

Primerjava metod za izboljšanje prostorsko-časovne ločljivosti geoprostorskih podatkov

Jernej Cukjati, Borut Žalik

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

E-pošta: jernej.cukjati@um.si

The comparison of the methods for increasing the spatiotemporal resolution of geospatial data

This paper presents a statistical analysis and comparison of six approaches, which increase the spatio-temporal resolution of geospatial data. These approaches are: our method, Geographically and Temporally Weighted Generalized Regression Neural Network, Adaptive Graph Convolutional Recurrent Network, Attention Temporal Graph Convolutional Network, Message Passing Neural Networks with Long Short-Term Memory, and the Deep Ensemble Machine Learning framework. Our method performs the best, while the second most accurate method consisted of Message Passing Neural Networks and Long Short-Term Memory.

1 Uvod

Geoprostorske podatke pridobivamo iz različnih virov; najpogosteje iz lokalnih senzorjev in iz satelitskih instrumentov. Pri tem so podatki zajeti z različnimi ločljivostmi [1]. Lokalni senzorji imajo visoko frekvenco zajemanja podatkov, njihova pomanjkljivost pa je redka postavitev [9]. V nasprotju z njimi sateliti zagotavljajo veliko prostorsko pokritost, a nizko časovno ločljivost (do dvakrat na dan). Zaradi premajhnega števila lokalnih senzorjev in nizke časovne ločljivosti satelitskih slik so bile razvite metode, ki izračunajo vrednosti okoljskih spremenljivk na izbrani lokaciji v izbranem časovnem trenutku [5]. Te metode zapolnjujejo vrzeli, pri zajemu senzorskih podatkov.

V tem članku bomo med sabo primerjali različne metode za izračun izbrane okoljske spremenljivke, ki so bile predstavljene v zadnjih treh letih. Te so: naša metoda [8], geografsko-časovno obtežena nevronska mreža (angl. Geographically and Temporally Weighted Generalized Regression Neural Network, GTW-GRNN) [4], prilagodljiva grafavska konvolucijska povratna nevronska mreža (angl. Adaptive Graph Convolutional Recurrent Network, AGCRN) [3], časovno grafavska konvolucijska nevronska mreža z mehanizmom pozornosti (angl. Attention Temporal Graph Convolutional Network, A3T-GCN) [6], nevronska mreža za izmenjevanje sporočil (angl. Message Passing Neural Networks, MPNN), združeno z mrežo LSTM [7] in globoko ansambelsko strojno učenje (angl. Deep Ensemble Machine Learning framework, DEML

framework) [10]. Z omenjenimi metodami bomo povečali prostorsko-časovno ločljivost vrednosti koncentracije NO₂. Podatke meritev NO₂ iz lokalnih senzorjev bomo preslikali v vrednosti NO₂ pikselov obdelanih satelitskih slik. Točnost izračunov bomo nato ovrednotili z metriko povprečne kvadratne napake (angl. Root Mean Square Error, RMSE) in metode med sabo primerjali z izračunom točnosti vrednosti v istoležnih pikslih, v izbranih časovnih trenutkih. Statistično zančilnost razlik vrednosti RMSE metode predlagane v [8] z ostalimi bomo ocenili z izračunom korelacijskih koeficientov.

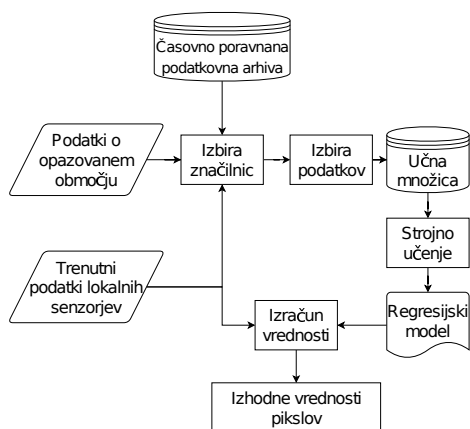
2 Opis metod

V članku primerjamo šest aktualnih metod:

- Našo metodo [8],
- GTW-GRNN [4],
- AGCRN [3],
- A3T-GCN [6],
- MPNN + LSTM [7] in
- DEML [10].

