Workshop on Visualisation in Environmental Sciences (EnvirVis) (2013)
O. Kolditz, K. Rink and G. Scheuermann (Editors)

Short Papers

Using K-Means Clustering for a Spatial Analysis of Multivariate and Time-Varying Microclimate Data

Kathrin Häb[1], Ariane Middel[2] and Hans Hagen[1]

[1]Computergraphics and HCI Group, TU Kaiserslautern, Germany
[2]Decision Center for a Desert City, Arizona State University, USA

Abstract

In this study, we propose a k-means clustering algorithm combined with glyph-based encoding method to analyze the spatial distribution and dependence of multivariate, time-varying 3D microclimate data. We obtained five climate variables, i.e. air and surface temperature, specific humidity, direct shortwave radiation and sensible heat flux, from an ENVI-met® simulation of a residential neighborhood in Phoenix, AZ. In a preprocessing step, we aggregated the 3D gridded simulation data by adding up value differences between two consecutive time steps for each grid cell over the entire simulation time to get a highly compressed view of the data without losing the spatial context. K-means clustering was then conducted in coordinate space by weighting each grid cell based on its difference to the spatial mean of temporal value differences. To reduce occlusion and to encode additional cluster member information, the visualization focused on the k-means cluster centroids. Resulting images show that the applied technique is suitable to provide a first insight into the spatial relationship of features based on their temporal variability.

Categories and Subject Descriptors (according to ACM CCS): I.5.3 [Pattern Recognition]: Clustering—Algorithms J.2 [Physical Sciences and Engineering]: Earth and atmospheric sciences—

1. Introduction

As urban population continues to increase, urban climatology research becomes more important for solving issues that may result from this growth. It not only seeks to describe and explain the effects of built structure on the atmospheric environment at different scales, it is also strongly connected to several areas of application such as urban planning and - in this context - air quality, human health and thermal comfort [MH87, Arn03]. Since the measurement of parameters contributing to the unique atmospheric conditions in cities can only be conducted pointwise in space and time and, therefore, lacks insight into the spatial and temporal continuity of meteorological processes, considerable research has been directed towards computational modeling of atmospheric processes in urban areas at various scales. The smallest scale is the so-called microscale. Models operating at that scale include, for example, the thermal comfort model Rayman® [MRM07] and the three-dimensional model ENVI-met® [Bru13, BF98]. As these simulation models become more and more complex due to advanced computing power, algorithms and visualizations that facilitate the analysis of increasingly large data sets need to be developed. Sophisticated visualizations contribute to the understanding of interdependencies between the factors responsible for feedbacks between urban form and the surrounding atmosphere and therefore need to be spatially explicit.

In the context of atmospheric research, the field of forecast verification provides methods that can also be applied to the urban microscale - not only for verification purposes. Traditionally, forecast verification compares corresponding grid cells of predicted and observed data, but this approach disregards the spatial connection between forecast and reality. If, for example, a predicted meteorological field is offset from the real event, but intensity and extension are computed correctly, the use of the traditional verification method would lead to a higher error rate than necessary. Therefore, it is not easy to interpret the verification results with regard to the physical properties of forecast performance [CWS*08]. In order to overcome this issue, forecast verification methods such as the so-called feature-based approach have been developed [CWS*08, GAB*09]. These techniques detect and isolate matching features in the forecast and observation

DOI: 10.2312/PE.EnvirVis.EnvirVis13.013-017

fields by different criteria, e.g. by a treshold. Then, the specific properties of such feature pairs are compared with regard to their size, position or intensity [CWS*08, LK10].

The challenge of finding spatial correlations and potential feedbacks between different atmospheric parameters due to physical processes such as advection can also be transferred from weather forecasting to the urban microscale. In this study, we applied an easily implementable feature-comparison to an ENVI-met® simulation output resulting from a microclimate study of typical neighborhoods in the Phoenix metropolitan area [MHB*12]. The time-varying character of the data was taken into account by defining "features" as areas with an above-average temporal variability over the entire simulation time. These features were retrieved individually for each variable. In order to gain insight into possible spatial correlations within the multivariate dataset, we applied a standard k-means clustering routine to each regarded variable independently. We ran the algorithm in the simulation's coordinate space, which allowed us to structure features within their spatial context. By using the cluster centroids as glyphs, we highlighted the features' locations, created a visual summary of each cluster's properties, and avoided clutter, facilitating an overview of the complex dataset.

