Project Report:
Milan Thermal Comfort Analysis

Geospatial Analysis and Representation for Data Science.

Davide Vandelli
davide.vandelli@studenti.unitn.it

Abstract - **This report aims at analyzing the effects associated with different urban settings in the densely populated city of Milan, using several clustering techniques (k-means, Decision trees and Random Forests, DBSCAN) and statistical descriptive analysis. The scope of the latter analysis is to focus on the impact of different kinds of land use on PET, utilizing existing literature on the topic and Transform Transport Foundation's indications to process and analyze data.**

Key words - DBSCAN, k-means, PCA, Python, Physiological Equivalent Temperature, QGIS, Spatial regression.

Data origin and processing

As global temperatures keep increasing, many humans that occupy urban landscapes heavily depend on the livability of cities. Extreme temperatures are capable of causing an increase in mortality if not tackled or prevented, and are likely to decrease the quality of life for densely populated areas, such as the city of Milan.

The data for this report was obtained from Transform Transport (TT from now on), a non-profit research organization.

It is partially derived from Zuretti, Pedrazzoli, Ceccarelli, and De La Hoz's work, as an article published in 2023. The research focused on computing and analyzing the Physiological Equivalent Temperature index (PET) using Urban Multi-Scale Environmental Predictor (UMEP), focusing on the comfort of pedestrians in the streets, gathering features on spatial (e.g. buildings, walls, tree canopy heights), meteorological (e.g. temperature), environmental (e.g. emissivity of building walls and roofs) and human settings (e.g. mean radiant temperature, clothing levels) to compute the PET index[1]. The data was made available by request.

The area of focus is a grid with a 5x5m resolution, representing a subset of the center of Milan and does not cover the whole city area, ranging from being close to the main train station to its Duomo.

The data available requires several processes to be clustered and analyzed, as summarized in the next section.

PET data is expressed in the form of 160.000 text files, each representing one cell (5x5m) of Milan, containing a small dataset of 24 rows (representing the hours) with 35 features, which need to be aggregated together into one dataframe. Secondly, a grid shapefile is provided by TT foundation, which simply needs to be populated with PET data. Since the text files had 24 rows each to represent hourly changes, the solution that was found is to add 24 features to each cell in the grid, each representing the PET at a specific hour; the other features were left for clustering and descriptive analysis. This solution took a few attempts as no better interpretation of TT's indications was found.

The pre-analysis processing of the data ends with selecting streets, parks, jaywalks and pedestrian routes - in other words, any cell that was not overlapping with a building, which is another shapefile provided by TT. This was done by utilizing QGIS's *intersection* feature, and later removing the cells that were classified as mostly overlapping with a building. Due to the size of the cells, there was no hard threshold to select them, as 5m^2 are small enough for a fine analysis of an urban area, but not fine enough to be separating exactly where a building ends and where the street starts. There are some cells that are partly overlapping with buildings, but upon inspection always less than a half of the whole area.

Ultimately, we may end up with 67,053 cells (335,265 m^2, less than 0.2% of the whole metropolitan area of Milan).

Analysis

The thermal perception and physiological stress are two factors that are often utilized to compare different settings simply using the PET index and furthermore make it closer to human perception. The thresholds for categorizing PET are reported in (I).

[1] For a more detailed description of the 30+ features, please refer to Zuretti *et al.* (2023).

TABLE I

DATA SIMPLIFICATION USING PET. ADAPTED FROM MATZARAKIS (2008)

Threshold	Grade of physiological stress
PET ≤ 8.0	betw. Extreme cold stress and Moderate cold stress
8.1 ≤ PET ≤ 35.0	betw. Slight cold stress and slight heat stress
35.1 ≤ PET	betw. Moderate heat stress and strong heat stress

The original categorization had multiple grades of stress, but in order to simplify they were actually resumed to three main categories.

The data was representative of the summer in Milan, hence there is not one single hour with PET ≤ 8.0. If the thresholding is done as shown above, we should see that the distribution of moderate heat stress or above takes place from 8 a.m. to 18 p.m., peaking at 14 p.m..

