A neural network regression model for estimating maximum daily air temperature using Landsat-8 data

A. Nascetti [1*], C. Monterisi [1], F. Iurilli [1], A. Sonnessa [1]

[1] Department of Civil, Environmental, Land, Construction and Chemistry (DICATECh), Polytechnic University of Bari, Italy

Commission III, ICWG III/IVc

KEY WORDS: UHI, Neural Network, Land Surface Temperature, Air Temperature, Regression Models

ABSTRACT:

Urban Heat Islands (UHI) phenomenon is a pressing problem for highly industrialized areas with serious risks for public health. Weather stations guarantee long-term accurate observations of weather parameters, such Air Temperature (AT), but lack appropriate spatial coverage. Numerous studies have argued that satellite Land Surface Temperature (LST) is a relevant parameter for estimating AT maps, exploring both linear regression and Machine Learning algorithms. This study proposes a Neural Network (NN) regression model for estimating the maximum AT from Landsat-8 data. The approach has been tested in a variegated morphological region (Puglia, Italy) using a large stack of data acquired from 2018 to 2020. The algorithm uses the median values of LST and Normalized Difference Vegetation Index (NDVI) computed using different buffer radius around the location of each reference weather station (250 m, 1000 m, and 2000 m) to train the NN model with a K-fold cross-validation strategy. The reference dataset was split into three sets using a stratified sampling approach considering the different station categories: rural, High- and Low-density Urban areas respectively. The algorithm was tested with different learning rates (LR) (0.001 and 0.005). The results show that our NN model accuracy improves with the increase of the buffer radius, minimizing the difference in terms of R^2 between training and evaluation data, with an overall accuracy consistently higher than 0.84. Future research could investigate more input variables in the NN model such as morphology or climate variables and test the algorithm on larger areas.

1. INTRODUCTION

The Urban Heat Islands (UHI) phenomenon is a pressing problem in highly industrialized areas with high population pressure, as its effects are amplified by climate change. UHI occurs when the temperature in urban areas is higher than that of surrounding rural areas (Chen et al. 2020). It is generated by the greater absorption of electromagnetic energy and the slow cooling of the urbanized surfaces compared to the surrounding areas with vegetation (Imhoff et al. 2010). The increase in heat creates risks for public health, rising the percentage of city mortality due to high thermal stress (Stocker et al. 2013). An accurate, continuous and multi-temporal study of UHI could help urban planners to manage associated risks. For this purpose, it is crucial to have reliable and accurate data.

Weather stations guarantee long-term accurate observations of weather parameters but their ability to describe the spatial variations of the Air Temperature (AT) in heterogeneous areas, such as cities, is limited by their lack of appropriate spatial coverage (Ding et al). On the other hand, Earth Observation (EO) satellites collect data on the interaction of solar energy with the earth's surface in terms of reflected, absorbed and transmitted energy, continuously and for decades. Numerous studies have argued that satellite Land Surface Temperature (LST) is highly appropriate to determine the radiative load of the Earth's surface at large scale and to estimate air temperature maps (Benali et al 2012, Yoo et al 2018, Xu et al 2014). Several studies have explored different types of regression algorithms to estimate AT from LST data: from the simple linear regression methods (Bechtel et al. 2017) to the more advanced Machine Learning algorithms such as Random Forest (Dos Santos et al. 2020, Zhang et al. 2016) Support Vector Machine (Yoo et al. 2018) and Neural Networks (Jang et al. 2004, Ho et al. 2014, Zeng et al. 2021) also using auxiliary data for the AT estimation, like Normalized Difference Vegetation Index (NDVI), Digital Elevation Model (DEM) and zenith angle of the Sun (Jang et al. 2004, Zeng et al. 2021, Halder et al. 2021).

Most of the studies have focused on using medium resolution data (e.g. MODIS) and have concentrated on regions characterized by the presence of large cities. However, these studies, with coefficient of determination or goodness of fit (R^2) values between 0.54 and 0.68, do not reach acceptable accuracy levels for continuous monitoring of AT variations.

Furthermore, although UHI is currently the most documented phenomenon of climate change in the urban environment, the studies applied in areas with a Mediterranean climate have mainly focused on cities with a high population density, neglecting the phenomenon in small cities (Otgonbayar et al. 2019).

