** do rozpoznawania obrazów (konwolucyjnych sieci neuronowych) i **ogromnej mocy obliczeniowej**
- najnowsze badania (Chen i in. 2016, Tingzon i in. 2019) sugerują, że wykorzystanie **bezpłatnych i łatwo dostępnych** danych z Open Street Map pozwala na osiągnięcie porównywalnie dokładnej predykcji wskaźników społeczno-ekonomicznych
- podobnie jak zdjęcia satelitarne, dane OSM są **aktualizowane znacznie częściej** niż konwencjonalne dane statystyczne
- dzięki modelowi precyzyjnie przewidującemu dobrostan na podstawie zdjęć satelitarnych lub OSM, jego pomiar na poziomie regionalnym i lokalnym może być wykonany na długo **przed zebraniem oficjalnych danych** statystycznych
- ponadto takie wskaźniki można łatwo przewidzieć dla **obszarów nieadministracyjnych**


DSLab

Zastosowane podejście empiryczne

1. **stworzenie miary lokalnego dobrostanu** na poziomie dzielnic Warszawy inspirowanego LHDI (UNDP, 2012)
2. **wyodrębnienie cech** widocznych z kosmosu ze zdjęć satelitarnych dla poszczególnych dzielnic Warszawy
   - wykorzystano narzędzia uczenia głębokiego (ang. *deep learning*) – **konwolucyjne sieci neuronowe** (CNN) i tzw. *transfer learning*
3. **zastosowanie modeli predykcyjnych** wykorzystujących te cechy oraz dane pobrane z OSM jako predyktory dobrostanu na poziomie dzielnic Warszawy
   - wykorzystano algorytmy **uczenia maszynowego** pozwalające na **automatyczną selekcję** ważnych zmiennych objaśniających oraz umożliwiające identyfikację **potencjalnie nieliniowych zależności** (LASSO, regresja wektorów nośnych, las losowy i xgboost).
   - ponadto stosujemy innowacyjne narzędzia **wytłumaczalnej sztucznej inteligencji** (XAI) do identyfikacji najważniejszych czynników dobrostanu oraz do odkrycia kształtu zależności.


Hipotezy badawcze

W pracy weryfikacji poddano trzy hipotezy badawcze:

1. Cechy widoczne z kosmosu pozwalają z **dużą precyzją** przewidywać poziom dobrostanu wewnątrz miasta.
2. Dane pozyskane z Open Street Map są **równie dobrymi** predyktorami dobrostanu jak cechy wyodrębnione ze zdjęć satelitarnych.
3. Modele uczenia maszynowego oferują **większą dokładność** w przewidywaniu poziomu dobrostanu w porównaniu z modelami liniowymi.

Źródła danych

- Open Street Map – lokalizacja budynków, ulic, rzek i jezior, terenów zielonych, transportu publicznego, wypożyczalni rowerów, stacji benzynowych, supermarketów i centrów handlowych
- Google Maps: 6000+ dziennych zdjęć satelitarnych Warszawy (wysokiej rozdzielczości)
- wszystkie dane pozyskane z OSM i GM zostały **zagregowane do poziomu dzielnic** i **przeliczone na 1 km2** (gęstość)
- wskaźniki społeczno-ekonomiczne dla warszawskich dzielnic:
  - *Panorama dzielnic Warszawy w 2017 r.* – udział dochodów dzielnicy z PIT/CIT
  - raport *Stan zdrowia mieszkańców Warszawy* (2013)
  - Centralna Komisja Egzaminacyjna

Indeks dobrostanu

- Indeks dobrostanu (Well Being Index, **WBI**) dla warszawskich dzielnic zbudowany jest z trzech komponentów:
  - indeks **zdrowia** (Health Index, **HI**),
  - indeks **edukacji** (Education Index, **EI**),
  - indeks **zamożności** (Welfare Index, **WI**).
- **WBI** jest **średnią harmoniczną** poszczególnych komponentów
- ponieważ WI jest niedoskonałym przybliżeniem dochodów rozporządzalnych, dajemy mu mniejszą wagę:

$$WBI_i = \left(\frac{2HI_i^{-1} + 2EI_i^{-1} + WI_i^{-1}}{5}\right)^{-1}$$

- wszystkie **zmienne bazowe** zostały wyrażone w **wartościach relatywnych**, w odniesieniu do średniej dla całej Warszawy i pomnożone przez 100

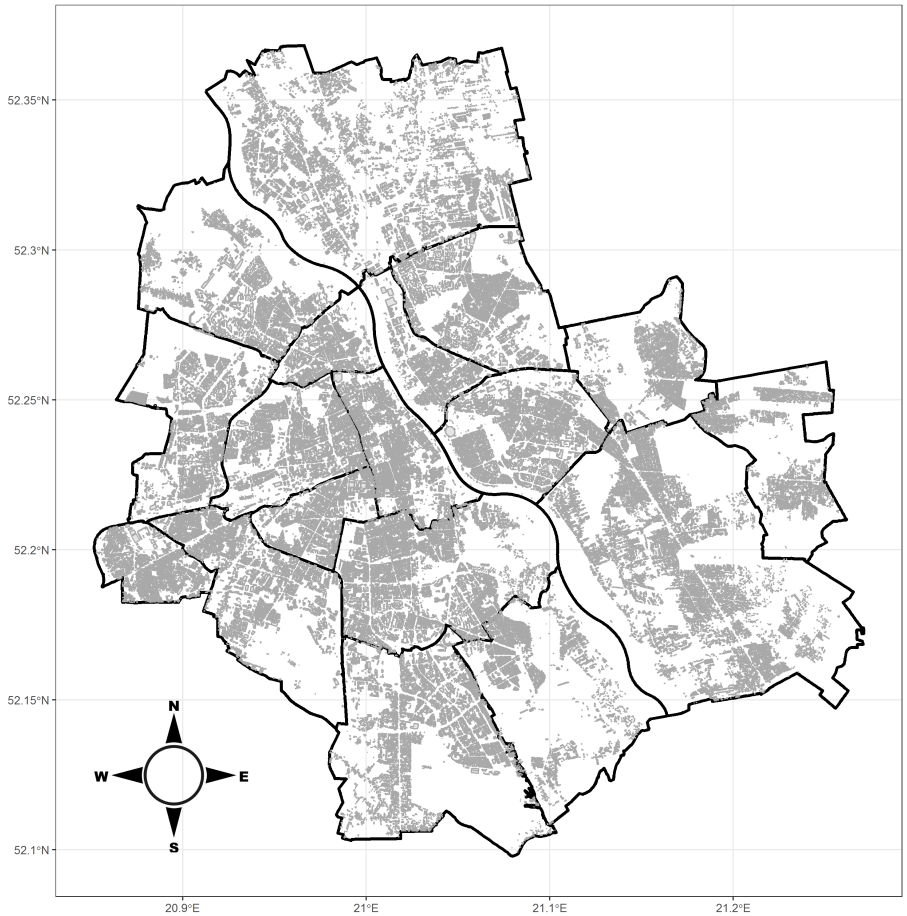
Indeks dobrostanu – c.d.

