High-resolution air temperature mapping in a data-scarce, arid area by means of low-cost mobile measurements and machine learning

High-resolution air temperature mapping in a data-scarce, arid area by means of low-cost mobile measurements and machine learning

Ahmed H M Eldesoky[1], **Nicola Colaninno**[2] **and Eugenio Morello**[2]

[1] Università IUAV di Venezia, Palazzo Badoer, San Polo 2468, 30125 Venice, Italy
[2] Laboratorio di Simulazione Urbana Fausto Curti, Dipartimento di Architettura e Studi Urbani (DAStU), Politecnico di Milano, via Bonardi 3, 20133 Milan, Italy

E-mail: ahmed.eldesoky@iuav.it, (nicola.colaninno, eugenio.morello)@polimi.it

Abstract. The availability of gridded, screen-level air temperature data at an effective spatial and temporal resolution is important for many fields such as climatology, ecology, urban planning and design. This study aims at providing such data in a data-scarce, arid city within the greater Cairo region (Egypt), namely the Sixth of October, where, to our knowledge, no such data are available. By using (i) air temperature data, collected from mobile measurements, (ii) multiple spectral indices, (iii) spatial analysis techniques and (iv) random forest regression modelling, we produced air temperature maps (for both daytime and nighttime) at 30-m spatial resolution for the entire city. The proposed method is systematic and relies on low-cost instrumentation and freely-available satellite data and hence it can be replicated in similar data-scarce, arid areas to allow for better spatial and temporal monitoring of air temperature.

1. Introduction

With heatwaves becoming more severe and frequent across many parts of the world [1], the interest in better understanding the urban micro- and local climate phenomena has been growing both in research and practice of urban planning and design. Furthermore, air temperature, measured at screen-level height (~ 1.5 m above ground), is an important variable for many fields such as climatology, ecology and hydrology [2,3]. However, monitoring air temperature in the urban canopy layer (beneath the roof level) has been always limited by the availability and spatial coverage of air temperature data from fixed weather stations [4,5]. Moreover, setting up a meteorological network of fixed weather stations can be expensive or not possible in some locations [6]. Alternatively, mobile measurements, using instruments mounted on vehicles (e.g. cars, bicycles) or carried by humans, can be used to complement observations from fixed weather stations or to observe places that are rarely explored or with spatial heterogeneity of air temperature [6] and have been used in many studies [e.g. 7–15].

Nevertheless, air temperature data obtained either from fixed weather stations or using mobile measurements are collected as point samples and cannot continuously describe the spatial variability of air temperature. Hence, providing gridded air temperature data at high spatial resolution has become of great importance and different modelling approaches have been used for this purpose such as interpolation, regression and simulation [3].

In particular, the random forest (RF) regression—a non-parametric machine learning model—is among the most recently investigated regression modelling techniques that have proven high predictive performance in many studies when using mobile measurements [8–11].

In this study, we explore the effectiveness of the RF regression in modelling air temperature in an arid area, using sample air temperature data, collected from low-cost mobile measurement campaigns (as dependent variable), and multiple spectral indices, derived from freely-available satellite imagery (as explanatory variables). The aim is to provide and make publicly available gridded air temperature data at high spatial resolution for a data-scarce area where, to our knowledge, no such data are available.

The paper is structured as follows: firstly, the study area is introduced in Section 2; then, the data and methods utilized for modelling air temperature are presented in Section 3; Section 4 presents the results of the study; and finally, the main conclusions and the future research are presented in Section 5.

2. Study area

The study area is a new desert city within the greater Cairo region (Egypt), namely the Sixth of October. The city is located approximately 32 km west of Cairo (29.9° N, 30.9° E) and covers an area of about 220 km² with a population of nearly 348,870 inhabitants as of 2017. According to the Köppen–Geiger climate classification system [16], the city is located in the hot desert climate zone (BWh).