Onto the clustering process. Several methods and versions of clustering were conducted to obtain results that, as solidly as possible, are not determined arbitrarily but with respect to the structures of similarity in the data.

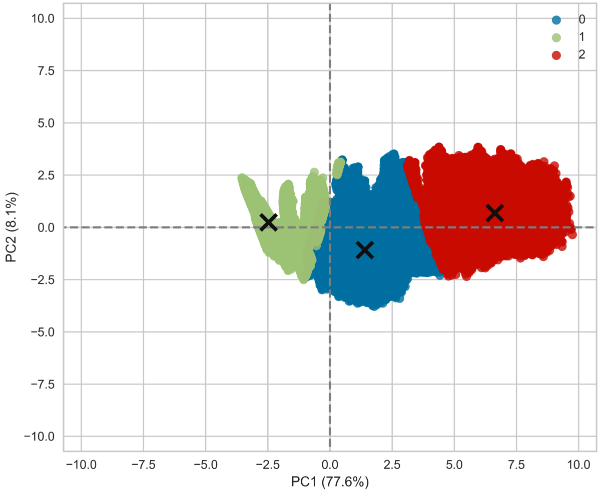
The first and last method used is K-means: first as an exploratory approach and second as a memory efficient method of reproducing other results.

Any clustering was produced with euclidean distance, and if later visualized with PCA (in its first 2 or 3 dimensions, depending on the importance of the clustering results).

For most clustering methods the feature space needs to be either standardized or normalized; in this case the "Robust Scaler" from sci-kit learn was applied.

FIGURE 2

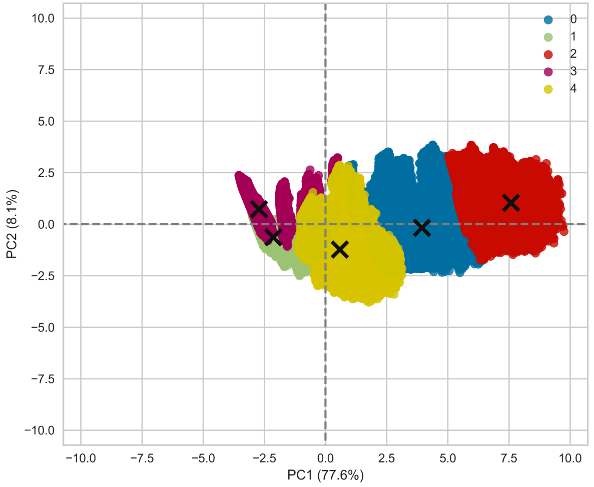
PCA (PC1 AND PC2) PROJECTION OF POINTS FOR K-MEANS, K = 3



In (2), the first K-means is visualized using PCA; there is not a linearly separated cluster, and all the points seem to be placed on a continuum - which also seems representative of their real-life spatial counterparts, where there are no hard thresholds between one 5x5 cell and another. Below, the 5-means version.

FIGURE 3

PCA (PC1 AND PC2) PROJECTION OF POINTS FOR K-MEANS, K = 5

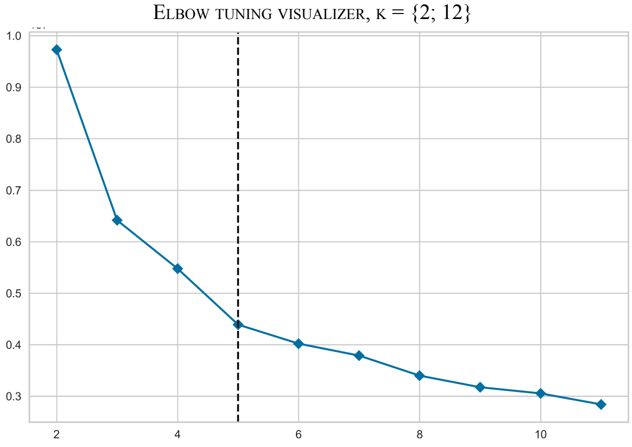


In this version, the clusters seem even cloudier, with some almost completely overlapping others. It must be noted that PCA's visualization is more an artifact than anything else, as its purpose is not to make the structures of the data more interpretable.

