Urban heat island: a statistical perspective using machine learning for Florianópolis, Brazil.

Rodrigo Nehara Moreira [a]

[a] Environmental Science And Technology, Federal University of ABC, São Paulo, Brazil, neharaaaa4@gmail.com.

Abstract. Surface Urban Heat Island (SUHI) is a phenomenon of urban areas being significantly warmer than their surrounding rural areas due to the built environment and human activities. Local Climate Zones (LCZs) are a recent classification scheme that partitions urban areas into zones based on their physical and functional characteristics, providing more nuanced information about urban environments than traditional land use categories. Machine learning (ML) and remote sensing (RM) techniques have been applied to analyze and model SUHI across LCZs, leveraging large volumes of spatial data such as remote sensing imagery and weather station observations. ML models have shown promising results in predicting SUHI intensities, identifying key LCZ features associated with SUHI, and estimating important variables for a better understanding of the climatic conditions in a given region. These findings can inform urban planning and design strategies aimed at mitigating SUHI effects and promoting sustainable urban environments.

Keywords. Surface urban heat island, machine learning, local climate zones, Landsat 8.

1. Introduction

A surface urban heat island (SUHI) is a phenomenon in which urban areas experience higher temperatures than surrounding rural areas due to the built environment and human activities. The heat island effect occurs as urban areas become more developed, with increased energy consumption, the use of heat-absorbing materials such as asphalt and concrete in buildings and pavement, and reduced vegetation. As a result, urban areas can experience surface temperatures several degrees higher than nearby rural areas. This temperature difference can create a microclimate within the city, with increased heat and reduced air quality.

Additionally, due to the lack of local information, SUHI assessment traditionally used to be retrieved by considering the homogeneity of urban fabrics regardless of the actual diversity of their occurrence. Thus, the definition of representative reference samples to conduct SUHI diagnoses was for a long time uncertain. To tackle previously existent gaps and to broaden the exchange of experiences regarding this type of analysis, the Local Climate Zones (LCZ) framework was developed [1]. It consists of a standardized system for the classification of landscape typologies based on surface structure, function, land cover, urban metabolism, etc. This scheme enables a more detailed understanding of the urban fabric and facilitates de comprehension of microclimate responses, as it allows correlational studies - e.g., the effect of vegetation and built environment geometry over temperatures [2; 3].

Recent advancements in machine learning (ML) have allowed researchers to better analyze and model the urban heat island effect [4]. ML algorithms can handle large datasets and identify complex patterns and relationships that are difficult to observe using traditional statistical methods. ML models can also be used to predict SUHI intensity in different LCZs and assess the effectiveness of different mitigation strategies.

This paper aims to analyse the SUHI effect in each LCZs using ML in five cities in Santa Catarina, Brazil. We will discuss the SUHI behavior of each LCZ, and the use of ML models in studying the SUHI effect. Finally, we will conclude with a discussion on the future directions of research in this field.

2. Material and Methods

2.1 Urban Heat Island

The SUHI was developed using pairs of images from the Landsat 8 satellite (Collection 2, Level 1).

The selection criteria included proximity of acquisition times and non-occurrence of precipitation between them. These assumptions allowed the retrieval of land surface temperature for nighttime scenes by considering that land surface emissivity doesn’t change between the two images [5].

The cloud cover was also considered in the selection, reason why only a single pair of daytime/nighttime scenes fitted the selection criteria. The meteorological data used in the study was downloaded from the Brazilian National Institute of Meteorology database (https://bdmep.inmet.gov.br/) for the weather station A806, located at coordinates 27°36’00” S and 48°37’12” W.

2.2 Local Climate Zones

Local Climate Zones (LCZs) is a classification system that categorizes urban and suburban areas based on their surface cover characteristics and the urban climate that they generate and is widely used in urban planning and climate research of SUHI.

LCZs consist of 17 categories, each with distinct surface cover characteristics and urban climate properties. These categories range from dense urban core areas, such as high-rise commercial and residential districts, to low-density residential areas, industrial areas, and natural areas such as parks and open spaces (fig 1).