Therefore, due to the complex behavior of the phenomenon and its dependence on variegated urban characteristics, it is essential to broaden this area of research, including cities with large, medium and small population densities and thus studying their behavioral correlations.

2. OBJECTIVE AND METHOD

The objective of this study is to assess the relevance of a Neural Network (NN) regression model for estimating the maximum AT from Landsat-8 data in a variegated morphological region (Puglia, Italy) characterized by small to medium-size cities. In particular, the selection of the Apulia region as a study area is

* Corresponding Author: andrea.nascetti@poliba.it

justified by its climatic diversity, in terms of temperature, speed and wind direction. Characterized by a Mediterranean climate, it has quite hot and dry summers and mild and moderately rainy winters, with an abundance of rainfall during the autumn season. A strong spatial variability of rainfall linked to the different spatial/geographic characteristics is associated, in each individual area, with a strong variability of rainfall and wind for the individual data collection stations, as it often occurs in Mediterranean climates (Cotecchia, ISPRA Ambiente, 2017).
Moreover, the Apulian territory is relevant from the point of view of the density of urban agglomerations present in the territory which combines rural areas, low-density urbanized areas and densely urbanized areas.

In order to evaluate the spatial distribution of AT over the study area, temperature maps were generated for 6 two-month periods per year (January-February, March April, May-June, July-August, September-October and November-December), computing pixel-based median of LST value (see Fig. 1).

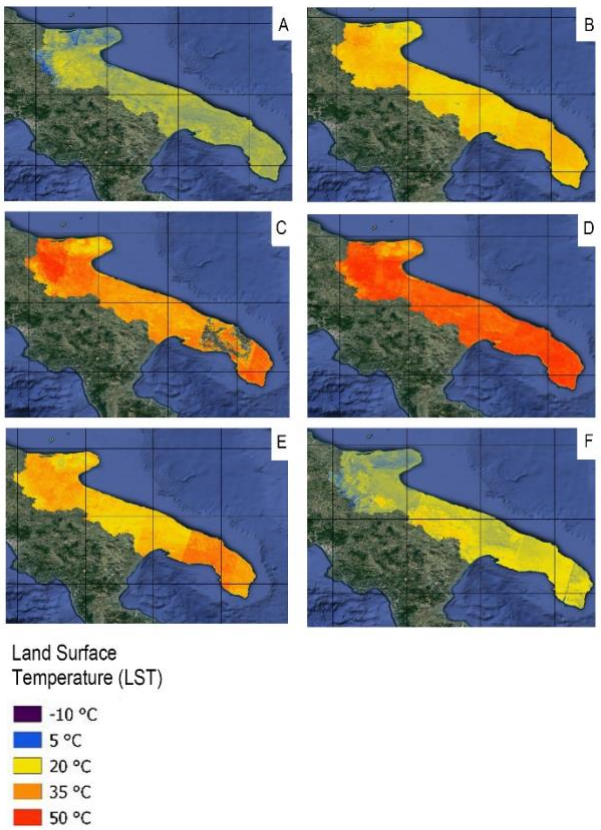

Fig. 1: Land Surface Temperature maps. A. January-February 2018, B. March-April 2018, C. May-June 2018, D. July-August 2018. E. September-October 2018, F. November-December 2018

We used the LST and the NDVI derived by the Landsat-8 imagery and the acquisition time converted in suitable time-periodic variables as input variables for the regression model. We used available weather stations AT data distributed throughout the Apulian territory to train and validate the NN regression models. We computed the median value of the LST and the NDVI in a radius area (250m, 1000 m and 2000 m) around the location of each weather station.

In order to manage the periodic acquisition time data in the neural regression model and therefore preserve the cyclical temporal meaning of the succession of days, weeks, months and years, a mathematical expedient was used by converting the acquisition time into two different variables: sin_time and cos_time respectively. In particular, we computed the sin_time and cos_time starting from the day-of-year both using the following formulas:

$$\sin_{time} = \sin(2\pi * DayofYear/365)$$
$$\cos_{time} = \cos(2\pi * DayofYear/365) \quad (1)$$

In Fig.2, the scatter plot using sin_time and cos_site for the valid (not cloudy) LST data acquired in 2018 is reported. It is visible that during the end of the spring and summer period (second and third quadrants) there are more available data (from green to yellow color) respect to the autumn and winter periods.