After trying the most common K values (3 and 5) in the literature, we move onto fine-tuning K using the elbow method.

FIGURE 5

ELBOW TUNING VISUALIZER, K = {2; 12}



The elbow is found automatically at 5, using the metric Distortion, which is the sum of squared distances from each point to its assigned center[2]. Another measure was implemented, named the Calinski-Harabasz metric, which is

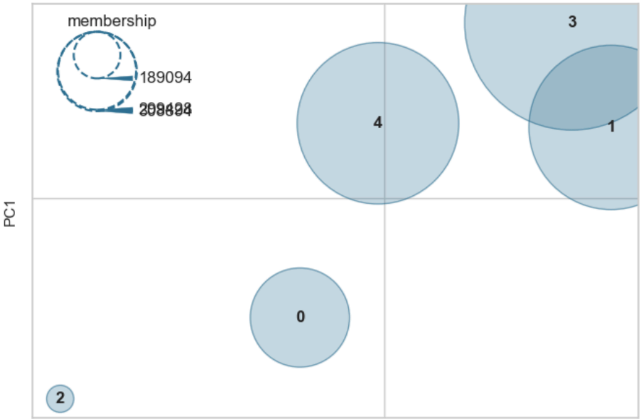
[2] For more information, find the formula in the Appendix of this paper.

the ratio of the sum of between-cluster dispersion and of within-cluster dispersion, but since the clusters are not linearly separable it was not possible to interpret it (as it also suggested optimal K=2).

In (6), another visualization technique is used, t-SNE, often accredited for its flexibility with visualizing inter-cluster distance.

FIGURE 6

T-SNE VISUALIZER FOR K=5

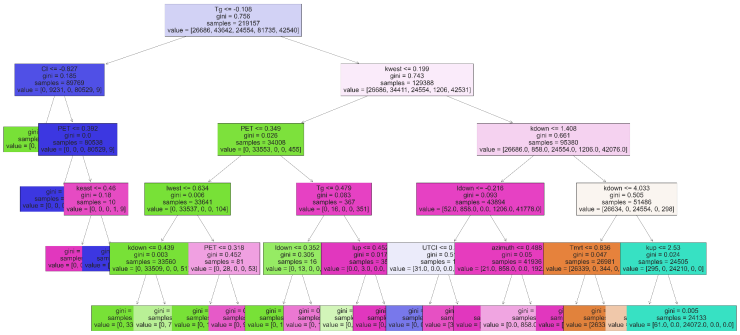


As shown above, clusters 1 and 3 are overlapped and sort of hard to separate even with t-SNE, although PCA is able to place cluster 4 away from the others, while PCA cannot, while clusters 0 and 2 are more separable. The results between visualization are comparable to a lesser extent.

To interpret the clustering results, and get a deeper understanding of what distinguishes these clusters from each other, it may be useful to implement decision trees.

FIGURE 7

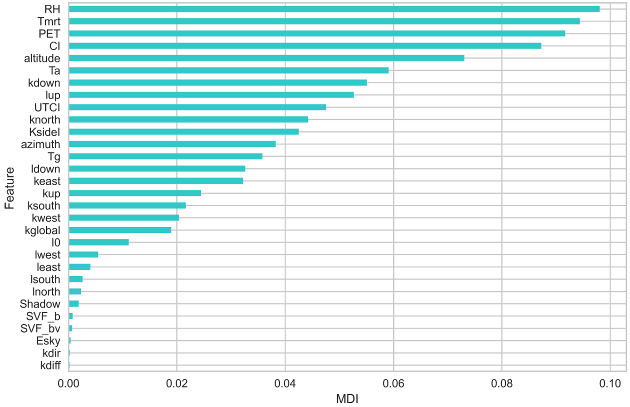
DECISION TREE, PREDICTING RESULTS FROM K=5



A higher resolution of this tree can be found in the Git of this project (FluveV/Milan_Thermal_Comfort_Analysis). The tree shown (7) has an accuracy of 99.03%, on unseen data separated from the training set (75/15). This tree has the issue that it might be overfitting. To obtain a better understanding of the relationship between variables and cluster assignment, a Random Forest is implemented.