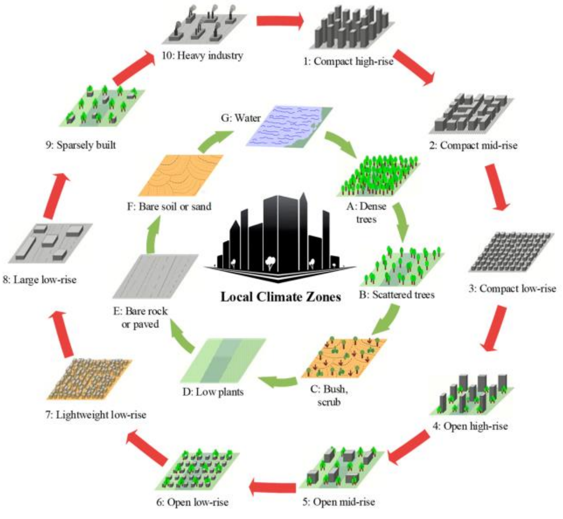

Fig. 1 - Description of LCZ types, containing 10 urban types on the outer circle and 7 natural types on the inner circle [6].

Each LCZ category has a unique set of surface cover characteristics, such as the percentage of impervious surfaces like pavement and buildings, the amount of vegetation, and the level of human activity. These characteristics influence the local climate of the area, affecting factors such as temperature, humidity, and wind patterns.

2.3 Statistical Aproaches

The statistical analyses in this project followed the theoretical perspective based on King and Roberts [7], where instead of ranking which statistical approach is appropriate, among parametric, robust, non-parametric, data mining, it will be assumed that they can be interpreted as complementary modeling perspectives on the same object of study. Thus, if different methods of approach show the same statistical pattern, there is greater evidence that these patterns are valid. Otherwise, if the results show discrepant results, it is a sign that there is still room for further in-depth analysis.

In R environment, the data from the Local Climate Zones categories, urban heat island (daytime and nighttime) and all indices and elevation deviated variables were converted into a data frame.

2.4 Univariate analysis

In the R environment, a Box-Plot with statistical details were generated for the SUHI variable, to analyze the values for each LCZ (daytime and nighttime, pairwise) and find out if the values are significantly different

For this, the parametric paired Student test, the non-parametric paired Wilcoxon test and the paired robust Yuen were performed.

After these tests, the parametric ANOVA test, the non-parametric Kruskal-Wallis [8] test and the robust version of ANOVA by Medians [9] were performed for all LCZs for daytime and nighttime separately, followed by the pairwise post-hoc tests (Games-Howell adjustment, Dunn adjustment and Yuen's trimmed means adjustment with Bonferroni correction method, respectively) to distinguish which LCZs are significantly different from the others for all tests. For the ANOVA test, Levene's Test (1960) were performed to check the variance in the groups, to decide between the Welch and Fisher type [10].

2.5 Bivariate analysis

Also following the theoretical perspective of King and Roberts, the Pearson correlation test, Spearman correlation test and Winsorized Pearson correlation test were performed for, daytime and nighttime, onsidering other variables such as the normalized difference vegetation index (NDVI), elevation, slope, aspect and emissivity.

The NDVI and emissivity were calculated using the Landsat 8 with 30-meter resolution image of the study area, while the elevation, slope and aspect will be obtained from the Copernicus DEM, also with 30-meter resolution.

2.6 Multivariate analysis

A multivariate linear regression method were performed to understand the relationship between the calculated indices and SUHI for daytime and nighttime and the other informations as independent variables.

The regression methods were calculated by the mean, median, robust and quantile (0.1 to 0.9, by 0.1) estimators. The last one interacting daytime and nighttime for all LCZs. The first three regression methods results were compared by the Akaike information criterion [11] and both parametric tests were by R^2.

3. Results and Discussions

The hypotesis tests results for all LCZs showed that all three tests had p-values less than 0.05, indicating that the different between the groups is statistically significant and the results is not due to chance or random variation, while the effect size statistic showed this difference is large. After performing a Levene´s test, the p-value lower than 0.05 determined that there is unequal variance across the groups. As a result, the Welch ANOVA test was performed.

The same statistical pattern was obtained for the paired student's t-test, the paired Wilcoxon test, and the paired Yuen robust test (fig. 2).

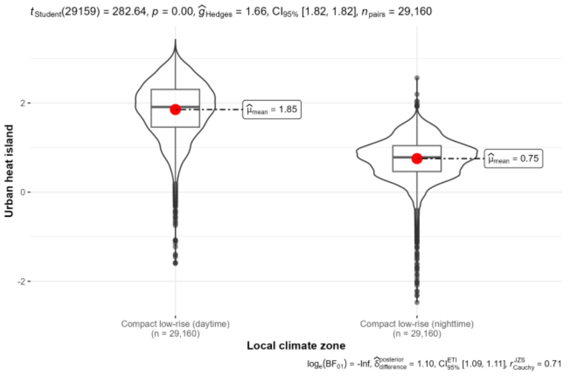

Fig. 2 – Student's t-test for Compact low-rise type with statistical details, for daytime and nighttime.

