Machine learning-based emulation of land cover effects at sub-hectometric scale using crowd-sourced weather observations

Andrei Covaci[1], Thomas Vergauwen[2,3], Sara Top[3], Steven Caluwaerts[2,3], Lesley De Cruz[1,3]

(1) Vrije Universiteit Brussels (VUB), Brussels, Belgium (2) Ghent University (UGent), Ghent, Belgium (3) Royal Meteorological Institute of Belgium (RMI), Brussels, Belgium

Introduction

Traditional numerical weather prediction models assume a standardized "short grass, open field" environment. Such an environment is far from representative of where most people live. Moreover, despite advances in **urban climate modelling**, even state-of-the-art weather forecasts and climate scenarios do not account for the **hyperlocal influence of land cover** on meteorological variables.

To bridge this gap, **machine learning (ML) models**, such **as random forest (RF)** or neural networks, can be used to **correct standardized temperature** (e.g., from analysis, forecast, or even climate projection data) **for non-standardized environments**, as demonstrated by research from Venter et al. [1].

The crowdsourced data used in this work is the **VLINDER network**, managed by Ghent University. This network consisted of 60 stations in 2020 and was expanded to **78 weather stations** in 2023, placed in non-standardized environments, and equipped with **high-quality, calibrated sensors** [2].

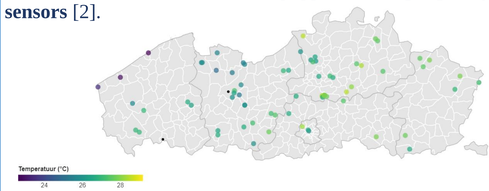


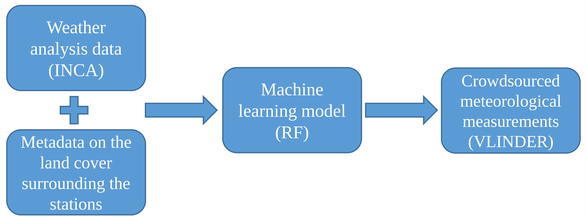
Fig. 1: A map of Flanders (the North half of Belgium) with the locations of the VLINDER stations.

Objectives

The main goals of this research are the following:

- Train ML model that predicts the difference between the INCA analysis and the VLINDER observation (residual) as a function of the local land cover.
- Evaluate the model performance on unseen data.
- Analyze the interpretability of the model (Shapley values).
- Use the model to map out the strength of the Urban Heat Island (UHI) effect during a diurnal cycle.

Methods



Features used as input for the model

From analysis data (INCA [3]):

- 2-meter temperature
- 10-meter wind speed
- Relative humidity
- Maximum temperature of the previous 24 hours

From the metadata:

- Altitude
- Land cover fractions (water, green & impervious) at a radius of 50 m, 250 m, 1 km around the station

Other:

- Diurnal cycle expressed as a cosine function with a 24 hour period

The RF model is trained with data from 06/2020-08/2020 from all the stations from the VLINDER network.

Results

The performance of the RF model was tested on data from unseen stations over 09/2020 using the root mean squared error (RMSE) metric. **Table 1 and Fig. 2** show that the RF model can correct the INCA data by capturing the characteristic land cover-related impact on the temperature. However, there is still room for improvement in the RF model, especially in capturing the nighttime UHI effect.

	Impervious dominant station	Water dominant station	Green dominant station	All stations
Total RMSE	1.1 (1.8)	1.0 (1.3)	1.0 (1.0)	1.0 (1.3)
Night RMSE	1.3 (2.4)	1.1 (1.7)	0.8 (0.8)	1.1 (1.5)

Table 1: RMSE values (in °C) over 09/2020 between the RF model prediction and the measurements and in brackets the RMSE between the INCA data and the measurements.

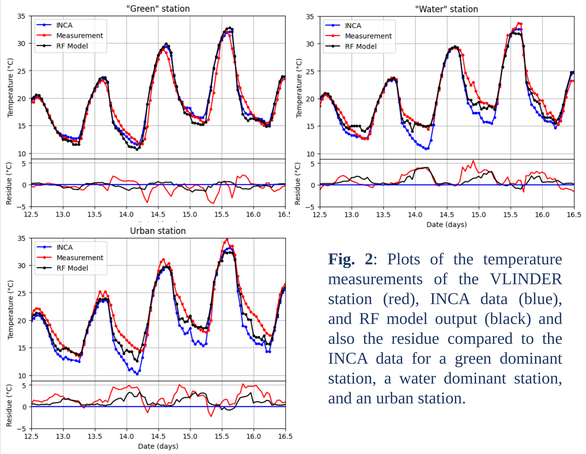


Fig. 2: Plots of the temperature measurements of the VLINDER station (red), INCA data (blue), and RF model output (black) and also the residue compared to the INCA data for a green dominant station, a water dominant station, and an urban station.

