Predicting air temperature using multiple linear regression: a case study of McGill University

Priyadharshini Sakthivel (261098987)

Vaishali Mishra (261069994)

Final Project Report

submitted for

INFS 630 - Data Mining

April 8, 2023

McGill University

TABLE OF CONTENTS

Abstract

It is a well-known fact that climate change is a global concern. Studies suggest that urban areas are major contributors to climate change with their maximum greenhouse gas emissions which are the major drivers of global warming. The world population continues to rise and it is anticipated that more than half of the world population will reside in the urban areas spread across the globe. So the focus should be on climate change adaptation and in urban zones green urban infrastructure is the widely proposed adaptation technique. This project aims to predict air temperature around McGill University with the help of multiple linear regression when time and vegetation cover percentage are given.

Keywords: Multiple linear regression, climate change, air temperature, time series data

Introduction

The term "climate," derived from the Greek "klima" (inclination), describes the weather of a region averaged over a significant period of time (Heshmati, 2020). A change in either the average climate or the variability of the climate that continues for an extended period of time is referred to as "climate change." Although climate change has always occurred on Earth, the rapid rate and massive scale at which the changes are currently occurring are concerning because of its numerous effects on the ecosystem, biodiversity, and health (Heshmati, 2020; Riedy, 2016). Earth's mean surface temperature rises with an increase in heat-trapping greenhouse gases in the atmosphere, mostly carbon dioxide and water vapour. More water evaporates from the oceans and other water sources into the atmosphere as the temperature rises, which further raises the temperature (Heshmati, 2020; McMichael & Lindgren, 2011).

With over 90% certainty, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) stated that anthropogenic activity has been the main driver of the temperature increase since 1950 (McMichael & Lindgren, 2011). Even if we drastically cut GHG emissions starting right away, warming won't stop until at least the mid-2050s but may slow and reduce the degree of warming. This is because human activities that contribute to emissions cannot be stopped immediately and that putting policies in place to limit these emissions takes time (Masson-Delmotte et al., 2021). Global warming will most certainly reach 1.5°C between 2021 and 2040; it is necessary to note that we have already achieved 1.1°C in the recent decade. Limiting warming to 1.5°C -

2°C is impossible without rapid, robust, and sustained reductions in greenhouse gas emissions (Masson-Delmotte et al., 2021). This is in agreement with the analysis by English sociologist Anthony Giddens, which he refers to as the "Giddens Paradox" (Riedy, 2016). According to him, most people won't take action because there isn't a real, immediate threat from climate change. But, considering the delay between greenhouse gas emissions and their full warming impact, it will already be too late to take any action when the risk is readily obvious. Further warming will be ensured by emissions already present in the atmosphere when the threat becomes too big to ignore. The primary concern is figuring out a way out of this paradox (Riedy, 2016).

Since urban activities are significant contributors to greenhouse gas emissions, it is asserted that cities have a considerable role in climate change. According to estimates, cities account for 75% of all CO2 emissions, with transportation and construction being the two biggest sources (Bazrkar et al., 2015; UNEP, 2017). By 2050, 68% of the world's population, up from 55% today, is anticipated to reside in cities. This means that by 2050, an additional 2.5 billion people could live in urban regions due to urbanisation, with about 90% of this increase occurring in Asia and Africa (*UNDESA*, 2018). Future threats of exposure to consequences of climate change under a variety of climatic scenarios have increased due to projections of the number of people predicted to reside in urban areas (*UNDESA*, 2018). People of lower social classes and of racial and ethnic minorities are more likely to reside in warmer neighbourhoods than privileged ones, which causes a problem of environmental and climatic justice (Fan & Sengupta, 2022; Schlosberg & Collins, 2014) and drives more climate refugees into urban areas (DePaul, 2012). These figures clearly demonstrate the significance of researching urban climate dynamics.

When it comes to urban climate, the term ‘Urban Heat Island’ never goes unnoticed. Urban Heat Island (UHI) is a climatic phenomenon characterised by temperature difference between the urban and the rural zones. The urban areas experience a higher temperature relative to their rural surroundings, especially at night (Yang et al., 2020; Zhou et al., 2019). With recent developments in cities, there is no distinct borderline between ‘urban’ and ‘rural’ areas, therefore, the UHI can be considered in terms of the difference between the central parts of the city and its surrounding areas (Ngie et al., 2014). UHI occurs due to heat accumulation resulting from the physical properties of urban landscape and anthropogenic heat released into the atmosphere (Oke, 1973).