Correlation tests indicated a weak or moderate negative relationship of daytime with all variables for Pearson and Spearman tests. For the Winsorized test, strong relationship were identified with NDVI and elevation (fig. 3)

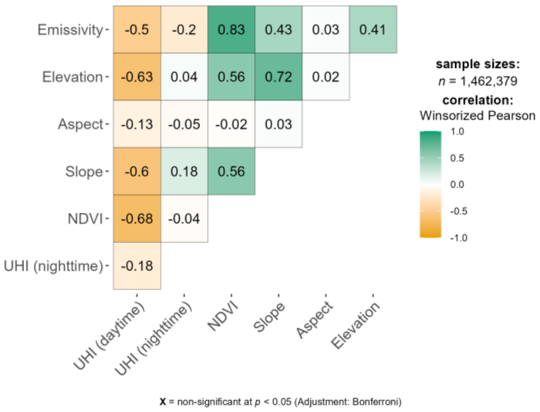

Fig. 3 – Winsorized Pearson correlation test for all considered variables.

For nighttime, only weak correlations were observed for all variables except NDVI for the Pearson test, which had a correlation of -0.34. However, for the Spearman test there was a null correlation (fig. 4).

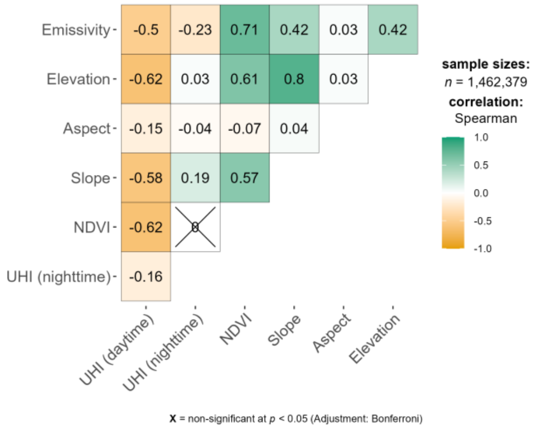

Fig. 4 – Pearson correlation test for all considered variables.

For the post-hocs, it has been found that there are significant differences between several groups for all methods. This suggests that there are statistically significant variations in the data that can't be attributed to chance alone.

Among 105 pairs compared, 82 have significant differences for the Welch ANOVA test, 78 for the Kruskal Wallis test, and 86 for the robust during daytime. For nighttime, it was 102, 92 and 104, respectively. With this it can be concluded that during the nighttime the temperature difference between the LCZs is greater than during the daytime.

For the linear regression using the mean estimator, the model yielded an R^2 of 0.71 and an AIC value of 2341764 for the daytime, while for nighttime the model yielded an R^2 of 0.44 and an AIC value of 3164334. This indicates that the mean estimator provided a good fit for daytime and explained a significant portion of the variance in the dependent variable, but such efficiency did not replicate for nighttime.

For the linear regression using the median estimator, the model yielded an AIC value of 2388706 for the daytime, while for nighttime the model yielded an AIC value of 3316945. This indicates that the median estimator provided a slightly worse fit compared to the mean estimator for daytime and nighttime.

For linear regression model using the robust estimator. The model yielded an AIC value of 2344747. This indicates that the robust estimator provided a good fit for the daytime, while for nighttime the model yielded an AIC value of 3166978. This indicates that the robust estimator provided a worse fit for both periods (tab. 1).

Tab.1 – R^2 and AIC values for each regression method.

Estimator	R^2 (daytime/ nighttime)	AIC (daytime/ nighttime)
Mean	**0.7128/0.4424**	**2341764/3164334**
Median	-	2388706/3316945
Robust	-	2344747/3166978

Quantile regression was able to promote a better understanding of the relationship between SUHI and the other variables, for all LCZs during the daytime and nighttime. Specifically, we found that the relationship between both SUHIs and the dependent variables was different from the mean for several quantiles. This suggests that the two dependent variables respond differently to changes in our independent variables, sometimes even possessing opposite behaviors (Fig. 5).