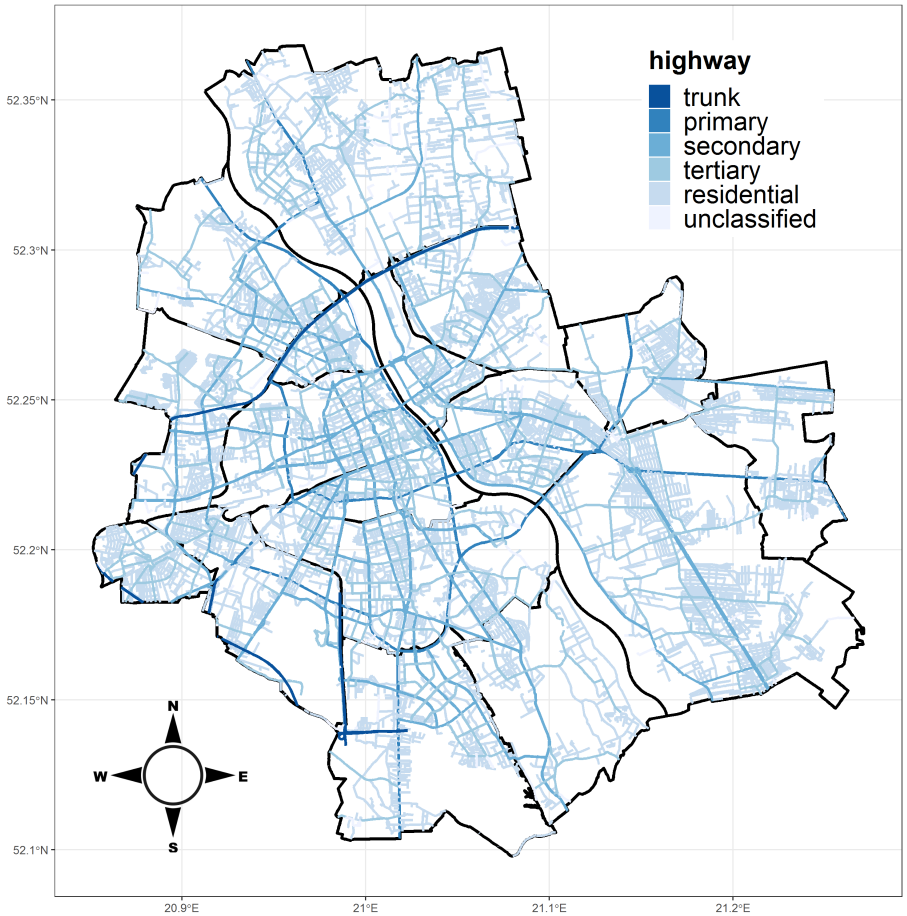
- indeks zdrowia (**HI**) jest **średnią harmoniczną** z **trzech** zmiennych bazowych:
  - oczekiwane dalsze trwanie życia w momencie narodzin (**H_LE**)
  - (zmormalizowana) stopa zgonów z powodu chorób układu krążenia (**H_CD**)
  - (zmormalizowana) stopa zgonów z powodu nowotworów złośliwych (**H_MN**)
- indeks edukacji (**EI**) jest **średnią harmoniczną** z **dwóch** zmiennych bazowych:
  - udział dzieci w wieku przedszkolnym (3–5) uczęszczających do przedszkola (**E_PS**)
  - średni wynik egzaminu gimnazjalnego (części matematyczno-przyrodniczej) (**E_EM**)
- indeks zamożności (**WI**) opiera się na jednej zmiennej bazowej – przychody dzielnicy z tytułu udziału w podatkach CIT/PIT *per capita*


Indeks dobrostanu – c.d.

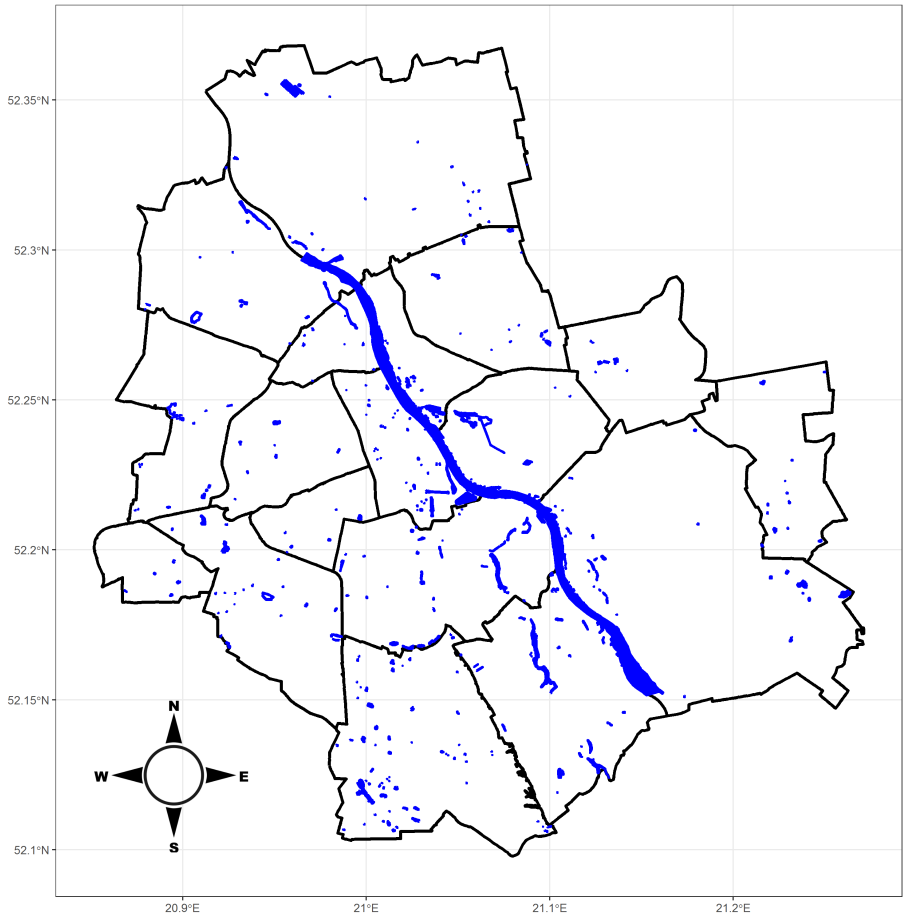
district	rank according to WBI	HI	EI	WI	WBI
Włochy	1	103.56	112.35	347.85	125.04
Śródmieście	2	94.13	106.95	222.98	112.53
Wilanów	3	123.90	100.63	118.61	112.49
Mokotów	4	99.78	100.60	125.25	104.37
Ochota	5	102.20	102.57	106.71	103.22
Wawer	6	112.57	96.59	99.17	102.97
Białołęka	7	99.81	96.65	109.32	100.24
Rembertów	8	107.43	97.38	94.71	100.58
Wola	9	87.03	95.74	114.99	95.12
Ursynów	10	119.40	105.96	75.36	102.26
Wesoła	11	107.73	106.55	73.31	98.08
Ursus	12	106.12	100.70	75.34	96.19
Targówek	13	98.71	93.81	76.88	91.60
Praga-Południe	14	98.14	97.73	60.13	86.99
Bemowo	15	108.28	92.09	54.62	85.47
Żoliborz	16	94.43	98.97	55.83	84.32
Praga-Północ	17	82.54	83.58	62.40	77.90
Bielany	18	102.37	109.28	41.37	80.63
min		82.54	83.58	41.37	69.62
max		123.90	112.35	347.85	139.98
mean		102.67	99.90	106.38	96.67
CV		9.64	6.66	66.51	17.79