FIGURE 8

MEAN DECREASE IN IMPURITY: FEATURE IMPORTANCE



By cross validating Sci-kit's Random Forest on ⅛ of the data (for memory limits purposes), we may obtain a mean score of 0.998 on 515,664 examples. We can also find that some features are effectively better at discerning the data than others. The features used more for fitting trees were: **'RH'**, **'Tmrt'**, **'PET'**, **'CI'**, **'altitude'**, **'Ta'**, **'kdown'**, **'lup'** (others had importance less than 5%)[3].

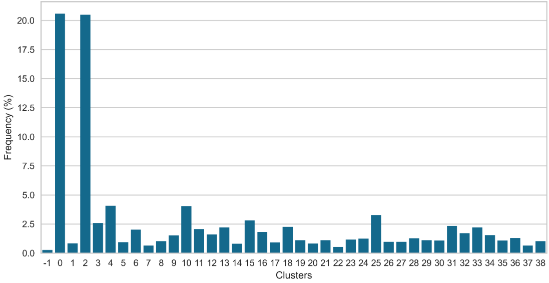
Ultimately, we might not assume a priori the k at all: using DBSCAN, short for Density-Based Spatial Clustering of Applications with Noise, only requires two hyperparameters: MinPts, which is the number of nearest samples for a point to be considered a core point, and ε, which is the maximum distance between two samples for one to be considered in the neighborhood of the other. This clustering application is particularly useful when lacking a domain expert for specific kinds of data; it marks some observations as outliers when they lie alone in low-density regions, while marking observations to the same cluster when packed in high-density regions.

Tuning DBSCAN requires several different approaches; in this case, what worked optimally was to set MinPts as the number of dimensions or features used plus one, while ε is set to twice MinPts - 2; the chosen distance metric is still euclidean. The features in this case were not normalized as it may be particularly difficult to tune hyperparameters if the feature space changes its scale. Due to the worst case scenario of this technique's performance, O(n^2) (Schubert *et al.*, 2017), the data had to be limited down to a sample size that is 1/100 of the original sample size.

[3] Extended names of the features: outgoing longwave radiation (lup), air temperature from meteorological data (Ta), mean radiant temperature (Trmt), incoming shortwave radiation (kdown), universal thermal comfort index (UTCI), altitude, relative humidity from meteorological data (RH), clearness index (CI).

FIGURE 9

DBSCAN'S CLUSTER DISTRIBUTION. N=15,469

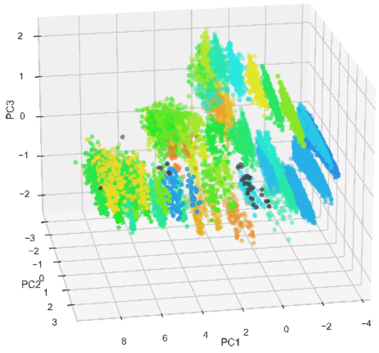


In (9), the observations falling within cluster "-1" are deemed as outliers by distance definition of ε = 60 and MinPts = 31.

DBSCAN estimated a total of 39 clusters, with only 41 observations appearing to be outliers. In (10) we visualize using three-dimensional PCA the results.

FIGURE 10

PCA ON DBSCAN CLUSTER ESTIMATION. N = 15,469



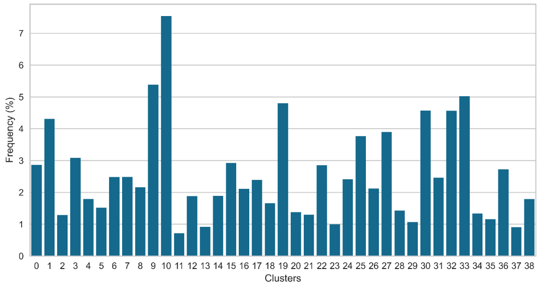
It is not extremely clear, but it appears as if observations of a certain cluster are horizontally distributed over what we may hypothesize is a principal component axis that is greatly influenced by time, as it is very linear. In terms of colors, how blue or how yellow a cluster is does not matter, as the order in which the clusters are colored is not representative of anything - it's just to distinguish them. But we can see that some blue-er clusters are closer together, and some yellower clusters are closer together. We can see the black observations indicating the outliers.