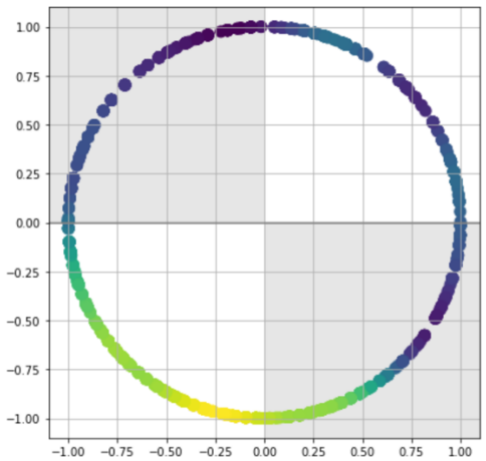

Fig. 2: Sin_time and Cos_time functions, related to the radiant angle

The NN model is therefore based on four input variables computed for each available weather station. We used a relatively small sequential fully connected neural network with six hidden layers having a decreasing number of nodes (i.e. 25, 20, 15, 10, 5,1 nodes), see Fig.3. The NN model returns the estimated air temperature at each location using a pixel-wise buffered approach.

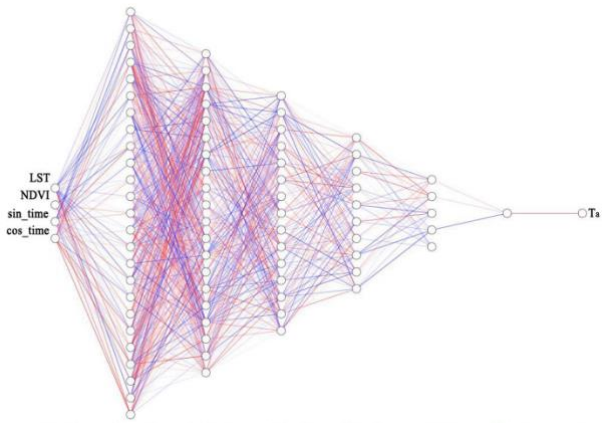

Fig. 3: Neural Network (NN) Graphic diagram

3. DATA AND EXPERIMENTS

All the radiometrically calibrated and atmospherically corrected surface reflectance and land surface temperature images acquired by Landsat-8 OLI/TIRS sensors over the Puglia region (Italy) during the period from 2018 to 2020, with cloud cover less than 50%, were collected using the Google Earth Engine (GEE) (Gorelick et al. 2017) platform. The images are from the USGS Landsat 8 mission Level 2- Collection 2- Tier 1 collection, with a return time of 16 days and a spatial resolution of 30 m. Landsat 8 SR products are created with the Land Surface Reflectance Code (LaSRC). All Collection 2 ST products are created with a single-channel algorithm jointly created by the Rochester Institute of Technology (RIT) and National Aeronautics and Space Administration (NASA) Jet Propulsion Laboratory (JPL).

A pre-processing step was performed to remove cloud covered pixels using the Landsat quality band with GEE Landsat Collection 1 LandsatLook 8-bit Quality Image in-based algorithm. We collected as reference data the daily max air temperature of 31 weather stations distributed throughout the Puglia region (Italy), see Fig.4, during the years 2018, 2019, and 2020, from Apulian civil protection database: https://protezionecivile.puglia.it/centro-funzionale-decentrato/rete-di-monitoraggio/.

We classified the weather stations into three classes according to the surrounding build-up and population densities: high density urban area, low density urban area and rural area (see Fig.4). We imported the weather station locations in the GEE platform as a feature collection.