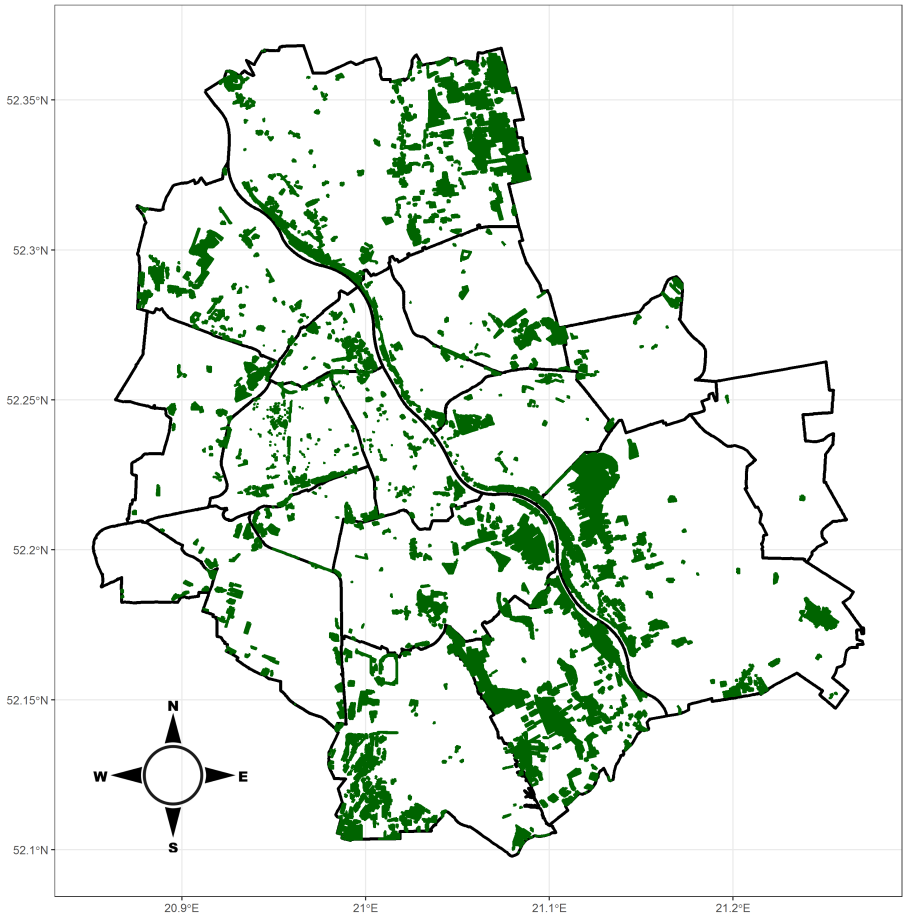
Dane OSM – budynki i ulice wg typu



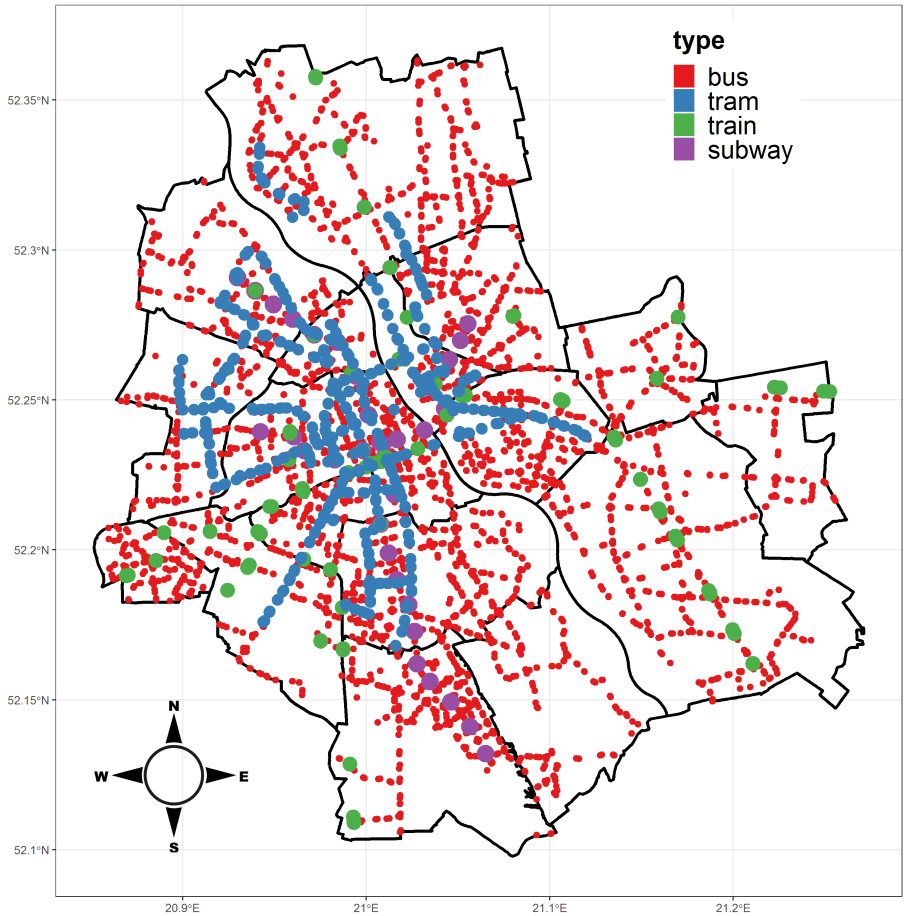


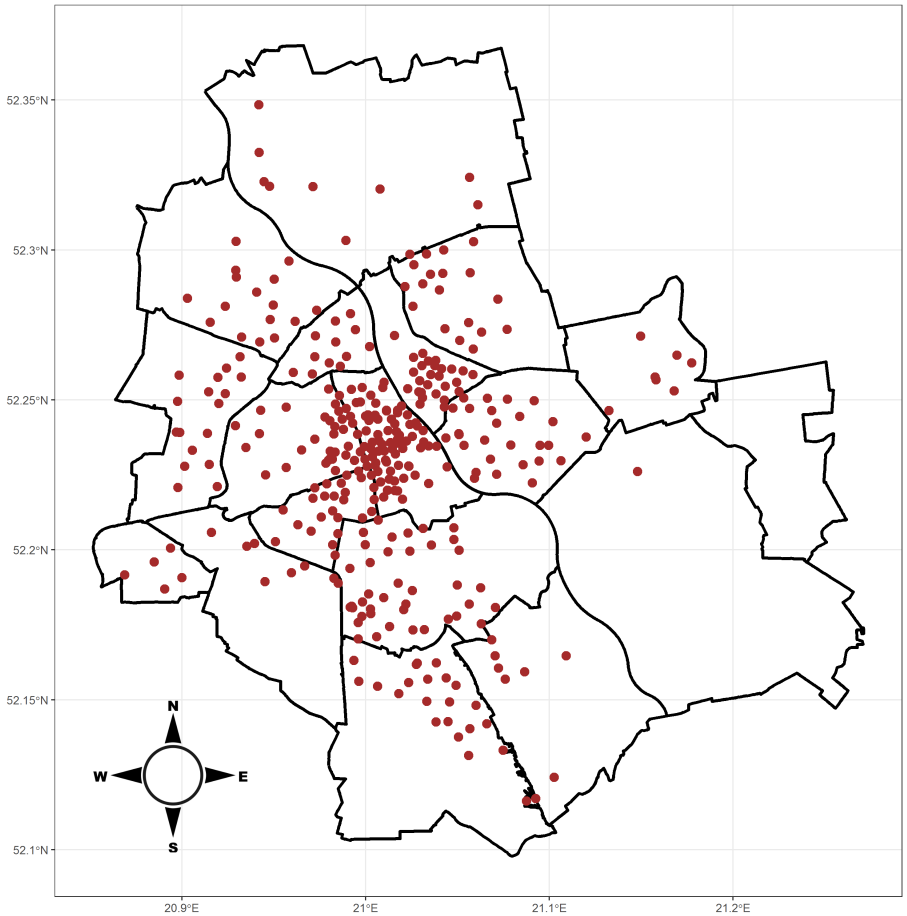
Dane OSM – woda i obszary zielone



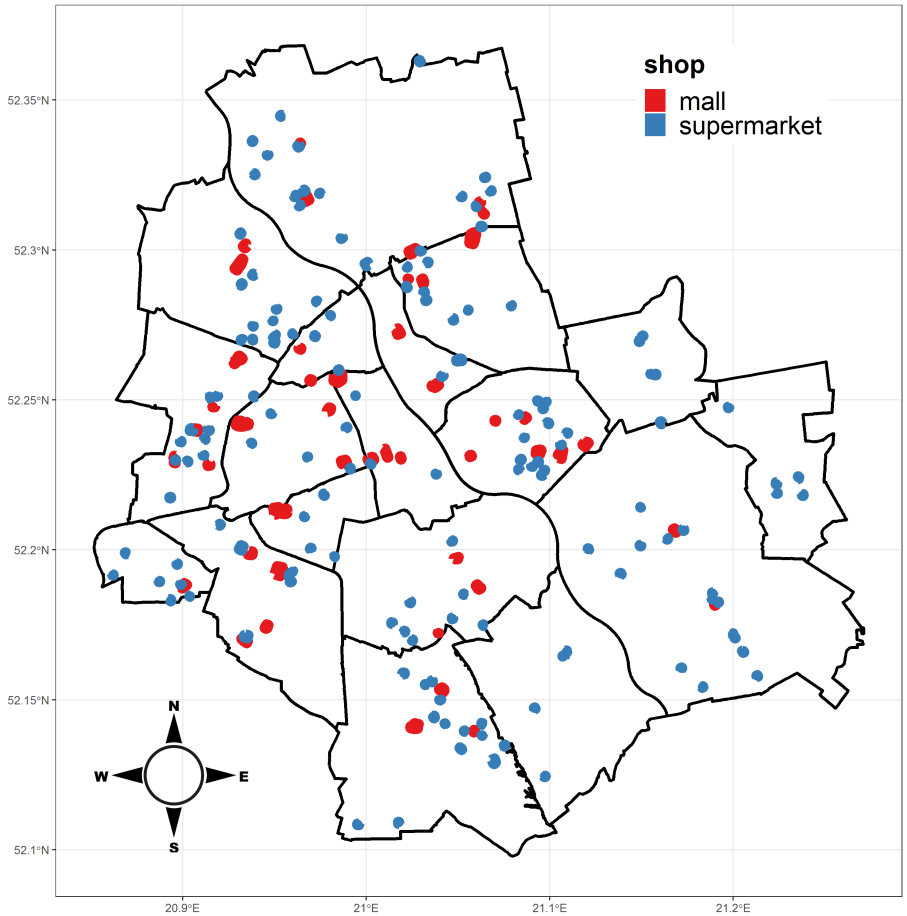


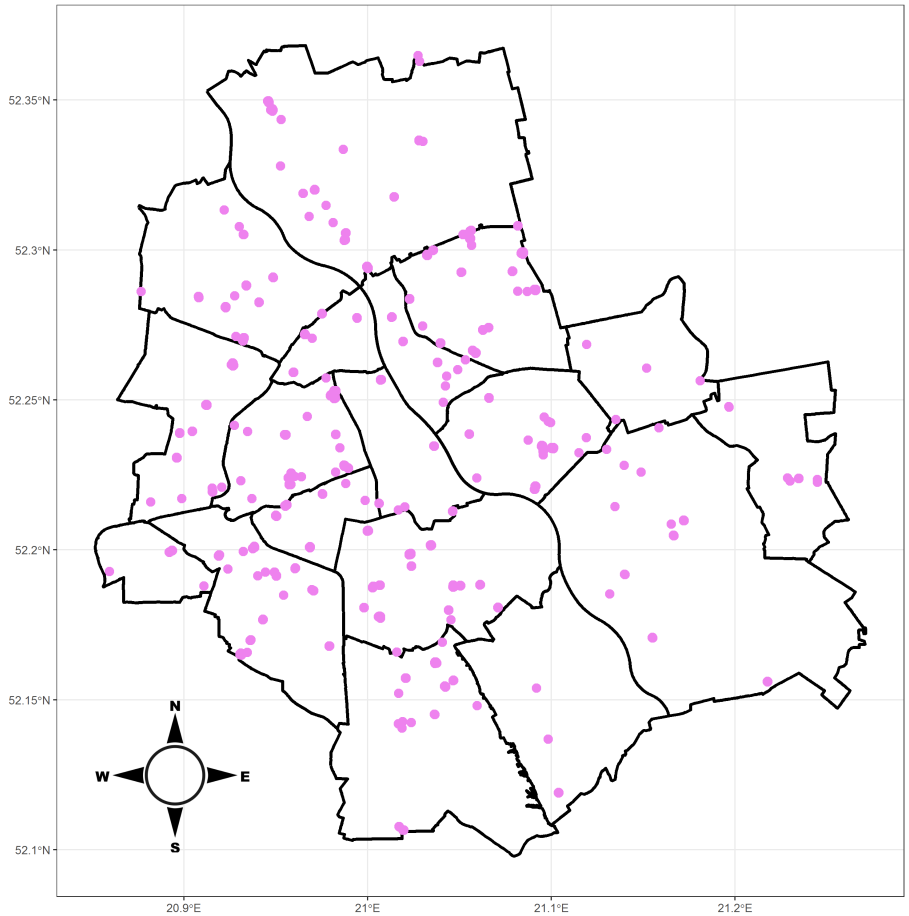
Dane OSM – transport publiczny i stacje Veturilo





Dane OSM – supermarkety i stacje benzynowe





Zdjęcia z Google Maps

Wyodrębnianie cech ze zdjęć satelitarnych

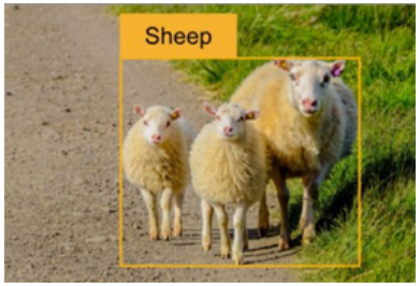
- dane z OSM zostały użyte do **oznaczenia obiektów** na zdjęciach satelitarnych (w próbie uczącej)
- zastosowano **transfer learning** przeprowadzony z wykorzystaniem konwolucyjnych sieci neuronowych
- model **nie jest trenowany od zera**, ale używa modelu zastosowanego wcześniej do podobnego problemu, wstępnie wytrenowanego na dużym zbiorze danych (tu: baza obrazów Imagenet)
- metodę wykorzystano do rozpoznania na zdjęciach Warszawy **budynków**
- ze względu na **nieprecyzyjne oznaczenia terenów zielonych** w OSM wykorzystano **alternatywny sposób** rozpoznawania ich na zdjęciach