3. Data and methods

Three main steps were required to map air temperature at high spatial resolution. Firstly, mobile air temperature data were collected and processed. Secondly, satellite data were acquired and multi-spectral, multi-scalar indices were derived. Finally, RF regression models were fitted and used to produce air temperature maps at high spatial resolution. Figure 1 shows an overview of the air temperature modelling procedure and Sections 3.1, 3.2 and 3.3 explain the aforementioned steps in more detail.

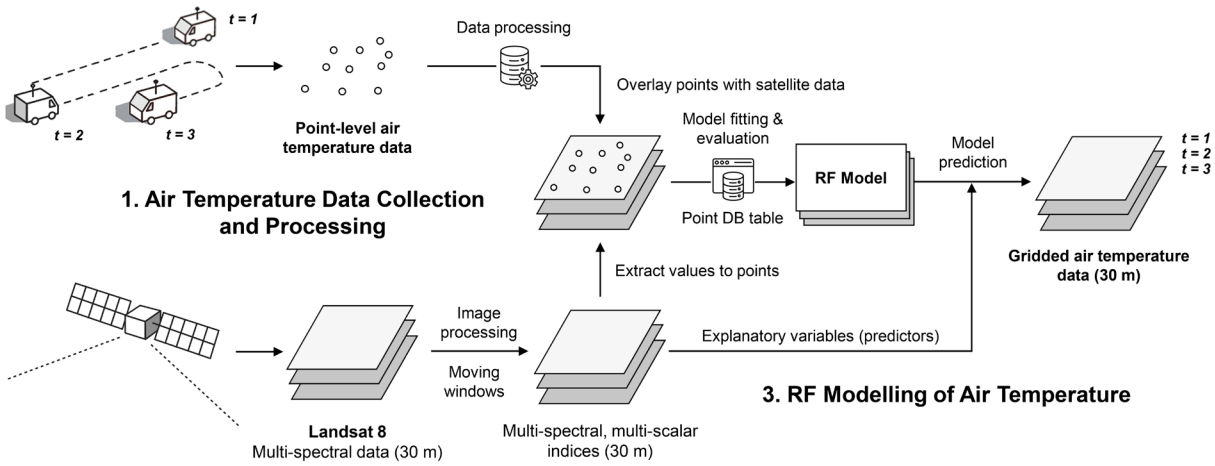

Figure 1. Schematic overview of the air temperature modelling procedure.

3.1. Air temperature data collection and processing

In this study, a total of four automobile-based measurement campaigns were carried out on September 2nd, 29th and 30th, 2020 under suitable meteorological conditions of low nebulosity and low wind speed (below 9 m/s at 10-m height). The measurement campaigns were conducted using a relatively low-cost (~ $400) portable, wireless weather station with Global Positioning System (GPS), manufactured by PASCO (PS-3209), to measure the ambient air temperature at a one-second interval along a predefined route that crosses different land uses/covers (LULCs) and morphologically different built-up areas. Each measurement campaign was completed between two and three hours to minimize any changes in the background climate conditions during the campaign time [12,14]. In particular, the temperature sensor has an accuracy of ± 0.2 °C and 0.1 °C resolution and it was mounted, at a screen-level height, on top of a car and shaded by a cardboard sheet. The car moved at an average speed of 20 to 30 km/h (minimum 15 km/h and maximum 60 km/h) which was sufficient to (1) ensure adequate ventilation for the sensor

to rapidly adjust to local temperature changes [15,17]; (2) minimize any radiation-induced errors [15]; and (3) reduce the influence of vehicular anthropogenic heat [12,14].

Further, in order to remove any confounding effects in the collected temperature data, three processing steps were applied as recommended by Oke *et al.* [6]. This included, firstly, correcting for the local temperature changes that occurred during the time of the campaign. This was done by returning to the measurement starting point and calculating an average cooling/warming rate to adjust all the temperature records to the reference start time. Secondly, removing the effect of altitude changes by applying an average lapse rate of 0.64 K per 100 m [6]. Thirdly, excluding measurements recorded when the car speed was very low/high (below 15 km/h and above 60 km/h). Table 1 shows the summary of the air temperature data for each measurement campaign after applying all the aforementioned steps.