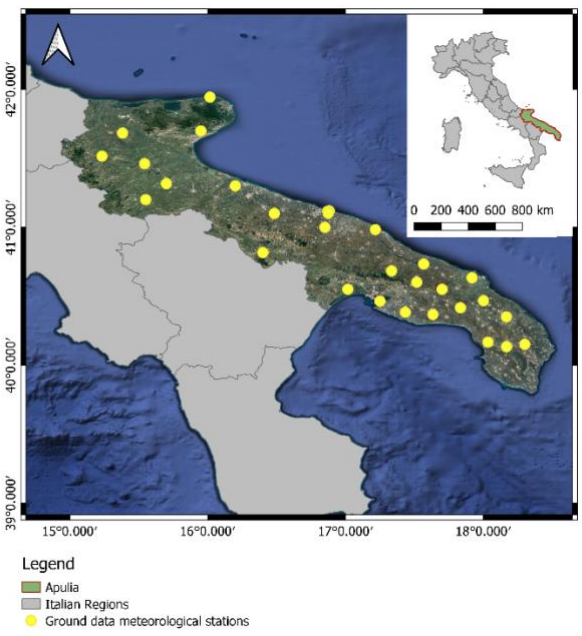

Fig. 4: Study area– Puglia Region (Italy)

Several studies have evaluated the spatial variability of AT data, finding that decisions made based on the AT recorded at a weather station at less than 1 km – 2 km buffer radius have the lowest uncertainty associated with the effect of spatial variability on the air temperature (Quiñones et al. 2018). The radius of influence is considered the maximum distance between the sampling point and the point where the differences in the values of the variables that are measured begin to be significant.
We computed the median values of LST and NDVI using different buffer radius (250 m, 1000 m, and 2000 m) around the location of each station for all the available imagery. Considering the monthly average LST values recorded by meteorological stations ashore during the years under examination, it is generally possible to verify a good correlation between them in the different locations, while the NDVI parameter is characterized by an important variability as it depends on the land cover of the precise location of the station.

We used a K-fold cross-validation strategy (K=3) to train and validate the NN model. In Fig. 5 the overall cross-validation schema with the three combination is depicted.

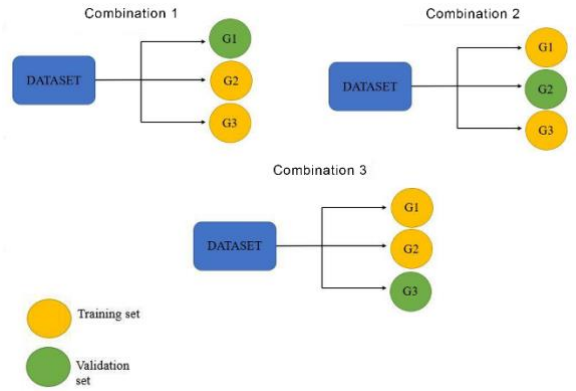

Fig. 5: Dataset cross-validation with three combinations

Using the cloud open source Google Colaboratory platform, we split the reference dataset into three sets using a stratified sampling approach considering the three station categories in relation to their location in rural areas, low density areas and dense urban areas, see Fig. 6. The first group includes the stations of Adelfia, Ascoli Satriano, Barletta, Ceglie Messapica, Corigliano, Galatina, Gravina, Latiano, Lizzano, Manduria, Monte Sant'Angelo, Nardò, Ortanova, Ostuni Palagiano, Pietramontecorvino, Polignano a Mare, Ruvo, San Pietro Vernotico and Spinazzola.
The second group includes the stations of Brindisi, Martina Franca, Peschici, San Pancrazio Salentino and Sansevero, while the third group includes the meteorological stations of Bari Idrografico, Bari Osservatorio, Bari campus, Foggia, Lecce and Taranto.

The experiments were conducted using the 3 different combinations of training data and validation data and varying, for each combination, the buffer size, and the learning rate. We performed several tests with different learning rates (LR) (0.001 and 0.005). We trained the NN model using the Adam optimizer for 10000 epochs for all the K-fold combinations and buffer radius. The cross-validation method consists in dividing the training set into k parts of equal size, and in selecting a part 1/K to use it as a validation set, while the remaining parts k-1/k continue to compose the training set. The procedure is repeated k- times, each time selecting a different subset as validation set which allows to eliminate the problem of overfitting in the training-sets.
An example of the results obtained is presented in the scatter plots Fig.7, with LR=0.001 and LR=0.005, where the consistency in terms of R^2 between the estimated and reference air temperature is shown.