Despite the significant role of urban areas in climate change, the majority of the existing studies focus on global macro-climate and very few studies have been done with respect to urban microclimate. While climate change and urbanisation are uncontrollable, what we need to focus on is climate adaptation by strategic planning of urban spaces, like green infrastructure. This project has been designed with an objective to predict urban air temperature across McGill University, at a given time and vegetation cover percentage. Through this, we can get to know the influence of green spaces (if any) on the air temperature and will help in decision making towards the adoption of green infrastructure concept.

Dataset

There are multiple temperature sensors installed around the McGill campus by Prof. Raja Sengupta from Bieler School of Environment, Department of Geography for his research on Urban Heat Island (*Fig 1*).

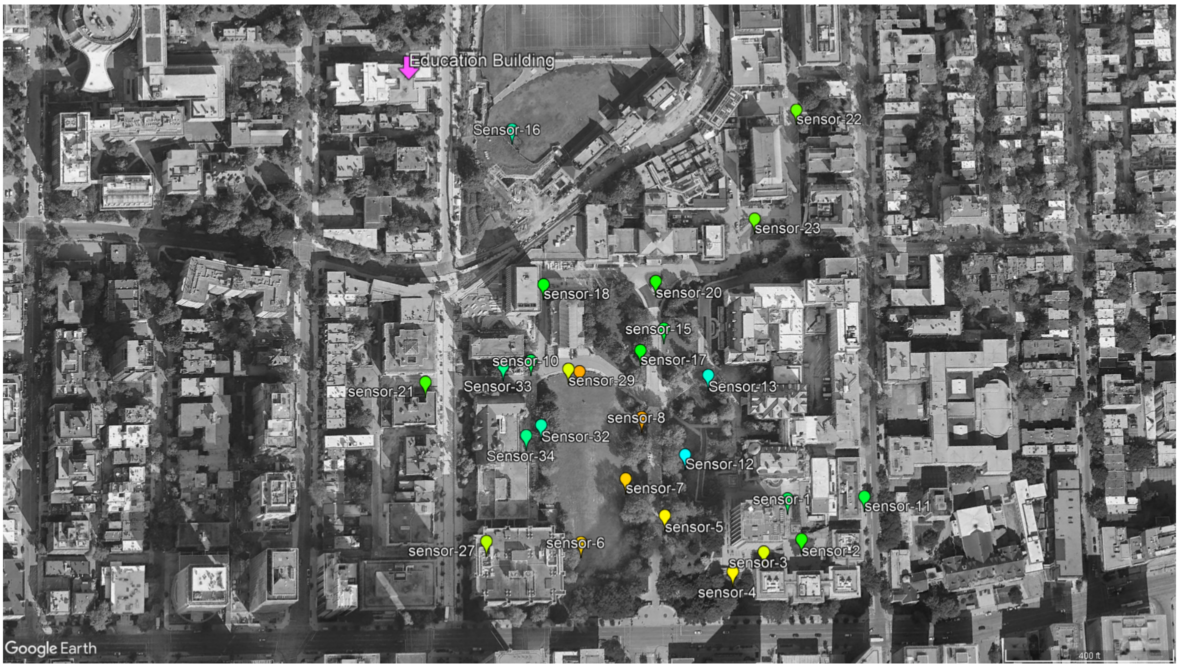


Fig 1. Distribution of temperature sensors across McGill University used for this study.

Each of these sensors collects data on temperature and humidity every 15 mins, and sends it via LoRaWAN, “a Low Power, Wide Area (LPWA) networking protocol designed to wirelessly connect battery operated ‘things’ to the internet” to an Amazon Web Service database. With GIS the distribution of impervious surfaces around each sensor as opposed to trees and grasses is estimated and recorded as a csv (comma separated values) file. The main dataset is also a csv file containing temperature measurements taken across McGill University with the help of sensors placed in chosen locations. The temperature observations are made every fifteen minutes on a daily basis from June 2022 to February 2023 with more than 400,000 records. Three variables from these datasets namely temperature, observation time and vegetation cover percentage are considered for this project which are all continuous numerical variables except vegetation cover percentage being a discrete numerical variable.

Methodology

The overall methodology of this project is illustrated as a flowchart in *Fig 2* below. The dataset, although majorly cleaned immediately after acquisition, still required some cleaning. It was followed by multiple linear regression in two phases and at last, accuracy assessment was done to estimate the performance and correctness of the models. All of these steps were done in jupyter notebooks using libraries and modules such as pandas, seaborn, matplotlib, numpy, sklearn, time and datetime. Each of these stages will be discussed in detail in the next sections.