Furthermore, the cluster numbers were automatically ordered according to the hour: cluster 38 (shown in (9), not in (10)) is later in the day, and cluster 0 is earlier.

To finally make more use of the clustering results, the DBSCAN clustering estimate of 39 was applied to k-means, as the latter method allows us to perform a clustering analysis on much more data.

FIGURE 11

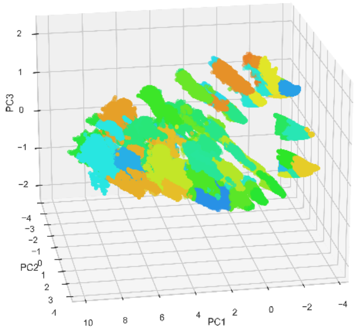
K-MEANS CLUSTERING, K=39. N = 1,546,992



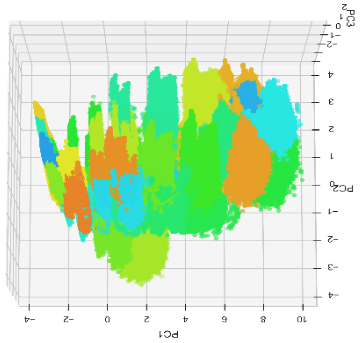
As shown in (11), k-means clustering found less observations to be falling in only two clusters, but nonetheless assigned observations in a spread manner. On average, based on DBSCAN, each cluster should hold 2.56% of the whole distribution and most clusters fall between 1-3%, with some exceptions. The most staggering similarity is found visualizing the 39-means results.

FIGURE 12

PCA, K=39. N = 1,546,992



(12) appears to be very close or identical to DBSCAN's PCA (10), on a much bigger scale. The patterns are a bit more sharp and identified, but the overall shape that reminds of tank tracks is preserved. From "above", PCA starts to remind of the very first few visualizations, with K=5.



Passing onto more interpretable data analysis, now some of the main aspects between the type of land use and PET will be analyzed.

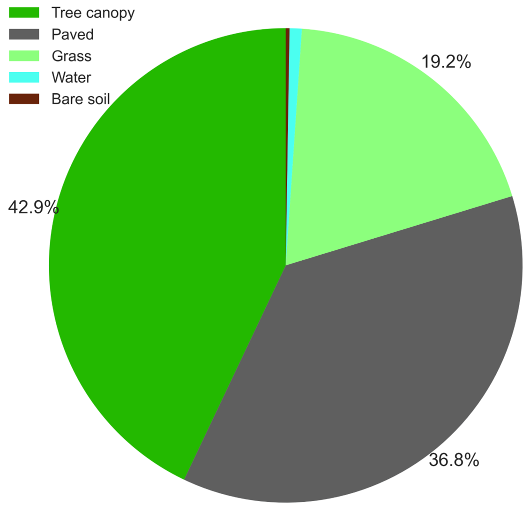
Several geodataframes require to be merged, in order to obtain data that features: the hourly PET, the belonging of a specific cluster for a specific phase of the day, and the type of land use, all associated with a specific geometry. To fix geometries, geopandas' "buffer" algorithm was used, and QGIS was used first intersect land use types (grass, pavement, water, bare soil, buildings) with only the subset of the center of Milan this report focuses on, and later it was also used to vectorize the data regarding trees (more specifically, the tree canopy covering certain areas with different intensities), to match the scale of the grid cells. Ultimately, every geodataframe geometry and attributes was merged on Python, in one dataset with the geometry, the land use, the tree presence, the hourly PET, and cluster belonging to different areas of the day. The clusters were simplified to the three solar phases of the day: morning (7 a.m. - 11 a.m.), midday (12 p.m. - 14 p.m.) and afternoon (15 p.m. to 19 p.m.).

Since the clusters are hour sensitive (the data was clustered also using the hour as a feature), in order to assign to a cell one cluster per morning, one per midday and one per afternoon, each cell was assigned to the statistical mode of the clusters it belonged to for specific hours[4].