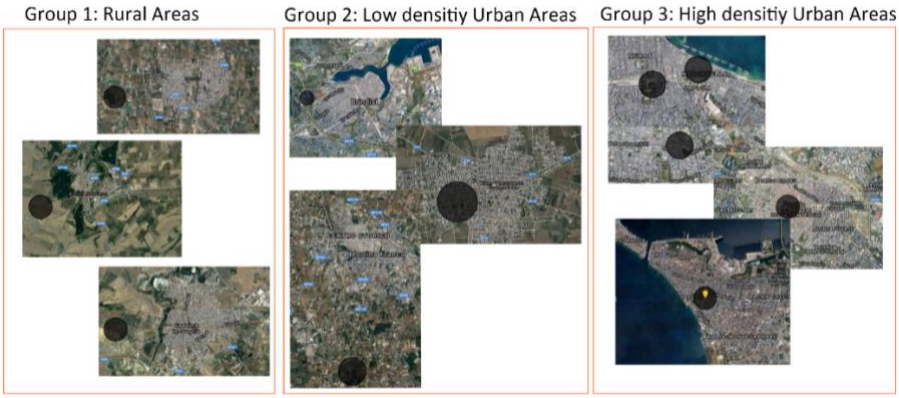

Fig. 6: Urbanization groups.
Group 1: Rural Areas (Adelfia, Ascoli Satriano, Gravina in Puglia); Group2: Low density Urban Areas (Brindisi, San Pancrazio Salentino, Martina Franca); Group 3: High density Urban Areas (Bari, Foggia, Brindisi).