Identyfikacja obiektów na zdjęciach satelitarnych

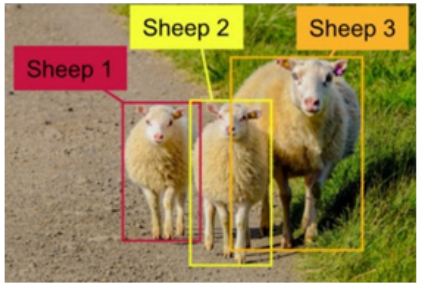
Rodzaje problemów w *computer vision*

Image classification

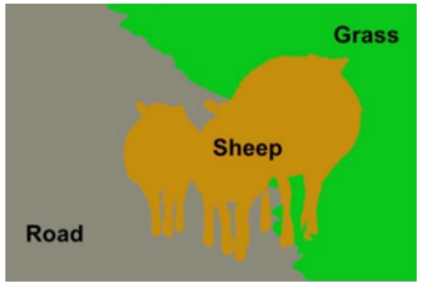
Classification with Localization



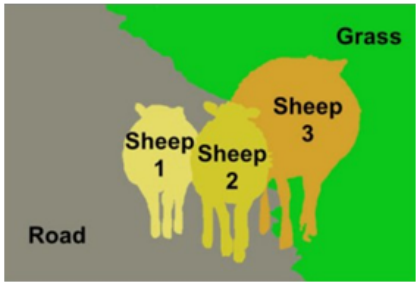
Object detection



Semantic segmentation

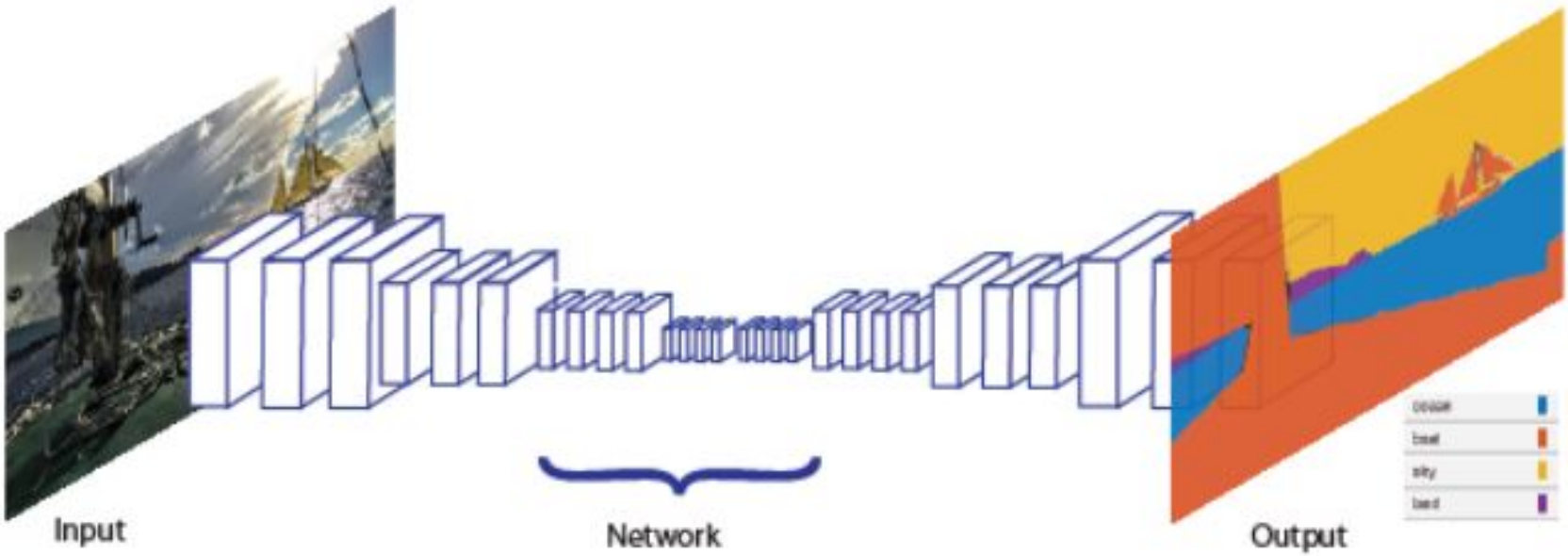


Instance segmentation



Segmentacja semantyczna

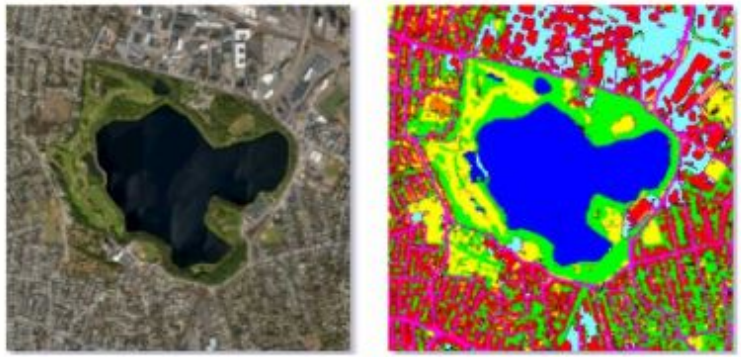
Segmentacja semantyczna to proces klasyfikowania każdego piksela obrazu do określonej grupy (klasy)