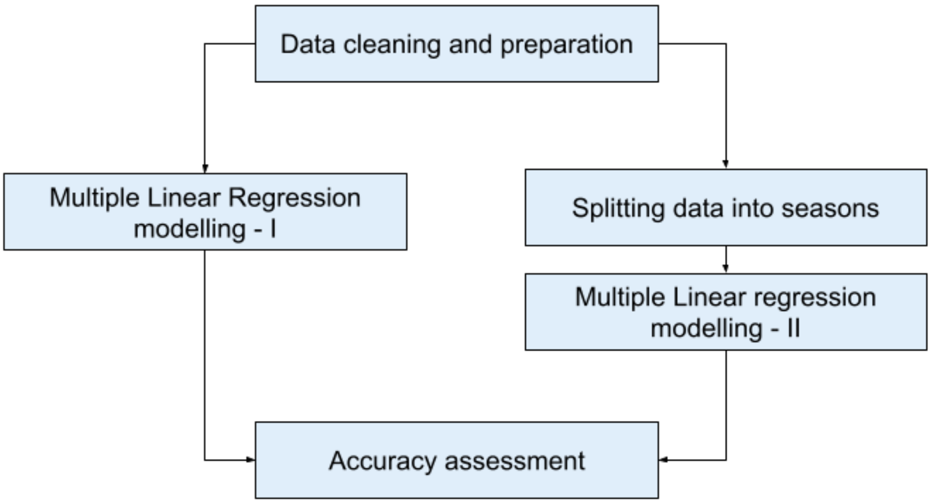

Data cleaning and preparation
Multiple Linear Regression modelling - I
Splitting data into seasons
Multiple Linear regression modelling - II
Accuracy assessment

Fig 2. Flowchart of overall methodology

Data cleaning and preparation

We had two csv files for the data. The first one had 7 columns for temperature, sensor number, readable time, humidity, battery of the sensor, etc. Out of these, we kept the columns for temperature, sensor number and readable time, while the rest of them were dropped as they were not relevant variables for our analysis. The temperature column stored the values in degrees celsius, sensor number represented the location at which that sensor was situated, and the readable time was a datetime column storing the date and time for the records. Next, we checked for outlier values for the temperature column by constructing boxplots *(Fig 3).* After observing the range of values in the column, we removed those outliers which were at a significant distance from the whiskers of the box plot or which represented impossible values which could have been stored as such due to errors in the data collection process. For instance, we removed all instances having temperature values of 125 ºC.

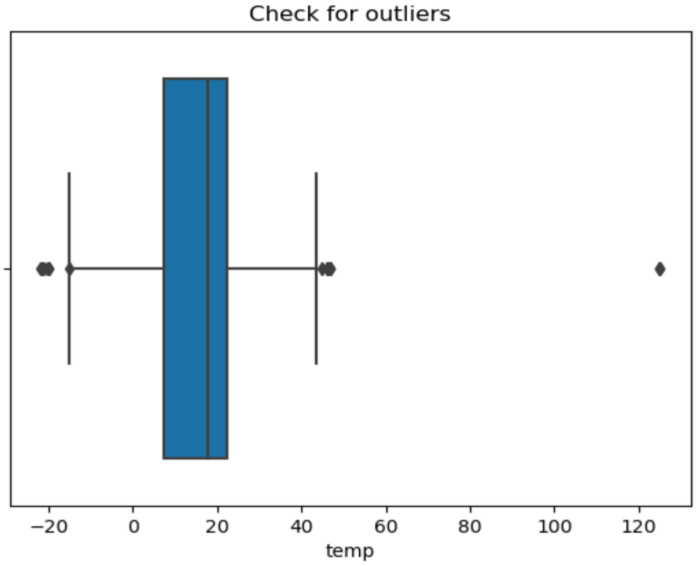


Fig 3. Boxplot to check for outliers for the temperature column

After the outlier removal, we resampled the readable_time column using the resample() function to get records at an hourly frequency rather than for every 15 minutes. Since we wanted to assess the impact that vegetation percentage had on temperature on an hourly basis, we did not further resample the values for bigger time units. The second csv file had columns for sensor number, the vegetation percentage and impermeable percentage for that sensor or location. The first and second csv files were then merged on the sensor number column using an inner join. The resulting merged csv did not have any null values. For the final step of the data cleaning and preparation process, we constructed plots to check for sensors which had an abnormal trend in terms of the temperature recorded as compared to the other sensors *(Fig 4)*. We were informed by Prof. Sengupta that these trends were a result of malfunctioning sensors. As such, they needed to be removed. Through these plots, we determined sensors 7 and 13 which had an abnormal trend. Therefore, the rows having these sensor values were dropped. The columns for impermeable percentage and sensor number were also dropped since they were no longer needed. After completing these data cleaning steps, we were left with 91,955 records.