Table 1. Summary of the processed air temperature data.

		Reference start time	Descriptive statistics (°C)			
Run no.	Date	(HH:MM)	Mean	Std.Dev.	Min.	Max.
1	2 September 2020	19:00 LT	31.36	0.45	30.35	32.60
2	29 September 2020	14:30 LT	36.01	0.50	34.59	37.92
3	30 September 2020	13:30 LT	35.34	0.62	32.88	37.68
4	30 September 2020	19:30 LT	29.35	0.38	28.49	30.53

Finally, all the processed air temperature data were exported to Geographical Information System (GIS), using the location information collected by the GPS sensor, and assigned a projection.

3.2. Calculating multi-spectral, multi-scalar indices

Modelling air temperature employing regression approaches requires using one or more variables as predictors (explanatory variables). However, there are many factors that influence air temperature (e.g. LULC, land surface temperature (LST), anthropogenic heat, solar radiation, altitude) [2,3]. In this regard, remotely-sensed data (e.g. from satellites) can provide information on many of the surface properties that influence air temperature [2]. For instance, several studies have statistically modelled air temperature based on satellite-derived LST and other spectral indices that distinguish LULC types such as the Normalized Difference Vegetation Index (NDVI) [e.g. 2,18].

Here, we used spectral indices, derived from Landsat 8 imagery, to model air temperature based on LULC characteristics. Landsat 8 carries two sensors, namely the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIR), which provide eight spectral bands at 30-m spatial resolution, one panchromatic band (15 m) and two thermal bands (collected at 100 m and resampled at 30 m) [19]. For this study, a Landsat 8 level-1 image of the study area, acquired on September 14, 2020, was freely downloaded from the United States Geological Survey (USGS) and atmospherically corrected.

In particular, three spectral indices, that distinguish LULC types in arid areas (e.g. vegetation cover, impervious surfaces and sandy desert), were derived [20,21]. These are the Soil Adjusted Vegetation Index (SAVI) [22], the Normalized Difference Built-up Index (NDBI) [23], and the Normalized Difference Sand Index (NDSI) [20].

More specifically, SAVI is a measure of vegetation density, and it is used in areas where vegetation cover is low (e.g. arid areas) to correct for the soil brightness. SAVI is defined as:

$$\mathrm{SAVI} = (((\mathrm{B}\,5 - \mathrm{B}\,4))/((\mathrm{B}\,5 + \mathrm{B}\,4 + \mathrm{L}))) \times (1 + \mathrm{L}) \quad (1)$$

where B 5 is the near-infrared (NIR) band, B 4 is the visible red band and L is a soil brightness correction factor (L = 0.5). On the other hand, NDBI is used to characterize built-up areas and bare soil which reflect more shortwave infrared (SWIR) than NIR and is defined as:

$$\mathrm{NDBI} = ((\mathrm{B}\,6 - \mathrm{B}\,5)/(\mathrm{B}\,6 + \mathrm{B}\,5)) \quad (2)$$

where B 6 is the SWIR 1 band (1.57-1.65 μm) and B 5 is the NIR band. Furthermore, to better distinguish between the sandy desert and the built-up areas or bare soil, Pan *et al.* [20] have proposed the NDSI based on Landsat 8 spectral bands 1 and 4. NDSI is defined as:

$$\mathrm{NDSI} = ((\mathrm{B}\,4 - \mathrm{B}\,1)/(\mathrm{B}\,4 + \mathrm{B}\,1)) \quad (3)$$

where B 4 is the visible red band and B1 is the coastal aerosol band (0.43-0.45 μm).