2.1 Naša metoda

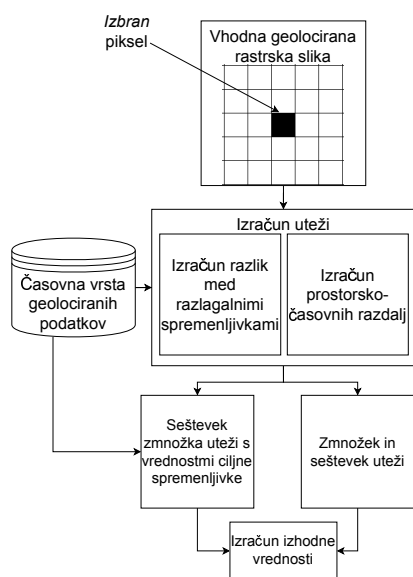
Naša metoda [8] pretvori prostorsko redke vrednosti, izmerjene z lokalnimi senzorji v vrednosti pikselov satelitske slike. Vhod v metodo so podatki o opazovanem območju, podatki trenutnih meritev lokalnih senzorjev in časovno poravnana arhiva preteklih podatkov meritev iz lokalnih senzorjev ter satelitskih slik. Kot je prikazano na sliki 1, je metoda sestavljena iz štirih korakov: izbire značilnic, izbire podatkov, strojnega učenja in izračuna vrednosti. V koraku izbire značilnic metoda rasterizirano opazovano območje razdeli na Voronoijeve celice, pri čemer so središča v danem trenutku aktivni lokalni senzorji. Glede na posamezno Voronoijevo celico metoda določi središčni senzor ter njegove naravne sosede. Pri izbiri podatkov vzame podatke meritev prej izbranih senzorjev iz obeh arhivov in zgradi učno množico. Metoda uporabi učni množici za učenje relacij med podatki iz lokalnih senzorjev v podatke meritev iz satelitskih slik. Na koncu z zgrajenim regresijskim modelom izračuna vrednosti nabora pikselov znotraj posamezne Voronoijeve celice.



Slika 1: Naša metoda

2.2 Metoda GTW-GRNN

Li s sodelavci [4] je predstavil GTW-GRNN. Kot je prikazano na sliki 2, so avtorji pri izračunu vrednosti izbrane spremenljivke uporabili časovno vrsto geolociranih podatkov iz različnih virov. Časovno in prostorsko poravnani podatki so bili vzorci učne množice. GTW-GRNN je uporabila učno množico pri izračunu vrednosti izbranega piksla vhodne geolocirane rastrske slike opazovanega območja v izbranem časovnem trenutku. Ta slika za vsak piksel vsebuje vrednosti razlagalnih spremenljivk. Metoda izračuna vrednost ciljne spremenljivke glede na uteži izbranega piksla iz vhodne geolocirane rastrske slike. Uteži piksla se izračunajo z razliko razlagalnih spremenljivk in evklidskimi prostorsko-časovnimi razdaljami v primerjavi s posameznim učnim vzorcem. Metoda nato vsebuje dve enoti, kjer sešteje zmnožke uteži oz. jih pomnoži še z vrednostmi ciljne spremenljivke iz učne množice. Izhod metode je količnik teh dveh vsot.



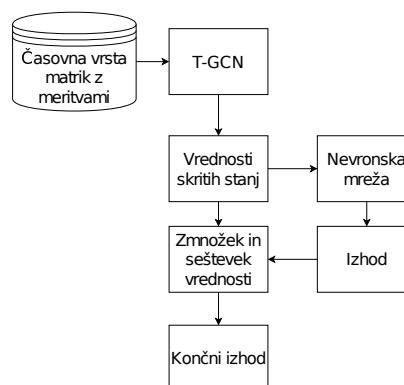
Slika 2: Metoda GTW-GRNN

2.3 Metoda AGCRN

V [3] so predlagali dve dopolnitvi k metodi GCN. Avtorji so ju z GRU združili v AGCRN. Osrednja metoda napoveduje časovne vrste podatkov meritev v vozliščih vhodnih grafov. Avtorji so v osnovno metodo dodali mehanizem prilagajanja uteži za posamezno vozlišče. Druga rešitev je izračun prostorskih relacij med posameznimi časovnimi vrstami vhodnih podatkov.

2.4 Metoda A3T-GCN

A3T-GCN [6] temelji na T-GCN, ki napoveduje vrednosti opazovane spremenljivke v izbranih grafovskih vozliščih. Kot je prikazano na sliki 3, so avtorji časovno vrsto matrik s podatki meritev uporabili v metodi T-GCN. Vrednosti skritih stanj iz T-GCN so vključili v nevronska mrežo z dvema skritima plastema, ki je imela na izhodu aktivacijsko funkcijo *Softmax*. Nevronska mreža je vrnila vrednosti, s katerimi so pomnožili prej omenjene vrednosti skritih stanj in jih sešteli, da so pridobili končno izhodno vrednost.