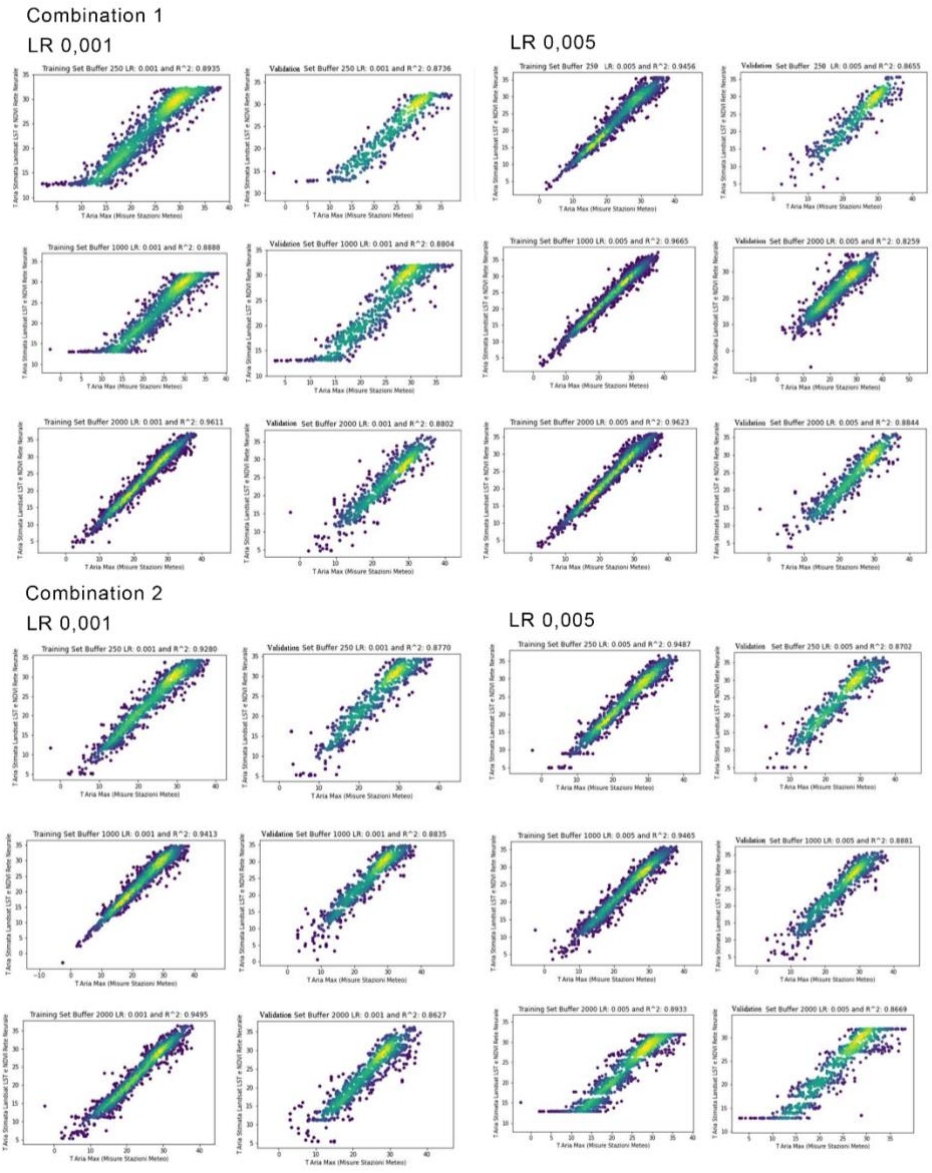

Fig. 7: Scatter plots estimated vs reference max air temperature with LR=0.001
(left column training, right column validation) – data combination 1 and data combination 2

In order to verify the network training process and test the accuracy of the method, two statistical parameters were considered: the Root Mean Square Error (RMSE) and the coefficient of determination or goodness of fit (R^2).

Fig. 8 represents the two learning curves of the first K-fold combination considering that (blue curve refers to the validation data and orange one refers to the training data) obtained with LR=0.001 (left) and LR=0.005 (right) respectively. The analysis of the graphs shows a more robust training in using the lower learning rate (less overfitting). In Tab.1 and Tab.2 all the R^2 results obtained with the two different learning rates, for all the combinations and for all the different buffer radius are presented.

By globally analyzing the average R^2 values obtained from the 3 combinations of data, as the learning rate varies (Table 2, Table 3), average values of R^2 are always greater than 0.8 and, in particular, between 0.9038 and 0.9482 for training data, and between 0.8478 and 0.8763 for validation data. The results show that using a lower LR=0.005 the NN Model improve the fitting in term of R^2 for all the combinations and buffer values. Moreover, the lower Learning Rate slightly improves the model's overall accuracy, which is always higher than 0.84 and generally better than the models presented in the literature.

The effect of the variation of the buffer size is not evident if the experimental results. The NN model accuracy slightly improves with the increase of the buffer radius only in few cases. A more stable behavior of the model could have been expected with the increase of the buffer, considering that maximum AT remains almost constant in a neighborhood of 2000 m from the weather reference station representing a more low-frequency spatial variation of the UHI with respect to the LST (Quiñones et al.).

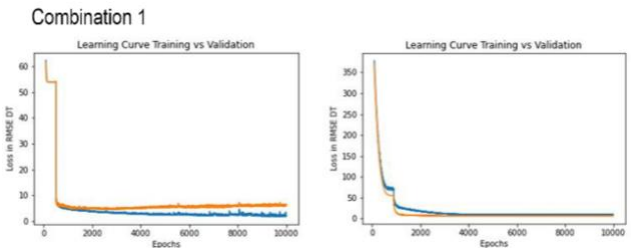

Fig. 8: Learning curves estimated vs reference max air temperature with LR=0.001 on the left and with LR=0.005 on the right- Combination 1
(left column loss in RMSE, right column epochs)

	Training Data - LEARNING RATE 0,001		
	BUFFER: 250 m	BUFFER: 1000 m	BUFFER: 2000 m
Combination 1 - R^2	0,8935	0,8888	0,9611
Combination 2 - R^2	0,928	0,9413	0,9495
Combination 3 - R^2	0,9804	0,8814	0,8854
R^2 Mean value	0,9340	0,9038	0,9320
	Validation Data - LEARNING RATE 0,001		
	BUFFER: 250 m	BUFFER: 1000 m	BUFFER: 2000 m
Combination 1 - R^2	0,8736	0,8804	0,8802
Combination 2 - R^2	0,877	0,8835	0,8627
Combination 3 - R^2	0,8355	0,865	0,8791
R^2 Mean value	0,8620	0,8763	0,8740

Tab. 1: Training and validation values of R^2 over the K-folds using Learning Rate 0.001.