Nevertheless, air temperature can be influenced by LULC characteristics at multiple scales (varying radii) [9,11]. To account for this spatial dependence in the regression models, several studies have further applied focal operations (also called neighborhood operations) to the spectral indices using a moving window approach with different radii [e.g. 2,8–11]. This creates a new dataset of the spectral indices where the value of each pixel is a function of the values of all the neighboring pixels within a specific radius (e.g. mean, minimum, maximum, median). One approach to identify the most appropriate spatial scales (radii) is to calculate the correlation coefficient between each predictor (calculated at each potential spatial scale) and air temperature and select the scale with the highest correlation coefficient or lowest Akaike information criterion (AIC) [10,24]. Alternatively, one can use multi-spectral, multi-scalar indices, where each spectral index is calculated at multiple meaningful scales [2,8,9].

In this study, we calculated each of the three derived spectral indices at multiple spatial scales using a moving average algorithm. More specifically, we used 15 potential radii that range from 50 to 1000 m as proposed by Voelkel and Shandas [8] and Shandas *et al.* [9], and hence a total of 45 predictors were used in modelling air temperature.

3.3. RF modelling of air temperature

RF is a non-parametric machine learning algorithm that uses ensemble learning for both classification and regression tasks and has a nonlinear nature [25]. It operates by constructing n_{tree} decision trees using n_{tree} bootstrap samples of the dataset with replacement and m_{try} random subset of candidate variables (predictors) at each node. Each new data point can be predicted by running it down through each of the n_{tree} decision trees and averaging all the predicted values from all trees (in case of regression) or taking the majority of votes (in case of classification).

To fit RF regression models and predict air temperature, the processed air temperature data points from each measurement campaign (Section 3.1) were overlaid with the multi-spectral, multi-scalar indices (Section 3.2), and each point was assigned the value of the pixels that it overlays. The result is four tables, each with 46 columns (the measured air temperature and 45 predictors) and a number of rows equaling the number of observations made in each measurement campaign. The tables were used as input in the *randomForest* function [26] in R [27] to fit RF models using the default number of trees (n_{tree}) and variables (m_{try}), i.e., 500 and 15 (the total number of variables divided by three), respectively. The RF models were then evaluated using the out-of-bag (OOB) dataset, i.e., the data that were not included in the bootstrap samples (around one third), and two measures for goodness of fit were calculated, namely the coefficient of determination (R^2) and the Root Mean Square Error (RMSE). Finally, the obtained RF models were used to predict an air temperature value for each location that was not visited by the vehicle, based on the values of the multi-spectral, multi-scalar indices, and air temperature maps (at 30-m spatial resolution) were produced for the entire city.

4. Results

All four models showed high performance with R^2 more than 0.87 and RMSE below 0.18 °C (Table 2). More specifically, the nighttime models outperform the afternoon ones which is in agreement with previous studies that recommended including other predictors (e.g. building heights) for better modelling air temperature during the afternoon time [8,9].

Table 2. Summary of the RF model results.

Run no.	Date	Reference start time (HH:MM)	Model goodness of fit	
			R^2	RMSE (°C)
1	2 September 2020	19:00 LT	0.96	0.09
2	29 September 2020	14:30 LT	0.87	0.18
3	30 September 2020	13:30 LT	0.92	0.18
4	30 September 2020	19:30 LT	0.93	0.10

Figure 2 shows the resulting maps at 30-m spatial resolution, where the impact of LULC and urban morphology is apparent on the spatial variability of air temperature. For instance, the industrial areas, located in the southwest of the city, exhibit higher daytime and nighttime air temperatures than the surrounding areas due to the extensive impervious cover, low-albedo construction materials and anthropogenic heat from industrial activity. In contrast, residential areas with abundance of vegetation cover (northeast of the city) are cooler than areas with sparse or without vegetation during both daytime and nighttime. During the daytime, central urban areas with relatively higher building density are cooler than the surrounding desert (urban cool island), but warmer at night which is typical for arid cities. This can be returned to the shadows cast by tall and compact buildings, which reduce the amount of absorbed solar radiation by surfaces.