Out of 67,053 cells that previously covered the streets, after merging and disposing of unmatching cells, we may end up with 59,322.

Figure 13

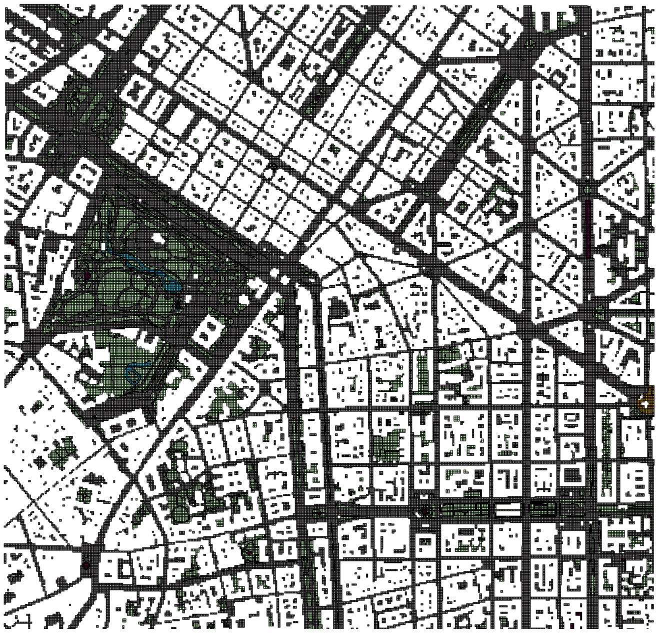
General pie chart of Land use (if covered by trees, it is classified as tree canopy)



For reference, the area that this paper focuses seems to have many trees, as displayed in (14).

Figure 14

Map of area of focus, categorizing on Land Use (grass, paved land, soil, water, except buildings)

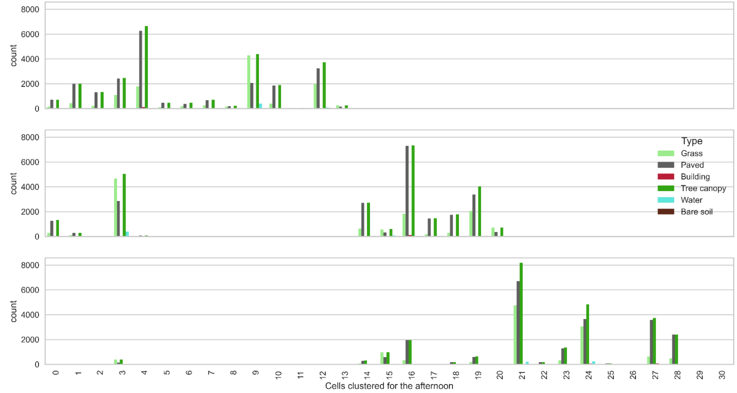


The latter map identifies the types of land in the focus area, colored to suggest different categories of use.

To prove that cluster assignment follows the hours of the day (given a certain influence from other factors), in (15) we can see how it shifts, for all types of land use.

Figure 15

Histograms by time of the day, hued on land use (Buildings excluded, see 14)



To understand better what is the role of land use towards PET, a multiple spatial lag regression model was implemented. The purpose was to consider the diffusion of the dependent variable (PET) across adjacent areas, since all of the cells are very close to each other. Since this model

[4] For example, if cell #45839 belongs to cluster 12 from 7 a.m. to 10 a.m., then its *majority class* or *mode* is the cluster 12, hence that cluster will represent the "morning cluster" for that cell.

required one value per cell, the median PET per cell was taken, regardless of the hour.
The types of land use were one hot-encoded to fit the numerical type of data required for the model.
The output of the model is included in the table below.

TABLE 16
Spatial lag Multiple Regression model: land use on PET

Variable	Coefficient	Probability
(Constant)	18.66240	0.00000
Building	-1.79987	0.09815
Grass	-0.85766	0.26907
Paved	-0.87079	0.25981
Tree canopy	-0.88003	0.25478
Water	-1.27084	0.16941
Median PET[5]	0.16901	0.00000

Buildings are more strongly negatively associated with PET, and also water. Although, these two land uses are barely present in the data, as most buildings were excluded from the analysis. More interestingly, grass is less negatively associated with PET than tree shade: this could indicate an error in handling the data on the author's side.
The model's AIC is equal to 4,498.900.