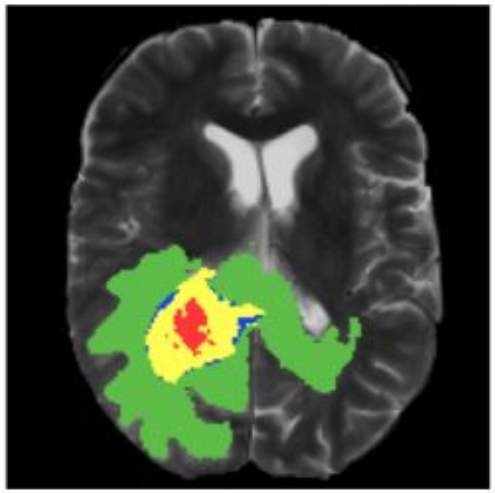
Zastosowania segmentacji semantycznej

Autonomous driving

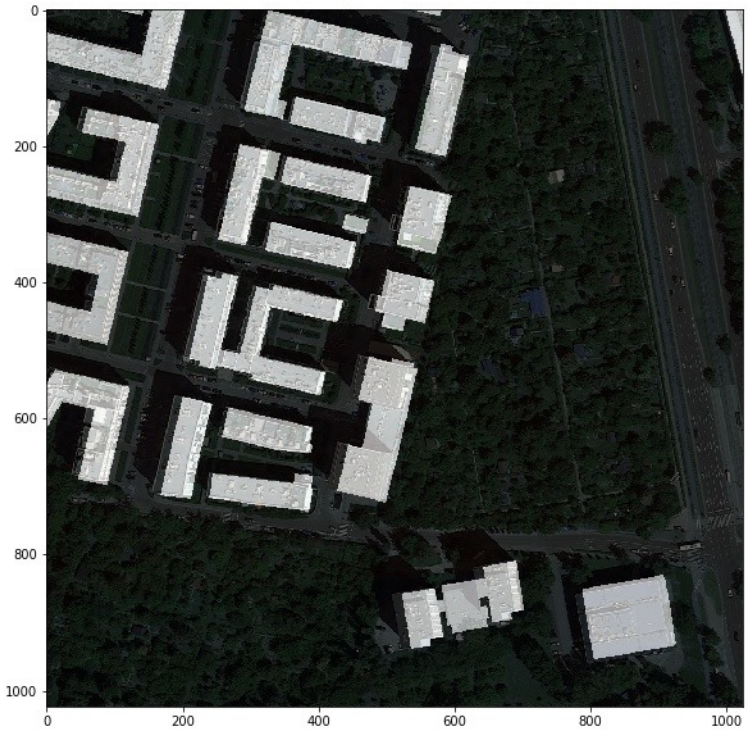
Satellite image analysis

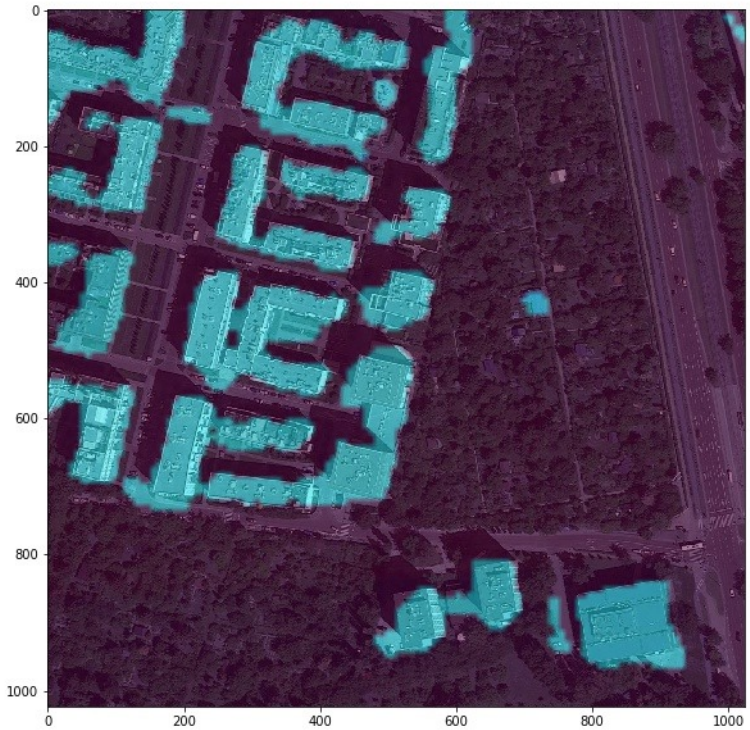


Medical imaging analysis

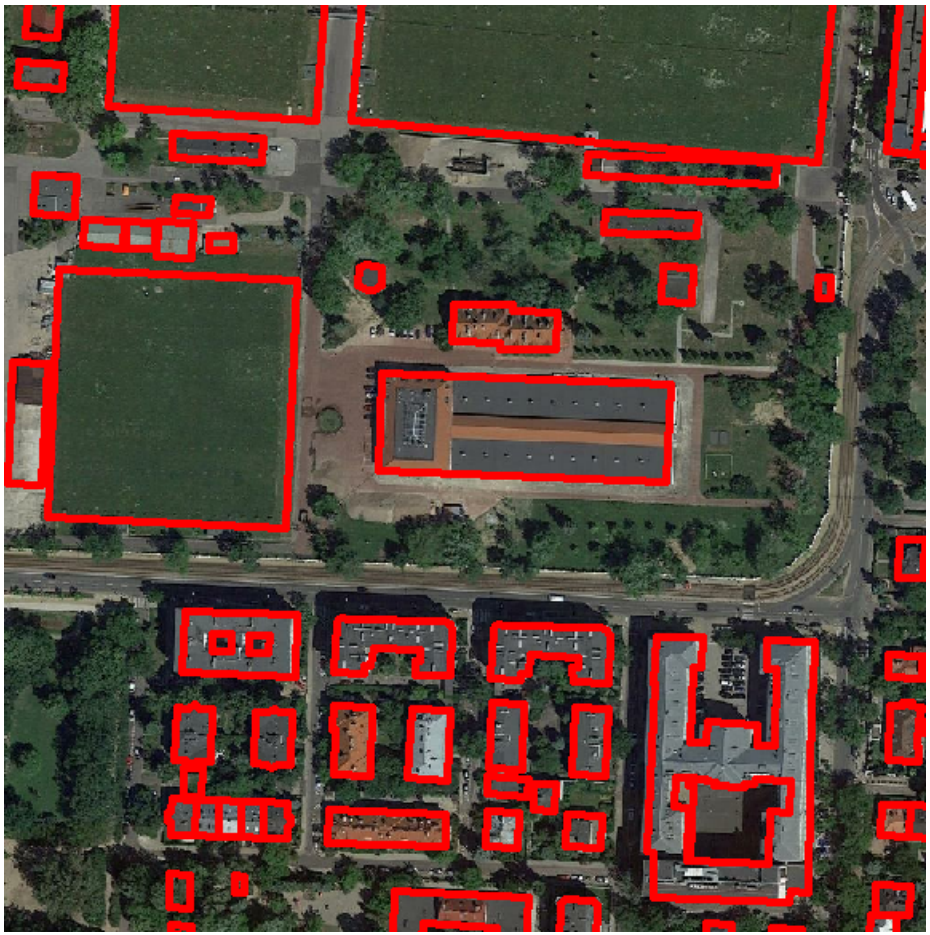


Przykładowe prawdziwe i prognozowane budynki



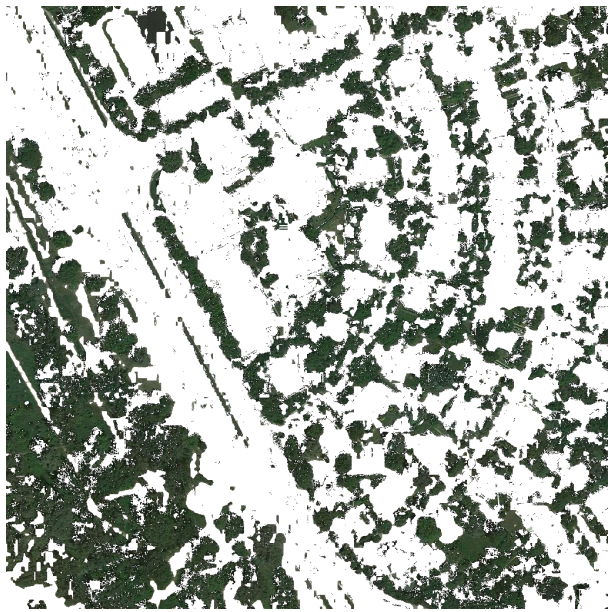


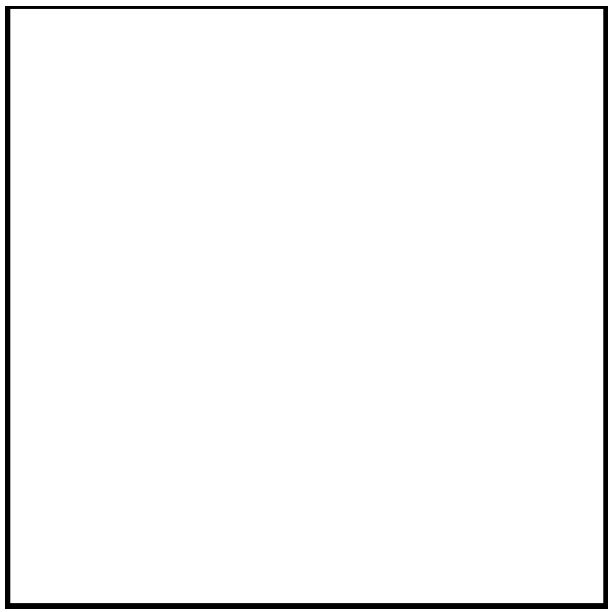
Przykładowe problematyczne budynki



Rozpoznawanie terenów zielonych

- obszary zielone w OSM są słabo oznakowane
- można to zrobić lepiej wyodrębniając tylko piksele w wybranych zakresach kolorów





Predykcja poziomu dobrostanu

- wyodrębnione cechy obszarów zostały wykorzystane w modelach predykcyjnych
- zastosowano **cztery nieliniowe modele uczenia maszynowego**: regresję wektorów nośnych (z wielomianową i radialną funkcją jądra, **SVR**), las losowy (**RF**) i *extreme gradient boosting* (**XGB**)
- model LASSO jest wykorzystany jako **liniowy benchmark**, ponieważ tradycyjna regresja liniowa nie może być stosowana, gdy liczba obserwacji jest równa liczbie predyktorów
- modele szacowano na **dwóch grupach predyktorów**:
  - wyłącznie zmienne z Open Street Map (**OSM**)
  - alternatywnie wartości dwóch predyktorów (powierzchnia budynków i terenów zielonych) zastąpiona danymi pozyskanymi ze zdjęć satelitarnych Warszawy; pozostałe predyktory niezmienione (**OSM + predicted**)
- porównanie modeli przeprowadzono z wykorzystaniem **walidacji *leave-one-out***

Jakość prognoz (MAPE) na próbie walidacyjnej

data	model	dependent variable								
		H_CD	H_MN	H_LE	HI	E_PS	E_EM	EI	WI	WBI
OSM	LASSO	13.58%	5.33%	1.80%	5.99%	8.41%	5.28%	5.55%	48.99%	10.09%
	RF	11.86%	5.30%	1.45%	5.35%	8.55%	6.21%	6.06%	49.69%	10.20%
	SVRpoly	12.04%	5.57%	1.63%	5.92%	8.11%	5.67%	5.68%	41.18%	8.58%
	SVRradial	14.21%	6.40%	2.13%	7.10%	7.83%	5.53%	5.46%	37.47%	10.00%
	XGB	9.53%	4.82%	1.24%	4.81%	7.64%	6.38%	5.89%	35.99%	10.35%
OSM + predicted	LASSO	13.21%	6.63%	1.77%	6.03%	6.36%	5.38%	5.55%	48.99%	10.31%
	RF	12.12%	5.11%	1.47%	5.32%	8.46%	5.54%	6.02%	50.77%	10.44%
	SVRpoly	11.73%	6.07%	1.53%	5.81%	5.25%	5.67%	5.68%	37.59%	9.95%
	SVRradial	13.25%	6.01%	1.80%	6.42%	7.50%	5.66%	5.48%	38.98%	9.91%
	XGB	10.22%	4.28%	1.27%	4.80%	6.35%	5.47%	5.00%	33.07%	10.65%

Jakość prognoz (MAPE) na próbie walidacyjnej

- **WBI** przewidywany ze średnim błędem około 10%
- Indeks zdrowia (**HI**) i indeks edukacji (**EI**) mają błędy prognozy ok. 5–7% w zależności od modelu
- Najtrudniejszą zmienną do przewidzenia jest **WI** (błąd prognozy 33–50%)
- Na podstawie skrajnie zróżnicowanych błędów predykcji dla **WI** widać, jak ważny jest dobór odpowiedniego algorytmu
- XGB jest najlepszy w przewidywaniu prawie każdego składnika **WBI** (szczególnie na danych **OSM + predicted**) – poza poszczególnymi składowymi indeksu edukacji (**E_PS** i **E_EM**), ale paradoksalnie, nieco gorszy niż inne modele w prognozowaniu ogólnego **WBI**
- Ogólnie **XGB** można uznać za **najlepszy z analizowanych** nieliniowy model predykcyjny