2. Exploring the temporal variability in coordinate space using k-means clustering

The k-means clustering algorithm is a traditional clustering method based on the Euclidian distance, which makes it particularly suitable for identifying areas of similar data behavior in coordinate space. After defining an initial set of cluster centers at random spots within the investigated space, the algorithm allocates the surrounding data points to the nearest cluster and calculates the new coordinates of each center by averaging the allocated points' positions. Thus, the cluster centroids are iteratively refined until they represent the center of a local point pattern [WFH11].

2.1. Data preprocessing

The underlying ENVI-met® dataset is organized on a regular and almost completely equidistant grid. The simulation area comprises 215 x 195 x 34 grid cells and a spatial resolution of 1 m in each direction (Figure 1). The lowest grid cells between 0 m and 1 m height are each subdivided into five sub-grid cells with a vertical extent of 0.2 m each for a better resolution of surface-atmosphere exchange processes [Bru13]. The dataset contains both three- and two-dimensional components: surface parameters (0 m height) are organized on a two-dimensional grid and atmospheric parameters are simulated in a three-dimensional space (0-30 m height). Details about the creation of the underlying dataset as well as its microclimatic analysis can be found in [MHB*12].

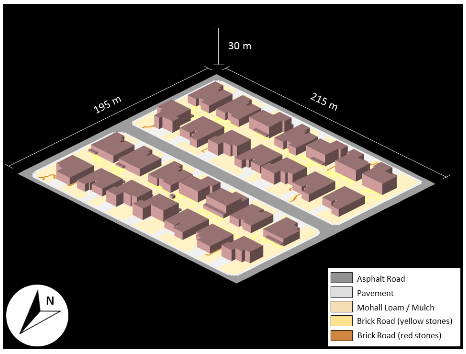


Figure 1: *The Raw Area, which served as a base for the microclimate simulation with ENVI-met®.*

We chose five sample variables for our study: air temperature (3D), specific humidity (3D), direct shortwave radiation (3D), surface temperature (2D) and sensible heat flux (2D). Since the purpose of our study was the detection of areas with similar behavior over time, which could then serve as "features" for a multivariate data analysis, it was necessary to preprocess the data. We used a comprehensive approach in order to take the time-varying component of the data into account. For each grid cell within the simulation area and each variable under investigaton, we added the absolute difference between two simulation timesteps (1 h) over the entire simulation time of 24 h. As a result, each grid cell has a single value for each variable representing its temporal variability. Since we defined "features" as areas with an above-average temporal variability, we computed the variable-specific spatial average of these values and extracted those grid cells where the mean was exceeded by more than 20%. Finally, we classified the extracted grid cells and their associated data according to the magnitude of difference ΔV to the area average with a stepsize of 10%. Fig-

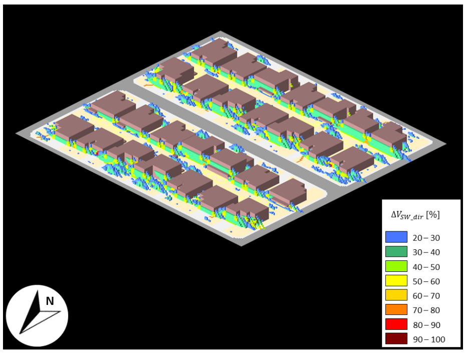


Figure 2: *The total sum of hourly differences in direct shortwave radiation, classified according to the distance to the area average.*