Ultimately, we may visualize the cluster's disposition for morning (17), midday (18) and afternoon (19) over the map.

Figure 17 (morning, clusters in {0;13}, from blue (lowest) to yellow (highest)

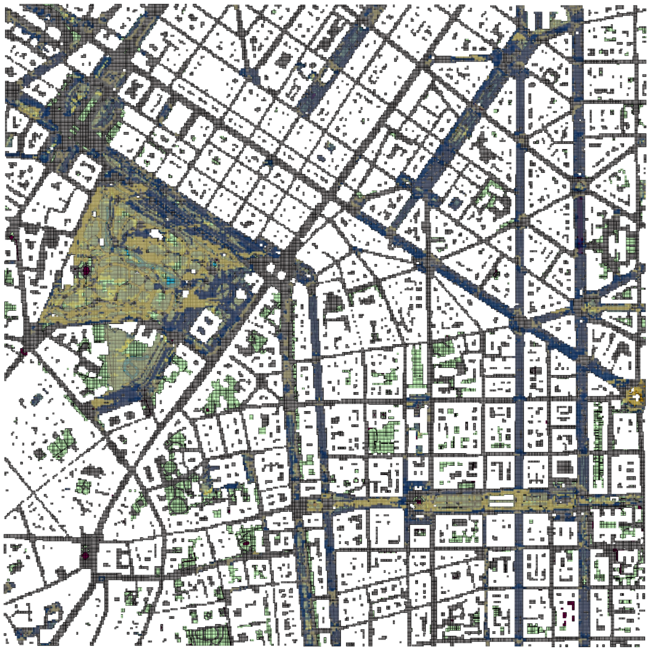
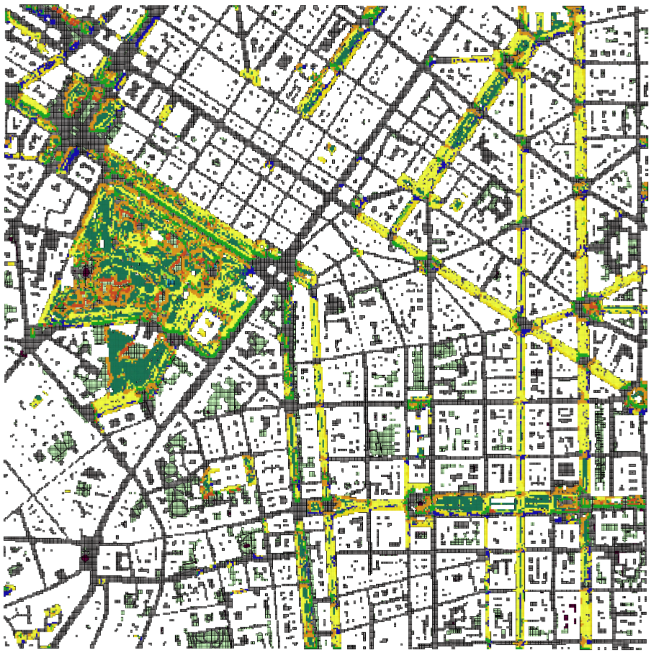
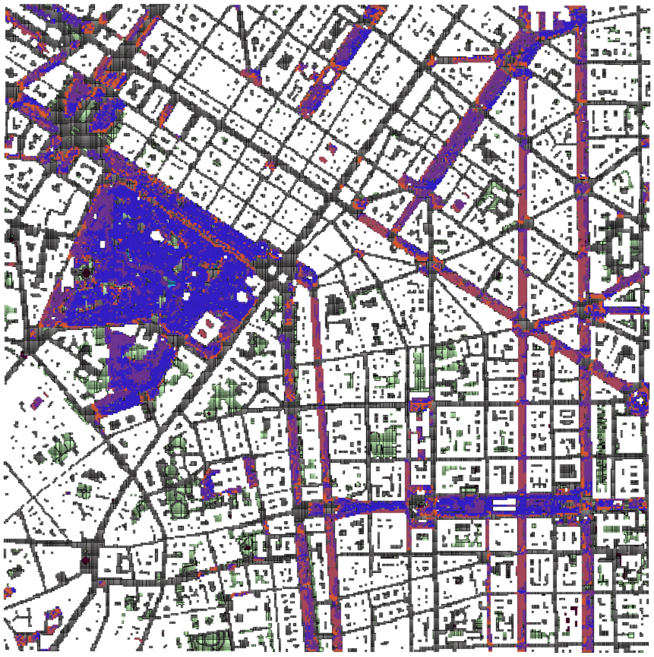