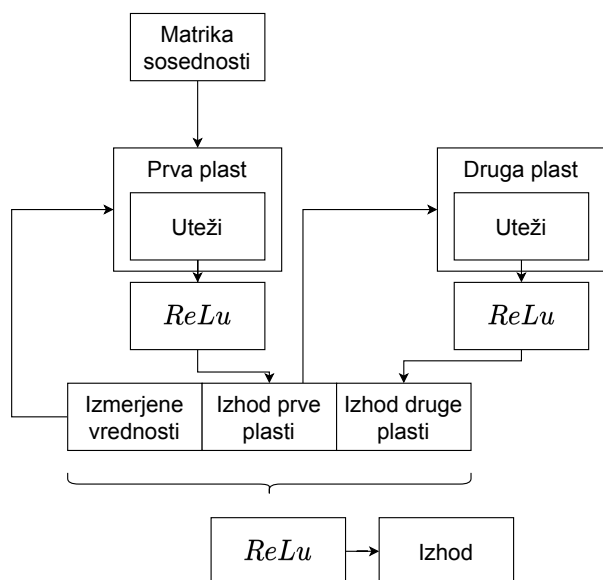


Slika 3: Metoda A3T-GCN

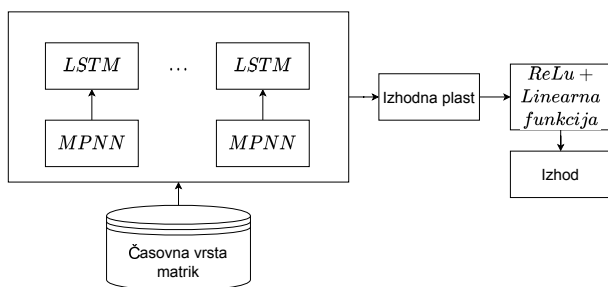
2.5 Metoda MPNN+LSTM

Panagopoulos s sodelavci [7] je združil MPNN z mrežami LSTM. Kot je prikazano na sliki 4, so MPNN sestavljene iz dveh plasti. Prva prejme matriko izmerjenih vrednosti in matriko sosednosti. Matrika izmerjenih vrednosti je v vsaki vrstici vsebovala vektorje podatkov meritev opazovane spremenljivke v posameznem vozlišču. Matrika sosednosti pa je za posamezen časovni trenutek hranila prostorsko razdaljo med posameznimi vozlišči. Naučene uteži iz prve plasti metoda preslika s funkcijo *Relu*. Druga plast na vohodu prejme preslikane uteži iz prejšnje plasti in vrne uteži, ki jih metoda zopet preslika. MPNN združi matriko izmerjenih vrednosti z izhodi posamezne plasti in jo z nelinearno funkcijo preslika v končno vrednost.

Metodo MPNN so avtorji združili z mrežami LSTM, kot je prikazano na sliki 5. Matrike podatkov (izmerjene vrednosti in matrike sosednosti) v izbranem časovnem obdobju so vključili v nevronska mreža MPNN z dvema dodatnima plastema nevronske mreže LSTM. Izhodne vrednosti so združili v izhodni plasti z ostalimi vhodnimi matrikami in jih preslikali s funkcijo *ReLU* ter linearno funkcijo.