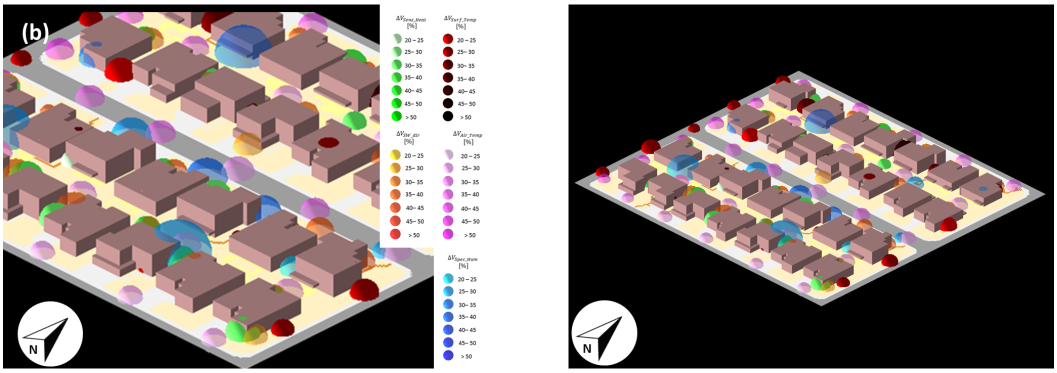


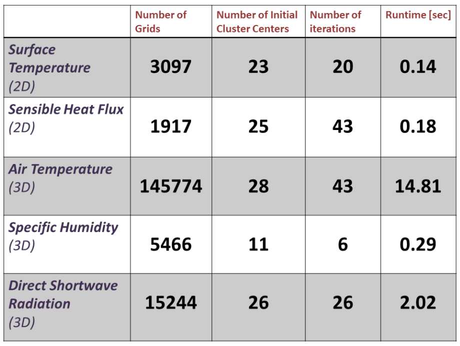
Figure 3: *Results: (a) close up view; (b) overview over area.*

ure 2 shows the extracted and classified grid cells for the direct shortwave radiation.

2.2. Initializing and running the k-means algorithm

The space coordinates of the extracted grid cells served as an input for the k-means clustering, which was conducted individually for each variable. Since the space coordinates alone do not account for the intensity of a variable's temporal variability at its location, we weighted the input positions according to the difference ΔV (in percent) to the variable's spatial average. This was implemented by decomposing the data into units with one unit corresponding to 10 % difference between the particular value and the area average. These units were then represented as a discrete dataset $\{x_1, x_2, \ldots, x_n\}$, where each x_i corresponds to a two- or three-dimensional vector of space coordinates for each variable and n is the total amount of data units per variable. This implies that each x_i can occur several times in the set of vectors, pulling the center of the k-means clustering to the location with the highest ΔV.

A common problem when applying the k-means algorithm is finding the proper initialization. Since the minimal distance from a point assigned to a center only minimizes the local Euclidian distance, the results of the algorithm are very sensitive to the sample of initial cluster centers [WFH11]. Inspired by two initialization routines described in [HLT*04], we chose the cluster centers $k_{i,var}$ individually for each variable var by subdividing the model area into 100 subareas of size 39 x 43 x 8. For each subarea, we used the local maximum as an initial $k_{i,var}$. We checked for other grid cells in each subarea with values ranging around this maximum and chose these positions as another cluster center if their distance to each local maximum exceeded half of the subarea's diagonal. If the subarea was only filled less than 1%, it was skipped and no k was added for this section. Once the initial cluster centers were found, the k-means clustering algorithm was run. The number of necessary iterations diverged between the individual variables, which is reflected in the specific runtimes (Table 1).



	Number of Grids	Number of Initial Cluster Centers	Number of iterations	Runtime [sec]
Surface Temperature *(2D)*	3097	23	20	0.14
Sensible Heat Flux *(2D)*	1917	25	43	0.18
Air Temperature *(3D)*	145774	28	43	14.81
Specific Humidity *(3D)*	5466	11	6	0.29
Direct Shortwave Radiation *(3D)*	15244	26	26	2.02

Table 1: *Number of extracted grid cells as described in section 2.1, number of cluster centers, number of necessary iterations and total runtime for each considered variable on an Intel® Core^{TM} i7 (2.5 GHz) with 8GB RAM (average of 10 runs).*

3. Visualizing the results

To visualize the results of the k-means clustering, we focussed only on the cluster centers. This approach avoids clutter when displaying multiple variables. In addition, the cluster centers can be used to encode summarizing information about the cluster properties. Hence, we rendered the centroids as spheres and designed their visual appearance according to three of their clusters' characteristics. The underlying microclimate variable is encoded as the base color of the sphere: pink spheres stand for air temperature, red spheres for surface temperature, yellow spheres for direct shortwave radiation, and so on. The second characteristic is the cluster's mean value, represented by the shade of base color used for the particular variable, with lighter colors representing lower mean values. The third property illustrated by the cluster centroid is the spatial spreading of the cluster's members, which is visually encoded as the diameter of the rendered sphere. For this purpose, the maximum Euclidian distance between each cluster's center and the cluster's members was measured and used as a base for the spheres' radius.