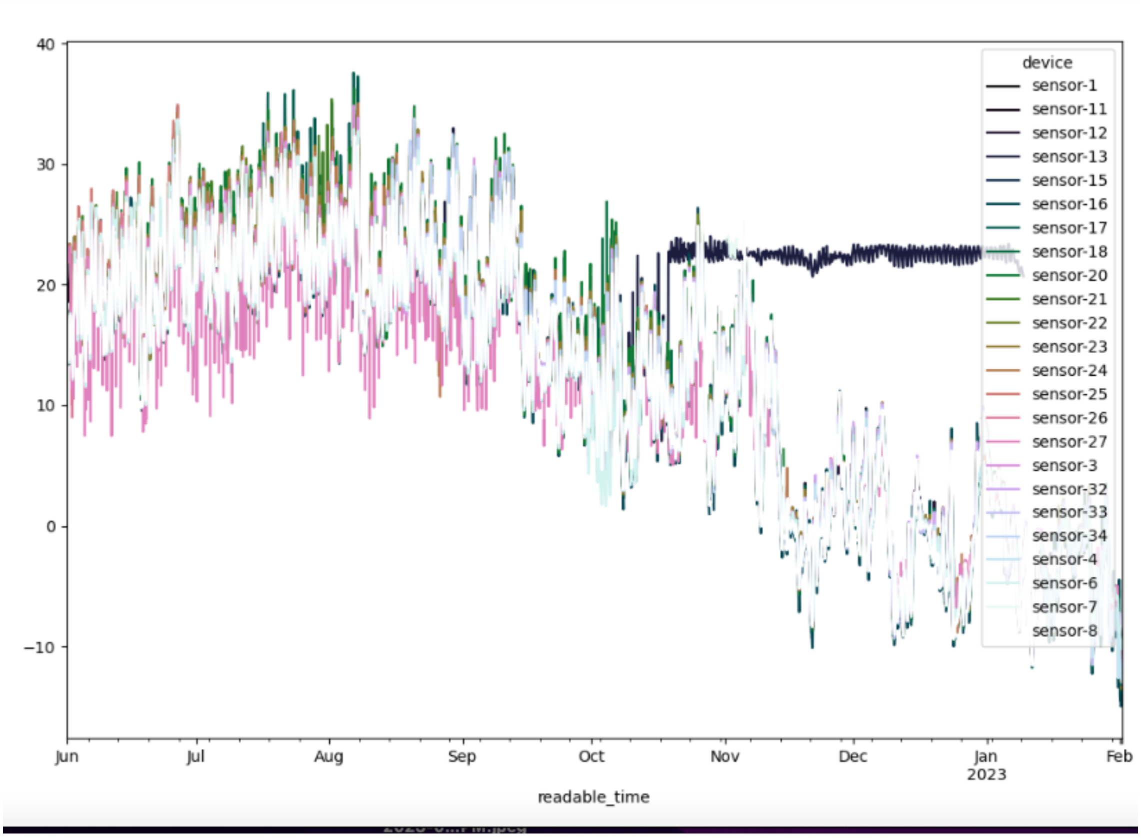
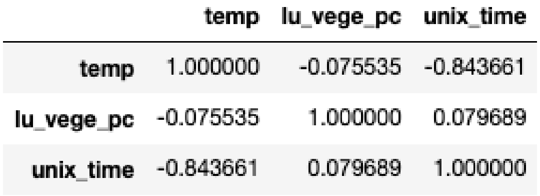


Fig 4. Plot to check for malfunctioning sensors

Assessing relationships between variables

As a pre-step in the model building phase, we analysed the relationships between our three variables of interest i.e. temperature, time and vegetation cover percentage. For this, we constructed plots for temperature vs. time, and temperature vs. vegetation cover percentage. While temperature vs time had a strong linear relationship, temperature vs vegetation percentage did not have any apparent strong relationship between them *(Fig 5)*. The time values were converted from a datetime variable to unix time values so that it could be used for multiple linear regression as the datetime values could not be used for modelling.