Slika 4: Metoda MPNN



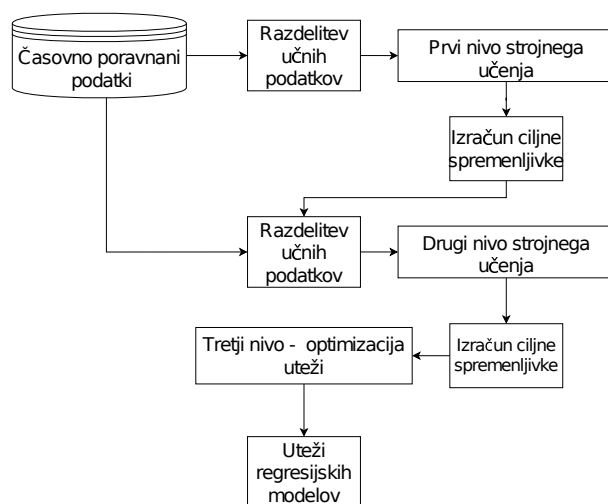
Slika 5: Metoda MPNN + LSTM

2.6 Metoda DEML

Yu s sodelavci [10] je predlagal DEML. Vhodne podatke meritev so avtorji interpolirali in časovno povprečili na dnevno ločljivost. Kot je prikazano na sliki 6, je metoda sestavljena iz treh nivojev. Vhodne podatke so avtorji na začetku delili na 10 podmnožic, ki se izmenjaje uporabljajo pri strojnega učenja na prvem nivoju ansambla. Naučene regresijske modele metoda uporabi za izračun vrednosti ciljne spremenljivke. Prvi nivo ansambla vsebuje štiri metode strojnega učenja: algoritem pospeševanja gradientov, metodo podpornih vektorjev, naključni gozd in XGBoost. DEML izračunane vrednosti iz prvega nivoja in prvotne učne podatke zopet razdeli in uporabi za učenje regresijskih modelov na drugem nivoju. Ta je sestavljen iz naključnega gozda, XGBoost in posplošenega linearnega modela. Izračun regresijskih modelov drugega nivoja vstavijo v tretji nivo, kjer z metodo nenegativnih najmanjših kvadratov naredijo optimizacijo uteži. Zgrajen regresijski model se uporabi pri računanju ciljne spremenljivke.

3 Ovrednotenje metod

Metode, predstavljene v drugem poglavju, smo uporabili za izračun toplogrednega plina NO_2 . Vhodne podatke



Slika 6: Metoda DEML

smo omejili na opazovano območje republike Slovenije v časovnem obdobju od 1. januarja 2020 do 1. junija 2021.

3.1 Vhodni podatki

Prosto dostopne vrednosti koncentracije NO_2 izmerjene z lokalnimi senzorji smo dobili na spletni platformi [11]. Satelitske podatke smo prenesli s programskim modulom SentinelSat [12]. To so bile obdelane geolocirane rastrske slike, zajete z instrumentom satelita Sentinel-5P. Posamezen piksel slike je vseboval časovno značko, vrednost opazovane spremenljivke in atribut *qa_value*. Atribut *qa_value* ima razpon med 0 in 1 ter označuje morebitno neveljavnost podatka. Mi smo uporabili piksele, ki imajo *qa_value* višji od vrednosti 0,75. Satelitske rastrske slike smo zaradi neenotne ločljivosti pikselov prevzorčili na enakomerno mrežo s prostorsko ločljivostjo povprečnega piksela satelitske slike, ki je velikosti $5,5 \times 5,5$ km. Podatke iz obeh virov smo pretvorili na enoten koordinatni sistem in časovni pas.

3.2 Konfiguracija metod

Kot smo pokazali v [8] je naša metoda vračala najboljše rezultate, v kombinaciji z večplastno nevronske mrežo. Ta je bila sestavljena iz dveh skritih plasti, pri čemer je mreža vrednosti iz prve plasti preslikala s funkcijo Softmax. Mrežo smo učili 200 epoh, z velikostjo paketa 16 učnih vzorcev in algoritmom Adam. Vhodne vrednosti razlagalnih spremenljivk smo pred strojnem učenjem preslikali na razpon med -1 in 1, vrednosti iz pikselov satelitskih slik pa smo standardizirali. Izračunane vrednosti ciljne spremenljivke smo na koncu preslikali nazaj v prvotni razpon.

Ker ostale metode predvidevajo popolno veljavnost vhodnih podatkov smo vrednosti meritev iz lokalnih senzorjev, na začetku interpolirali z metodo *IDW*. Za optimalno delovanje metode GTW-GRNN smo s prečno validacijo določili parametre:

- $\lambda = 87800$,
- $h_{st} = 6$ in

- $spread = 0,49$.

AGCRN, A3T-GCN in MPNN+LSTM na vходу zahtevajo graf. V našem primeru so bila vozlišča grafa piksli opazovanega območja. Povezave pa so bile evklidske razdalje med središči pikslov. Manjkajoče podatke meritev smo interpolirali po času z linearno interpolacijo. Nevronske mreže smo učili 100 epoch z algoritmom Adam, hitrostjo učenja 0,001 in funkcijo izgube $L1$.

Strojno učenje smo izvajali nad podatki meritev, zajetimi med 1. januarjem 2020 in 18. februarjem 2021. Z naučenimi regresijskimi modeli smo izračunali ciljne vrednosti glede na testno množico podatkov meritev iz lokalnih senzorjev. Dobljene vrednosti smo na lokacijah, kjer ni aktivnih lokalnih senzorjev, primerjali z dejanskimi vrednostmi pikslov. Kot je prikazano v tabeli 1, smo z našo metodo dosegli boljše rezultate kot z ostalimi.