To better understand how features interact, we used a concept from game theory called **Shapley values** [4]. They represent how much the **value of a feature** tends to **impact the output of the ML model**. For the obtained Shapley values from the RF (residual) model for the training data, see **Fig. 3**. For most features, we see expected correlations (e.g., Green fraction (250 m), Impervious fraction (50 m), wind speed …). However, the relation with the model output is less clear for other features (the bottom half of the graph). This suggests that the model and its input features can still be refined further.

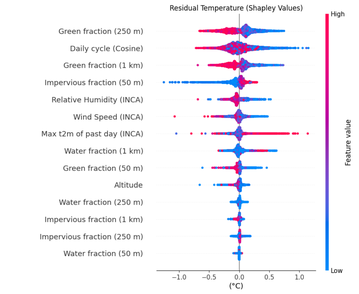


Fig. 3: Plot of the Shapley values for the training data, with large Shapley values impacting the model output more.

With the obtained RF model, the **diurnal cycle for a city** (Ghent, Belgium) containing all three land cover types was calculated, as seen in **Fig. 4**. These plots show the ability of the model to capture the temporal evolution of the UHI.

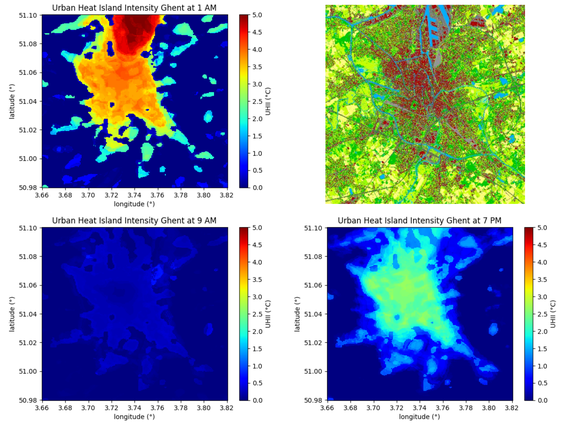


Fig. 4: The UHI intensity (difference between the temperature inside and outside the city) calculated for a clear sky and low wind day in September at 1 AM, 9 AM, and 7 PM **at 50 m x 50 m resolutio**n with, respectively, a map of the land cover of Ghent (impervious, water and green land cover, respectively as red, blue and green).

Conclusions

The random forest model shows a good performance for unseen locations and time periods. However, the current model can still be improved, especially for urban environments. The strength of the **urban heat island effect** tends to be **underestimated** by the model. The Shapley values also show that the feature choice can still be optimized. The power of the RF model is shown in its ability to calculate temperature corrections for large areas, such as **Fig. 4**, at a **low computational cost**.

Future perspectives

- Implementation / ingestion of synthetic data to get more variation in the land cover fractions in the data set.
- Evaluate the RF model against physically based models.
- More complex models
- More urban-related features (building height, sky view factor), more radiation-related features …
- Comparison with a higher density network of crowdsourced stations (e.g., Netatmo)

References

[1] Venter ZS, Brousse O, Esau I, Meier F (2020) Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. Remote Sensing of Environment, 242, 111791.

[2] Caluwaerts S, Top S, Vergauwen T, Wauters G, De Ridder K, Hamdi R, ... & Termonia P (2021) Engaging schools to explore meteorological observational gaps. Bulletin of the American Meteorological Society, 102(6), E1126-E1132.

[3] Reyniers M, Delobbe L, Kann A, Haiden T, Wittmann C, Deckmyn A (2010) The implementation of the nowcasting system INCA for Belgium: current status. ERAD 2010 conference (p 6-10).

[4] Merrick L, Taly A (2020) The explanation game: Explaining machine learning models using Shapley values. In Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4 (pp. 17-38). Springer International Publishing.

Acknowledgments

Supported by VUB - Onderzoeksraad grant OZR3893, FWO fellowship 1270723N, and BELSPO - FED-tWIN Prf-2020-017. We warmly thank Simon De Kock, Maarten Reyniers and the VLINDER team at University Ghent for their help. (VLINDER dashboard at https://vlinder.ugent.be/dashboard/)

This presentation participates in OSPP
EGU
Outstanding Student & PhD candidate Presentation contest

QR code to Abstract

Sharing is encouraged