Figure 18 (Midday, Clusters in {0;20} from blue (lowest) to orange (highest)



[5] In the code this variable is referred to as "daily avg PET", even though it is the median.

Figure 19 (afternoon, Clusters in {14;28}, from blue (lowest) to red (highest)



Due to lack of resources and time, the latter visualization could benefit from improving, as it seems to be strongly biased - but no short fix was found. There are differences between values and they are displayed accordingly, using a polarized color scheme, but it appears that some cells covering the streets went "lost" during handling of the data in previous sessions. It appears that each cell that is colored according to the color scheme is either underneath a tree canopy or very close to it, although they were not selected on this basis.

In the morning (17) there is not as strong of a difference as there is in later phases of sunlight, but with the midday and afternoon more dissonance occurs.

The clustering seems to be very detailed and sensible to the specific area a cell is. A case in which this seems to be true is the park area (which can be seen from the lack of buildings compensated by greener areas, in (14)): throughout the whole day, it stands out from other areas and streets, either as cooler (midday and afternoon) or as warmer (morning).

Conclusion

The types of land use and the physiological equivalent temperature are associated with changes throughout the hours. While a spatial lag regression model was implemented, the findings require more effort for contextual significance, as based on the results there seems to be little difference between paved roads and grassy land, while clustering techniques disagree.

Moreover, several clustering techniques can be applied to predict, or more accurately interpret, specific phenomena that influence PET. In terms of methods implemented, DBSCAN can be of great support to geospatial analysis in scenarios where the number of clusters possible is a widely unknown parameter.

The focus on a subset of the whole city of Milan made it possible to cluster and visualize that greener areas are significantly different from paved land. Many issues are found simply in the processing phase, which put more emphasis on handling highly dimensional data with more accuracy and precision.

While this paper delved deeper into clustering, the author recognizes that in the future more resources shall go towards a more visual representation of geospatial data.

Acknowledgments

The contribution to practical experience and real-life problem solving is greatly appreciated, from Transform Transport's efforts to the teachers' of this course.

References and appendix

Sum of squared distances from each point to its assigned cluster center

$$J(c, \mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

[1] Lindberg F, Grimmond CSB, A Gabey, L Jarvi, CW Kent, N Krave, T Sun, N Wallenberg, HC Ward (2019) *Urban Multi-scale Environmental Predictor (UMEP) Manual.* https://umep-docs.readthedocs.io/ University of Reading UK, University of Gothenburg Sweden, SIMS China

[2] Matzarakis, A., & Amelung, B. (2008). Physiological Equivalent Temperature as Indicator for Impacts of Climate Change on Thermal Comfort of Humans. In M. C. Thomson, R. Garcia-Herrera, & M. Beniston (A c. Di), *Seasonal Forecasts, Climatic Change and Human Health: Health and Climate* (pp. 161–172). Springer Netherlands.

[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[4] Schubert E, Sander J, Ester M, Kriegel H P, Xu X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 2017, 42(3): 19

[5] Zuretti, M., Pedrazzoli, A., Ceccarelli, G., De La Hoz, A. (2023). Multi-Disciplinary Perspectives on Pedestrian Thermal Comfort and Walkability. In: *51th European Transport Conference 2023* (ETC 2023), 6-8 September 2023, Milan (Italy)

Author Information

Github repository:
https://github.com/FluveV/Milan_Thermal_Comfort_Analysis