Tabela 1: Primerjava točnosti izračuna vrednosti NO_2 [mol/m^2] z različnimi metodami

Metoda	RMSE ($\times 10^{-6}$)
Naša metoda	15,3
GTW-GRNN	19,2
AGCRN	17,6
A3T-GCN	16,8
MPNN+LSTM	16,0
DEML	18,7

Za določitev statistične značilnosti razlike med našo metodo, in ostalimi smo uporabili metodo T - T est. Ničelna hipoteza je predpostavljala, da ni signifikantnih razlik med predlagano metodo in referenčnimi metodami. Najprej smo za izračunano in dejansko koncentracijo NO_2 nabora pikslov vsake satelitske slike določili RMSE. Pri tem smo upoštevali samo vrednosti pikslov, v katerih ni lokalnih senzorjev. Z metodo T - T est smo izračunali verjetnost p dobljene razlike RMSE ob predpostavki, da ničelna hipoteza drži. Manjši kot je p , manjša je verjetnost, da ničelna hipoteza drži. Skladno s priporočili [2] smo za mejno vrednost statistične značilnosti uporabili $\alpha = 0,05$. Kot je prikazano v tabeli 2, so vse izračunane vrednosti p manjše od α .

Tabela 2: Izračunana vrednost p z metodo T - T est med vrednostmi RMSE metode predlagane v [8] in referenčnih metod

Metoda	Vrednost p
GTW-GRNN	$2,07 \times 10^{-11}$
AGCRN	$2,36 \times 10^{-10}$
A3T-GCN	$4,78 \times 10^{-9}$
MPNN+LSTM	$1,30 \times 10^{-5}$
DEML	$2,60 \times 10^{-10}$

4 Zaključek

V prispevku smo med sabo primerjali našo metodo [8] s petimi aktualnimi metodami za izboljšanje prostorsko-časovne ločljivosti okoljskih geoprostorskih podatkov. Vse metode smo testirali na istem opazovanem območju in

istimi vhodnimi podatki. Točnost rezultatov smo ovrednotili z metriko RMSE in med izračunanimi vrednostmi pokazali statistično značilno razliko. To smo naredili z izračunom korelacijskega koeficienta, ki smo ga dobili z metodo T - T est.

5 Zahvala

To delo je omogočil raziskovalni program št. P2-0041, ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna po pogodbi o financiranju podiplomskega raziskovalnega usposabljanja mladega raziskovalca številka 0796-53590.

Literatura

- [1] B. P. L. Lau in sod. "A survey of data fusion in smart city applications". V: *Inf. Fusion* 52 (2019), str. 357–374.
- [2] R. L. Wasserstein, A. L. Schirm in N. A. Lazar. "Moving to a World Beyond " $p < 0.05$ ". V: *Am. Stat.* 73.sup1 (2019), str. 1–19.
- [3] L. Bai in sod. "Adaptive graph convolutional recurrent network for traffic forecasting". V: *Adv. Neural Inf. Process. Syst.* 2020.NeurIPS (2020), str. 1–12.
- [4] T. Li, Y. Wang in Q. Yuan. "Remote Sensing Estimation of Regional NO_2 via Space-Time Neural Networks". V: *Remote Sens.* 12.16 (2020), str. 2514.
- [5] S. Salcedo-Sanz in sod. "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources". V: *Inf. fusion* 63 (2020), str. 256–272.
- [6] J. Bai in sod. "A3t-gcn: Attention temporal graph convolutional network for traffic forecasting". V: *ISPRS Int. J. Geo-Information* 10.7 (2021), str. 485.
- [7] G. Panagopoulos, G. Nikolentzos in M. Vazirgiannis. "Transfer Graph Neural Networks for Pandemic Forecasting. (S. Singh, Ed.; S. Markovitch, Ed.)" V: *Proceedings of the AAAI Conf. Artif. Intell. AAAI. Zv.* 35. 6. 2021, str. 4838–4845.
- [8] J. Cukjati in sod. "IoT and Satellite Sensor Data Integration for Assessment of Environmental Variables: A Case Study on NO_2 ". V: *Sensors* 22.15 (2022).
- [9] M. V. Narayana, D. Jalihal in S. M. Shiva Nagendra. "Establishing A Sustainable Low-Cost Air Quality Monitoring Setup: A Survey of the State-of-the-Art". V: *Sensors* 22.1 (2022), str. 394.
- [10] W. Yu in sod. "Deep Ensemble Machine Learning Framework for the Estimation of $\text{PM}_{2.5}$ Concentrations". V: *Environ. Health Perspect.* 130.3 (2022), str. 1–11.
- [11] *okolje.info*. Dostopno na: <http://www.okolje.info/>. (Pridobljeno 4. 1. 2023).
- [12] *SentinelSat documentation*. Dostopno na: <https://sentinelSat.readthedocs.io/en/stable/>. (Pridobljeno 4. 1. 2023).