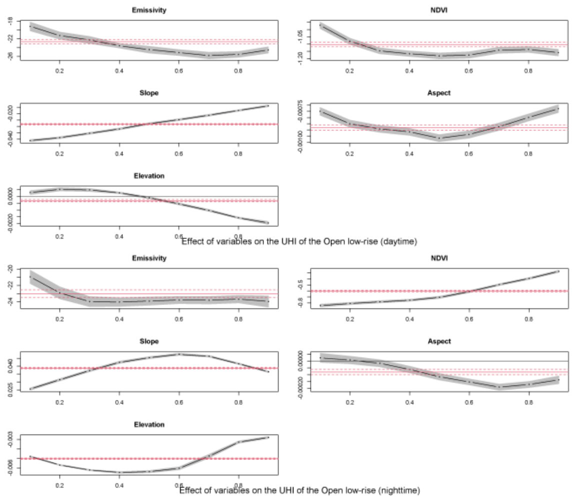

Fig. 5 – The effect of each variable on SUHI in the Open low-rise type, for daytime and nighttime.

4. Conclusion

The ML methods applied in this study was successful in demonstrating the behavior of SUHI in the different LCZs.

Our results showed that there was a significant difference between the SUHI during daytime and nighttime, indicating that period has a clear effect on the SUHI. Furthermore, by using quantile regression, we were able to better understand the behavior of all variables on the SUHI, just as regression by mean best represented daytime and nighttime.

In the parametric regressions, the model described the SUHI effects well for daytime (R^2 = 0.7128), but did not prove satisfactory for nighttime (R^2 =0.4424).

This allowed us to identify the key factors contributing to the SUHI and to gain insight into how these factors interact to create the SUHI phenomenon. Overall, our findings have important implications for urban planning and climate change mitigation efforts, and we hope that they will contribute to a better understanding of the complex relationship between urbanization and the SUHI.

5. Acknowledgment

I would like to express my gratitude to my supervisors, Jan Komárek and Tomas Klouček for their support and guidance throughout this research project. Their insightful comments, constructive feedback, and endless encouragement have been invaluable in shaping our work and pushing me to achieve my goals.

I would also like to extend my appreciation to my colleagues and friends, Bruno Rech and Tiago Mello, for their intellectual contributions, constructive criticism, and moments of fun during the development of our work.

6. References

[1] Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. Bulletin of the American Meteorological Society, 93(12), 1879-1900.

[2] Oliveira Borges, V., Carlos Nacimento , G. ., Celuppi , M. C. ., Lúcio , P. S. ., Tejas , G. T. ., & Gobo, J. P. A. (2022). Zonas climáticas locais e as ilhas de calor urbanas: uma revisão sistemática. Revista Brasileira De Climatologia, 31(18), 98–127. https://doi.org/10.55761/abclima.v31i18.15755

[3] Kaloustian, N., & Bechtel, B. (2016). Local climatic zoning and urban heat island in Beirut. Procedia Engineering, 169, 216-223. DOI: https://doi.org/10.1016/j.proeng.2016.10.026

[4] Salamanca, F., Tewari, M., & Martilli, A. (2017). Machine learning techniques for urban heat island analysis: New prospects in urban climatology. Frontiers in Environmental Science, 5, 66.

[5] Sekertekin, A., & Bonafoni, S. (2020). Sensitivity Analysis and Validation of Daytime and Nighttime Land Surface Temperature Retrievals from Landsat 8 Using Different Algorithms and Emissivity Models. Remote Sensing, 12(17), 2776. DOI: https://doi.org/10.3390/rs12172776

[6] Ma, Lei (2021). Advances of Local Climate Zone Mapping and Its Practice Using Object-Based Image Analysis. Atmosphere, 12(9), 1146. https://doi.org/10.3390/atmos12091146

[7] King, G., & Roberts, M. E. (2014). How robust standard errors expose methodological problems they do not fix, and what to do about it. Political Analysis, 23(2), 159-179.

[8] Mair P, Wilcox R (2020). Robust Statistical Methods in R Using the WRS2 Package. Behavior Research Methods, 52, 464–48.)

[9] KRUSKAL, William H.; WALLIS, W. Allen. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association, v. 47, n. 260, p. 583-621, 1952.

[10] Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. Statistical Science, 24(3), 343–360. doi:10.1214/09-sts301

[11] AKAIKE, H. Information theory and the maximum likelihood principle in 2nd International Symposium on Information Theory (B.N. Petrov and F. Cs ä ki, eds.). Akademiai Ki à do, Budapest. 1973.