To understand whether multiple linear regression was a suitable model to be used for our data, we put our entire dataset through a multiple linear regression model without splitting the dataset into training and testing data. Through this process, we obtained a R^2



	temp	lu_vege_pc	unix_time
temp	1.000000	-0.075535	-0.843661
lu_vege_pc	-0.075535	1.000000	0.079689
unix_time	-0.843661	0.079689	1.000000

Fig 5. Correlation matrix to assess the linear relationships between the variables

score of 0.71. The R^2 score or the coefficient of determination measures how well the data fits the model, and a score of 0.71 showed a decently strong fit of the data to the multiple linear regression model.

Regression Modelling I

As a result of the previous assessment, we decided to construct a multiple linear regression model for the entire dataset by splitting into training and testing data in order to predict the values of temperature based on the time and vegetation cover percentage. For this, we used the TimeSeriesSplit from the sklearn library which is a kind of train-test split recommended for time series data. After it split the dataset into training and testing data, we constructed a multiple linear regression model by importing the LinearRegression model from the scikit-learn (sklearn) library. We used the R^2 and RMSE values for assessing the performance of the model. These measures and their values are discussed in the performance assessment section below.

Regression Modelling II

Since the values for these two metrics were not strong enough for the model when we used the entire dataset in Regression Modelling I previously, we decided to construct separate multiple linear regression models for the different seasons i.e. summer, fall and winter. This was done because we wanted to see the difference between the model constructed for the entire dataset from June 2022 - February 2023 as opposed to the models constructed for the seasonal slices. TimeSeriesSplit uses past values for training the model and then tests the model based on the future date values. Due to this, we expected that since the variability in the temperature values for the seasonal slices would not be as strong as in the entire dataset, the evaluation metrics would show better results and the model would be better able to capture the differences in hourly temperature. To split the entire dataset into seasonal slices, we constructed a month column by extracting the month for all the records from the readable_time column, and then selected records pertaining to a particular season based on the month values. Three separate variables were used to store the data for the three seasons. We used the same process for constructing linear regression models for summer, fall and winter seasons as is used in Regression Modelling I mentioned above.

Performance assessment

As mentioned previously, we used the R^2 score and the RMSE (root mean squared error) as metrics to evaluate the performance of the model. R^2 or coefficient of determination compares the model's predictions to the mean of the targets. Usually, the values for R^2 range from 0 to 1 but it can also be negative if the model fits the data poorly. For example, if all the model does is predict the mean of the targets, then the R^2 value would be 0, while if the model perfectly predicts a range of values, then the value would be 1. The closer the R^2 value is to 1, the better is the model. However, this metric does not tell us how wrong the model is based on how far off each prediction is from the actual values.

For that, we have the RMSE values which gives an average of the absolute differences between the predictions and actual values. Ideally, the RMSE values should be as close to 0 as possible. The greater the RMSE values, the worse is the model in terms of predicting the values. *Table 1* given below shows the R^2 and RMSE values for the entire dataset and for the seasonal slices. As can be seen, the R^2 values for summer, fall and winter datasets show a slight improvement over the R^2 value of the entire dataset which has data for all 9 months, however, they are still poor. Similarly, even though the RMSE values for the seasonal slices also are comparatively less poor as compared to the 9 months dataset, they show a significant deviation in the predicted values from the actual values in all four datasets.

	9 MONTHS		SUMMER		FALL		WINTER	
	r-squared	RMSE	r-squared	RMSE	r-squared	RMSE	r-squared	RMSE
Train	0.41	5.44	0.21	4.57	0.28	3.84	0.014	4.46
Test	-2.8	8.85	-1.29	5.91	0.26	6.07	-0.34	4.07

Table 1. R^2 and RMSE values for the entire dataset and the seasonal slices

Results and Discussion

It is found from the accuracy assessment stage that the models constructed are very poor in capturing the relationship between the dependent and independent variables of this study. One obvious reason is that climate is temporally correlated; two observations that are close to each other in time are similar in measurement. Since the dataset is smaller which temporally covers a year from June, 2022 to February, 2023 and not an entire year, the model used summer and fall data to make predictions about winter temperatures.

Another reason is vegetation did not have much significance in this case. Urban areas are among the most altered environments on Earth. From cities to suburbs, the materials composing the surface are determined by their use to humans (Bazrkar et al., 2015). However, six key factors namely the level of urban density, street orientation, street aspect ratio, neighbourhood and building typology, size, type, and location of city parks, and building and paving materials are emphasised which are typically found under the authority of city planning departments that allow cities to better manage their urban climate (Erell, 2008). So using a larger dataset across a longer time period and considering additional factors might improve accuracy. Ideally, the time series dataset should at least cover two years completely, so that the model could train on one year of data and be tested on the other year.