Różnica w jakości prognoz (MAPE) – OSM + predicted vs. tylko OSM

model	dependent variable								
	H_CD	H_MN	H_LE	HI	E_PS	E_EM	EI	WI	WBI
LASSO	-0.38%	1.31%	-0.04%	0.04%	-2.05%	0.10%	0.00%	0.00%	0.22%
RF	0.26%	-0.19%	0.02%	-0.04%	-0.10%	-0.68%	-0.05%	1.08%	0.24%
SVRpoly	-0.31%	0.49%	-0.10%	-0.11%	-2.86%	0.00%	0.00%	-3.59%	1.37%
SVRradial	-0.96%	-0.39%	-0.33%	-0.68%	-0.33%	0.14%	0.02%	1.51%	-0.09%
XGB	0.69%	-0.54%	0.03%	-0.01%	-1.29%	-0.91%	-0.89%	-2.93%	0.30%

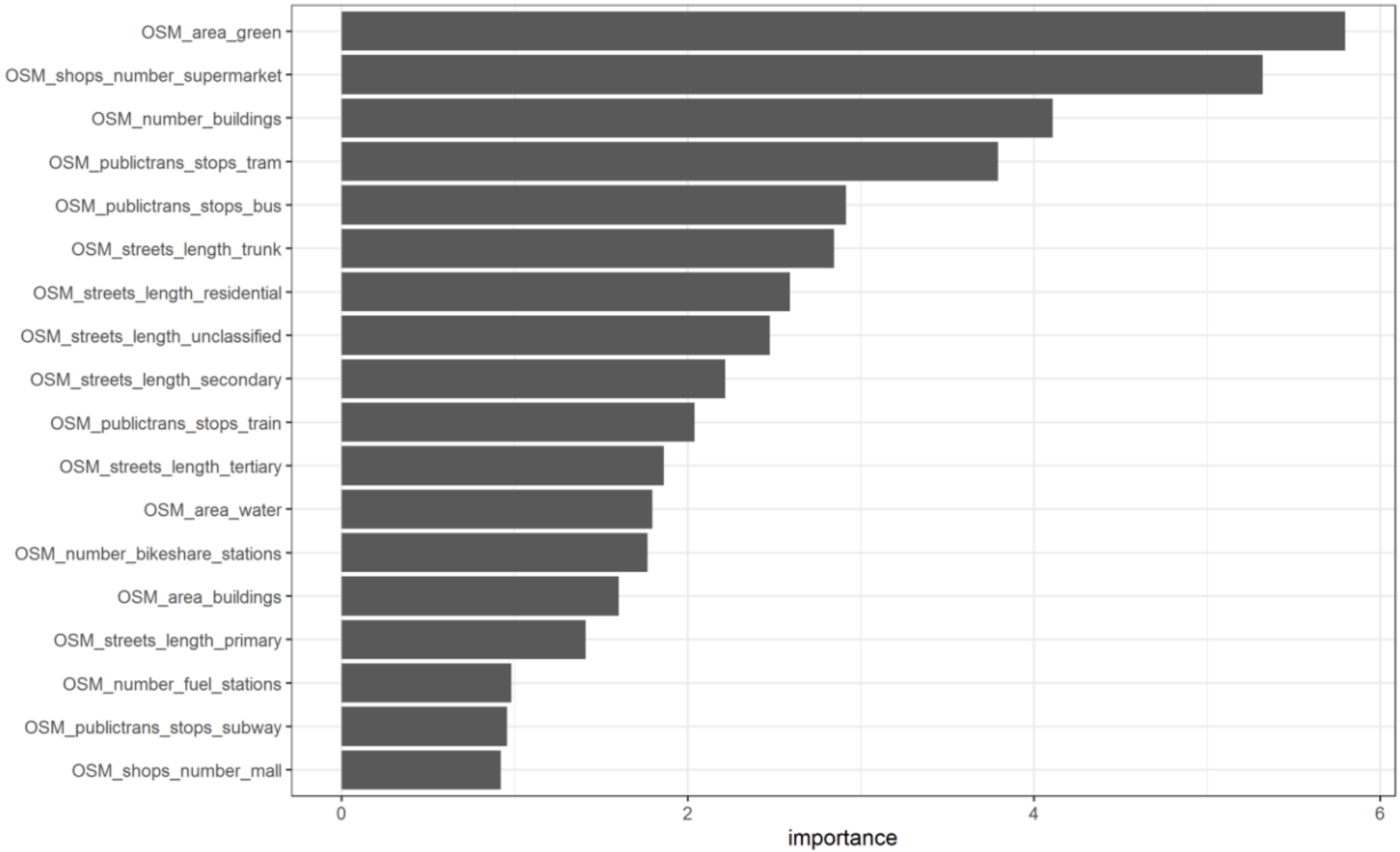
Ujemne wartości różnicy wskazują, że model oparty na cechach **OSM + predicted** miał większą dokładność (niższe MAPE) niż model oparty wyłącznie na danych OSM i *vice versa*

Różnica w jakości prognoz (MAPE) – LASSO vs. nieliniowe

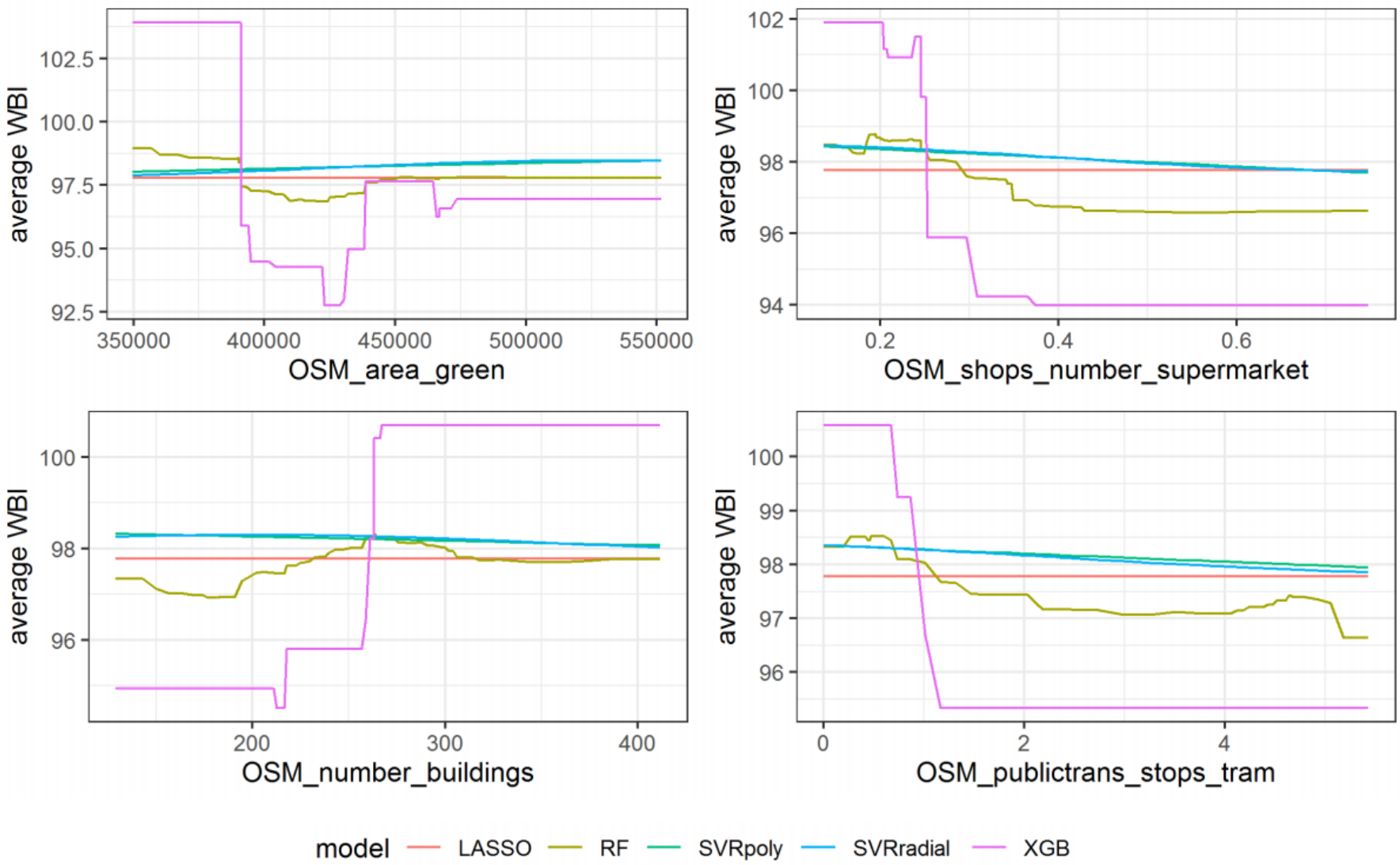
data	model	dependent variable								
		H_CD	H_MN	H_LE	HI	E_PS	E_EM	EI	WI	WBI
OSM	LASSO	13.58%	5.33%	1.80%	5.99%	8.41%	5.28%	5.55%	48.99%	10.09%
	XGB	9.53%	4.82%	1.24%	4.81%	7.64%	6.38%	5.89%	35.99%	10.35%
	average non-linear	11.91%	5.52%	1.62%	5.80%	8.03%	5.95%	5.77%	41.09%	9.78%
	non-linear - LASSO	-1.67%	0.19%	-0.19%	-0.19%	-0.37%	0.67%	0.22%	-7.91%	-0.31%
	XGB - LASSO	-4.05%	-0.51%	-0.56%	-1.18%	-0.77%	1.10%	0.33%	-13.00%	0.26%
OSM + predicted	LASSO	13.21%	6.63%	1.77%	6.03%	6.36%	5.38%	5.55%	48.99%	10.31%
	XGB	10.22%	4.28%	1.27%	4.80%	6.35%	5.47%	5.00%	33.07%	10.65%
	average non-linear	11.83%	5.37%	1.52%	5.58%	6.89%	5.58%	5.54%	40.10%	10.24%
	non-linear - LASSO	-1.38%	-1.27%	-0.25%	-0.45%	0.53%	0.20%	-0.01%	-8.89%	-0.08%
	XGB - LASSO	-2.99%	-2.35%	-0.50%	-1.24%	-0.01%	0.08%	-0.56%	-15.92%	0.34%