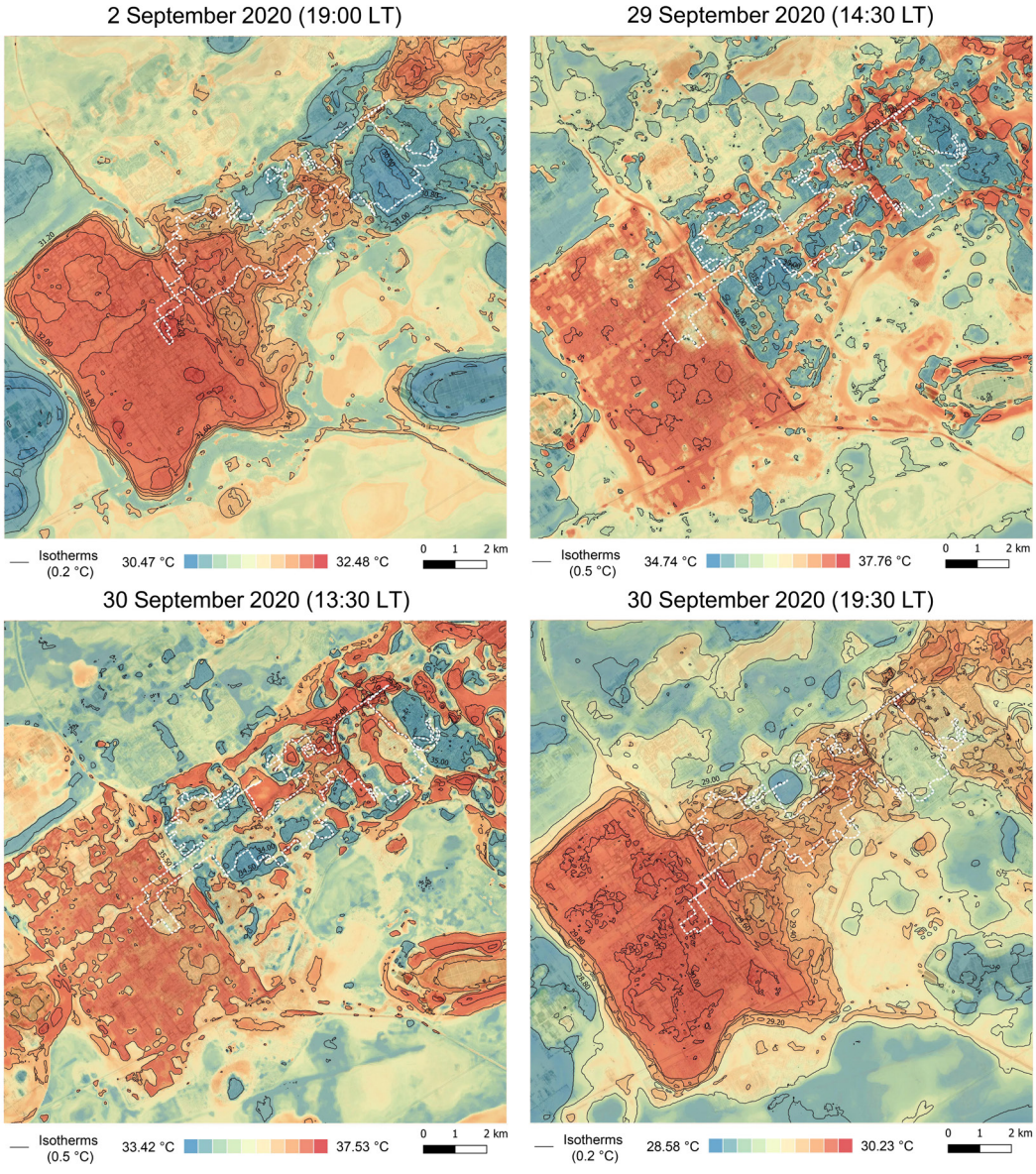

Figure 2. Spatial distribution of the modelled air temperature (30 m) with isotherm lines. The dotted white lines denote the routes of the measurement campaigns.