	Training Data - LEARNING RATE 0,005		
	BUFFER: 250 m	BUFFER: 1000 m	BUFFER: 2000 m
Combination 1 - R^2	0,9456	0,9665	0,9623
Combination 2 - R^2	0,9487	0,9465	0,8933
Combination 3 - R^2	0,9504	0,8434	0,9831
R^2 Mean value	0,9482	0,9188	0,9462
	Validation Data - LEARNING RATE 0,005		
	BUFFER: 250 m	BUFFER: 1000 m	BUFFER: 2000 m
Combination 1 - R^2	0,8655	0,8259	0,8844
Combination 2 - R^2	0,8702	0,8881	0,8669
Combination 3 - R^2	0,863	0,8295	0,8259
R^2 Mean value	0,8662	0,8478	0,8591

Tab. 2: Training and validation values of R^2 over the K-folds using Learning Rate 0.005.

4. CONCLUSIONS

This study assesses the relevance of a Neural Network (NN) regression model for estimating the maximum AT from Landsat-8 data in a variegated morphological region (Puglia, Italy) characterized by small to medium-size cities. In particular, we used as input variables for the regression model the LST and the NDVI index derived from the Landsat-8 imagery and the corresponding acquisition time converted in suitable time-periodic variables. We computed the median value of the LST and the NDVI in a circular buffer area around the location of each weather station to find the relation between the LST/NDVI and the max AT measured from the weather station.

We developed a complete automated procedure to collect, pre-process the Landsat 8 images and to train the NN model based on two Google cloud tools, Google Earth Engine and Google Colab. The jointly used of these two tools allows to considerably reduce the computational hardware costs as well as those associated to subsequent analyses, thanks to the versatility, replicability and shareability of the programmed code, as well as the provision of GEE of over 100 TB of open data within its dataset.

We performed some experiments using as reference data the daily max air temperature of 31 weather stations distributed mainly throughout the Puglia region (Italy) during the years 2018, 2019, and 2020. We classified the weather stations into three classes according to the surrounding build-up and population densities: high density urban area, low density urban area and rural area. We compute the median values of LST and NDVI using different buffer radius (250 m, 1000 m, and 2000 m) around the location of each station for all the available imagery. We used a K-fold cross-validation strategy (K=3). We split the reference dataset into three sets using a stratified sampling approach considering the three station categories. We performed several tests with different learning rates (LR) (0.001 and 0.005). and for all the K-fold combinations and buffer radius

The subdivision of the training and validation reference station data into three classes takes into account the territory variability of the Puglia region, which would otherwise have included unbalanced demographic and morphological characteristics during the training of the NN model. On the other hand, the choice of this territory represents a key point for experimenting with the method in differentiated areas and with variable climates, laying the basis for broader replicability and scalability of the model itself.

The results show that, with the increase of the buffer from the ground station, the model minimized the difference in term of R^2 between training and evaluation data reducing the overfitting. In addition, the increase in LR slightly increases the accuracy of the model, which in any case is always higher than R^2=0.84 and therefore more performing than the models presented in the literature. The model therefore demonstrates that it can be applied at regional scale and considering long time series of Landsat-8 data.

Future research developments may focus on the influence of other input variables in NN regression models such as Digital Elevation Models (DEMs) or buildings footprint to account for the elevation and morphology of the area or other climate variables such as wind direction and speed or precipitation. Moreover, it would be interesting to test the model on larger areas, perhaps on a national and pan-European scale.

OPEN SOURCE CODE AND DATA

The codes used for data pre-processing and analysis performed in GEE are available of the following links:

- https://code.earthengine.google.com/42a58f181d670f90cfe03bda203813bf
- https://code.earthengine.google.com/21934db6b319f916fc34cf4929eabf44

The Google Colab notebook to train the model and analyses the results is available at the following link:

- https://colab.research.google.com/drive/1iJnBnhYdgJjUf7JHA3zh_S5Rar5Qh50T?usp=sharing

ACKNOWLEDGMENTS

This research resulted from “EO4SDG – Earth Observation for Sustainable Development Goals: Big Data analytics for monitoring global land changes phenomena” project, funded by Fondazione CON IL SUD, as part of the 2018 highly qualified human capital call.
We also thank Prof. Maria Mancilla Garcia for her careful editing on this manuscript.