Modele nieliniowe (zwłaszcza **XGB**) są w stanie przewidzieć różne wymiary dobrostanu z nie mniejszą dokładnością niż liniowe podejście LASSO (z nielicznymi wyjątkami).

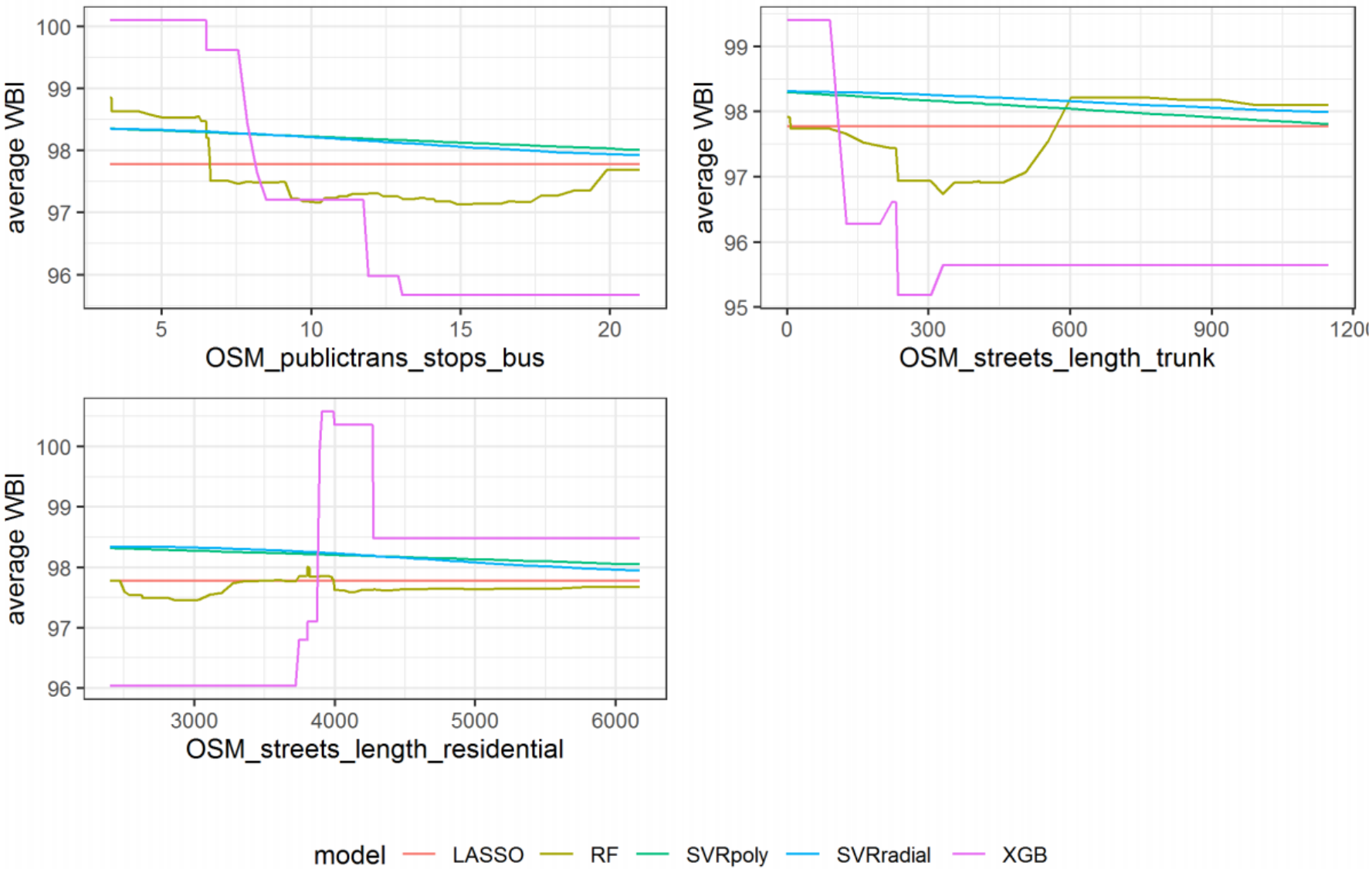
Ważność zmiennych w modelu XGB (OSM + predicted)



Kształt zależności dla najważniejszych predyktorów



Kształt zależności dla najważniejszych predyktorów – c.d.



Wnioski

- Cechy widoczne z kosmosu **pozwalają przewidywać** poziom dobrostanu wewnątrz miasta **z (relatywnie) dużą dokładnością** (potwierdzona 1. hipoteza).
- Cechy wyodrębnione z Open Street Map są **równie dobrymi predyktorami** dobrostanu w mieście, jak cechy wyodrębnione z dziennych zdjęć satelitarnych o wysokiej rozdzielczości (potwierdzona 2. hipoteza).
- **Zależności** między poziomem dobrostanu a jego predyktorami są **dalece nieliniowe**
- Zastosowanie nieliniowych algorytmów uczenia maszynowego nie tylko pozwala dokładniej przewidywać poziom dobrostanu (potwierdzona 3. hipoteza), ale daje również możliwość identyfikacji najważniejszych jego czynników i **kształtu ich zależności** ze zmienną objaśnianą.
- Najważniejsze czynniki **zgodne z wcześniejszymi badaniami i intuicją**: walory przyrodnicze, stopa urbanizacji, gęstość sieci transportowej, dostępność transportu publicznego

Dziękuję

DZIĘKUJĘ ZA UWAGĘ!