5. Conclusions and future research

In this study, we explored the effectiveness of the RF regression modelling in providing gridded air temperature data at high spatial resolution in a data-scarce, arid area, using air temperature data,

collected from low-cost mobile measurements and multiple spectral indices, derived from freely-available satellite imagery. The results showed a high predictive power of the RF models with R^2 more than 0.87 and RMSE below 0.18 °C. The produced air temperature maps can be useful for various applications such as better informing urban planning and design about the possible impacts of LULC and urban morphology on urban climate.

Nevertheless, this study has a number of limitations that may influence the accuracy of the produced maps and should be considered in future studies. Firstly, although the temperature sensor of the PASCO weather station is not directly exposed to sunlight and was further shaded by a cardboard sheet, some radiation-induced errors may remain, thus it is better placed in a solar radiation shield to ensure the highest data accuracy. Secondly, there is some uncertainty over using an average cooling/warming rate to correct for the local temperature changes that occurred during the campaign time, since different local areas may have different cooling/warming rates. One solution would be to calibrate the temperature data using observations from fixed weather stations along the route if they exist. Also, limiting the campaign time by using a shorter route or employing multiple vehicles can help to reduce this source of error. Lastly, although the RF approach has proven very effective in both classification and regression tasks, it is prone to overfitting. Hence, it is recommended to use an external dataset of air temperature for better evaluating the model performance rather than the OOB dataset. Cross-validation, using training and test subsets of the original dataset, can also be used for this purpose. Further measurement campaigns should be conducted over longer time periods and during different seasons for developing a dataset of air temperature at finer temporal resolution and thus allowing for better monitoring of air temperature.

6. References

[1] Perkins-Kirkpatrick S E and Lewis S C 2020 *Nat Commun*
[2] Ho H C, Knudby A, Sirovyak P, Xu Y, Hodul M and Henderson S B 2014 *Remote Sens Environ*
[3] Shahraiyni H T and Sodoudi S 2017 *Therm Sci*
[4] Hooker J, Duveiller G and Cescatti A 2018 *Sci Data*
[5] Cai M, Ren C, Xu Y, Lau K K L and Wang R 2018 *Urban Clim*
[6] Oke T R, Mills G, Christen A and Voogt J A 2017 *Urban Climates* (Cambridge: Cambridge University Press)
[7] Rajkovich N B and Larsen L 2016 *Int J Environ Res Public Health*
[8] Voelkel J and Shandas V 2017 *Climate*
[9] Shandas V, Voelkel J, Williams J and Hoffman J 2019 *Climate*
[10] Alonso L and Renard F 2020 *Remote Sens*
[11] Voelkel J, Shandas V and Haggerty B 2016 *Prev Chronic Dis*
[12] Leconte F, Bouyer J, Claverie R and Pétrissans M 2015 *Build Environ*
[13] Tsin P K, Knudby A, Krayenhoff E S, Ho H C, Brauer M and Henderson S B 2016 *Urban Clim*
[14] Shi Y, Lau K K L, Ren C and Ng E 2018 *Urban Clim*
[15] Cassano J J 2014 *Bull Am Meteorol Soc*
[16] Köppen W and Geiger R 1936 *Handbuch der Klimatologie: Das geographische System der Klimate*
[17] Unger J, Sümeghy Z and Zoboki J 2001 *Atmos Res*
[18] Cristóbal J, Ninyerola M and Pons X 2008 *J Geophys Res Atmos*
[19] USGS 2015 *Earth Resour Obs Sci Cent*
[20] Pan X, Zhu X, Yang Y, Cao C, Zhang X and Shan L 2018 *Sci Rep*
[21] Yang Y, Cao C, Pan X, Li X and Zhu X 2017 *Remote Sens*
[22] Huete A R 1988 *Remote Sens Environ*
[23] Zha Y, Gao J and Ni S 2003 *Int J Remote Sens*
[24] Bradter U, Kunin W E, Altringham J D, Thom T J and Benton T G 2013 *Methods Ecol Evol*
[25] Breiman L 2001 *Mach Learn*
[26] Liaw A and Wiener M 2002 *R News*
[27] Team R C 2019 *Vienna, Austria*