REFERENCES

Chen, M.; Zhou, Y.; Hu, M.; Zhou, Y. Influence of Urban Scale and Urban Expansion on the Urban Heat Island Effect in Metropolitan Areas: Case Study of Beijing–Tianjin–Hebei Urban Agglomeration. Remote Sens. 2020, 12, 3491.

Imhoff, M.L.; Zhang, P.; Wolfe, R.E.; Bounoua, L. Remote sensing of the urban heat island effect across biomes in the continental USA. Remote Sens. Environ. 2010, 114, 504–513.

Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M.M., Allen, S.K., Boschung, J. et al., 2014. Climate Change 2013: The physical science basis. contribution of working group I to the fifth assessment report of IPCC the intergovernmental panel on climate change.

Ding, L., Zhou, J., Zhang, X., Liu, S., Cao, R., 2018. Downscaling of surface air temperature over the Tibetan Plateau based on DEM. Int. J. Appl. Earth Obs. Geoinf. 73, 136–147.

Benali, A., Carvalho, A.C., Nunes, J.P., Carvalhais, N., Santos, A., 2012. Estimating air temperature in Portugal using MODIS LST data. Remote Sens. Environ. 124, 108–121.

Yoo, C., Im, J., Park, S., Quackenbush, L.J., 2018. Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data. ISPRS J. Photogramm. Remote. Sens. 137, 149–162.

Xu, W., Knudby, A., Ho, H.C., 2014. Estimating daily maximum air temperature from MODIS in British Columbia, Canada. Int. J. Remote Sens. 35 (24), 8108–8121.

Bechtel, B., Zaksek, K., Oßenbrugge, J., Kaveckisa, K., Bohnera, J., 2017. Towards a satellite based monitoring of urban air temperatures. Sustain. Cities Soc. 34, 22–31.

Dos Santos, R. S, 2020. Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. Int. J. Appl. Eart Obs Geoinformation. 88, 102066

Zhang, H., Zhang, F., Ye, M., Che, T., Zhang, G., 2016. Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data. J. Geophys. Res. Atmos. 121 (19).

Jang, J.D., Viau, A.A., Anctil, F., 2004. Neural network estimation of air temperatures from AVHRR data. Int. J. Remote Sens. 25, 4541–4554.

Ho, H.C., Knudby, A., Sirovyak, P., Xub, Y., Hodul, M., Henderson, S.B., 2014. Mapping maximum urban air temperature on hot summer days. Remote Sens. Environ. 154, 38–45.

Zeng, L., Hu, Y., Wang, R., Zhang, X., Peng, G., Huang, Z., Zhou, G., Xiang, D., Meng, R., Wu, W., Hu, S., 2021. 8-Day and Daily Maximum and Minimum Air Temperature Estimation via Machine Learning method on a Climate Zone to Global Scale. Remote Sens. 2021, 13(12), 2355.

Halder, B., Bandyopadhyay, J., Banik, P., 2021, Monitoring the effect of urban development on urban heat island based on remote sensing and geo-spatial approach in Kolkata and adjacent areas, India. Sustainable Cities and Society 74, 103186.

Otgonbayar, M.; Atzberger, C.; Mattiuzzi, M.; Erdenedalai, A., 2019. Estimation of Climatologies of Average Monthly Air Temperature over Mongolia Using MODIS Land Surface Temperature (LST) Time Series and Machine Learning Techniques. Remote Sens. 2019, 11, 2588.

Cotecchia, V. Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA Ambiente), 2017. Descriptive memories of the Geological map of Italy. Groundwater and marine intrusion in Puglia: from research to emergency in safeguarding the resource Volume 92.

Gorelick et al., Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment, 2017

Quiñones A. J. P., Cordoba B. C., Gutierrez M. R. S., Keller M., Hoogenboom G., 2018. Radius of influence of air temperature from automated weather stations installed in complex terrain. Theoretical and Applied Climatology