Conclusion

A widely proposed climate change adaptation technique in urban areas is promoting urban green infrastructure. Through this project we tried to build a model that predicts air temperature when parameters like time and vegetation cover percentage are given. Since the dataset we had access to covered a shorter period of time (nine months from June, 2022 to February, 2023), the model used summer and fall temperature records to predict winter records and hence exhibits poor performance with negative R-squared value and higher RMSE. To confirm that the model performed badly due to the aforementioned reason, we tried to make the dataset as homogeneous as possible in terms of season. On building models with seasonal datasets, the performance increased a little but not much, proving that temporal autocorrelation issue is one factor for the poor results. Another factor is that vegetation did not have a significant impact over air temperature. So considering a larger time series dataset with more parameters is likely to improve the accuracy of the model. Alternatively, this project could also be done using non-parametric classifiers like k-nn and decision trees, since they do not consider the concept of temporal autocorrelation. This can be made possible by converting the continuous dependent variable temperature into a categorical variable and using those non-parametric models.

References

Bazrkar, M. H., Zamani, N., Eslamian, S., Eslamian, A., & Dehghan, Z. (2015). Urbanization and Climate Change. In W. Leal Filho (Ed.), *Handbook of Climate Change Adaptation* (pp. 619–655). Springer. https://doi.org/10.1007/978-3-642-38670-1_90

DePaul, M. (2012). *Climate Change, Migration, and Megacities: Addressing the Dual Stresses of Mass Urbanization and Climate Vulnerability*. http://diplomatonline.com/mag/pdf/DePaul_Climate_Migration_and_Megacities.pdf

Erell, E. (2008). The Application of Urban Climate Research in the Design of Cities. *Advances in Building Energy Research*, *2*, 95–121. https://doi.org/10.3763/aber.2008.0204

Fan, J. Y., & Sengupta, R. (2022). Montreal's environmental justice problem with respect to the urban heat island phenomenon. *The Canadian Geographer / Le Géographe Canadien*, *66*(2), 307–321. https://doi.org/10.1111/cag.12690

Heshmati, H. (2020). *Impact of Climate Change on Life*. https://doi.org/10.5772/intechopen.94538

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., & Zhou, B. (Eds.). (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/9781009157896

McMichael, A. J., & Lindgren, E. (2011). Climate change: Present and future risks to health, and necessary responses. *Journal of Internal Medicine*, *270*(5), 401–413. https://doi.org/10.1111/j.1365-2796.2011.02415.x

Ngie, A., Abutaleb, K., Ahmed, F., Darwish, A., & Ahmed, M. (2014). Assessment of urban heat island using satellite remotely sensed imagery: A review. *South African Geographical Journal*, *96*(2), 198–214. https://doi.org/10.1080/03736245.2014.924864

Oke, T. R. (1973). City size and the urban heat island. *Atmospheric Environment (1967)*, *7*(8), 769–779. https://doi.org/10.1016/0004-6981(73)90140-6

Riedy, C. (2016). *Climate Change*. https://www.researchgate.net/publication/311301385_Climate_Change

Schlosberg, D., & Collins, L. B. (2014). From environmental to climate justice: Climate change and the discourse of environmental justice. *WIREs Climate Change*, *5*(3), 359–374. https://doi.org/10.1002/wcc.275

UNDESA. (2018). https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

UNEP. (2017, September 26). *Cities and climate change*. UNEP - UN Environment Programme. http://www.unep.org/explore-topics/resource-efficiency/what-we-do/cities/cities-and-climate-change

Yang, X., Peng, L. L. H., Jiang, Z., Chen, Y., Yao, L., He, Y., & Xu, T. (2020). Impact of urban heat island on energy demand in buildings: Local climate zones in Nanjing. *Applied Energy*, *260*, 114279. https://doi.org/10.1016/j.apenergy.2019.114279

Zhou, D., Xiao, J., Bonafoni, S., Berger, C., Deilami, K., Zhou, Y., Frolking, S., Yao, R., Qiao, Z., & Sobrino, J. A. (2019). Satellite Remote Sensing of Surface Urban Heat Islands: Progress, Challenges, and Perspectives. *Remote Sensing*, *11*(1), Article 1. https://doi.org/10.3390/rs11010048