Although the visualization is focused on the cluster centers to reduce clutter, occlusion problems can occur due to the size of the rendered spheres. If two cluster centers are located adjacent to each other, the spheres can intersect. We solved this problem by introducing transparency of the centroids depicting the cluster centers for the 3D data.

Figure 3 shows the results of the k-means clustering and their visualization as described above. Since the spatial context of the cluster centers is important for the analysis of interdependencies between the individual microclimate variables, we included the built environment and the soil types in the resulting image as well.

4. Evaluation

The visualization shown in Figure 3 illustrates the advantages of the described approach. First, our method facilitates the analysis of how maximum temporal difference is distributed within one variable. For example, the location of the cluster centers for the direct shortwave radiation are located adjacent to the buildings in the model area. These are the spatial locations where this parameter is most variable in the course of a day due to the shading patterns of the built structures. The cluster centers of the air temperature are equally distributed over the near-surface part of the simulation area, indicating a high feedback between diurnal surface temperature changes and temperature changes of the adjacent air masses. The cluster centers for the specific humidity are located over impervious surfaces within the simulation area, since the amount of evaporation is potentially higher in these sections due to an increased soil moisture reaching the surface. These findings also highlight a second advantage of the algorithm, i.e. the areas of maximal temporal differences can be related to the underlying urban form. A third benefit lies in the comprehensible interdependencies between the microclimate variables. Thus, the centroid patterns within the simulation area show a relationship between the local maximum surface temperature differences and the slightly offset local maximum air temperature differences at the western border of the area. The offset of the air temperature's centroids to the east compared to those of the surface temperature can be associated with advective effects due to the western wind direction.

Although these findings can easily be derived using the resulting images, the approach also shows several drawbacks, which will be addressed in future work. First, the interdependencies between different variables are not quantified. To address this issue, we will implement measures such as the Euclidian distance between the cluster centers and the volume difference between the rendered spheres, which can be applied both to an intra- or an intercomparison of microclimate variables. Similar methods are already used in the feature-based approach to forecast verification, e.g., in [LK10] based on a Gaussian Mixture Model. Another possible solution is to quantify correlations in a statistical manner as presented in [SWMW09]. Their approach, based on a Canonical Correlation Analysis, has the disadvantage that the analysis is restricted to a maximum of two different variables.

Another drawback of our approach is that the resulting images do not offer any information about the absolute values of the variables, nor do they exhibit the direction of changes over the meaasured time span. This issue can be solved by using more sophisticated time series analysis methods as described in [WS09], where the time-activity curve (TAC) of each grid is taken as a base for similarity measures at different time scales.

5. Conclusion

In this study, we explored the temporal variability of selected variables based on a multivariate microclimate dataset derived from simulations with the three-dimensional model ENVI-met®. For each grid cell, we added the absolute difference between two time steps over the simulation time of 24 h. On this basis, we ran the k-means clustering algorithm to determine regions of similar temporal behavior for each regarded variable. In order to visually compare the resulting clusters, we focused the visualization on the cluster centers, which were used as glyphs to encode central characteristics of each underlying cluster.

To address the drawbacks of our approach, future work is underway, aiming at

(a) quantifying the spatial relationship between the different variables' temporal dynamic by introducing measures such as spatial offset and volume differences,

(b) taking into account the absolute values of the variables under investigation, and

(c) and including the direction of the value difference between two time steps.

The described approach provides a simple, yet insightful, overview of the underlying dataset and helps highlight interesting sections in the whole simulation area that are worth analyzing more closely.

6. Acknowledgements

This research was supported by the German Science Foundation (DFG, Grant 1131) as part of the International Graduate School (IRTG 1131) at University of Kaiserslautern, Germany, and the National Science Foundation (Grant SES-0951366, Decision Center for a Desert City II: Urban Climate Adaptation). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

References

[Arn03] ARNFIELD A. J.: Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *International Journal of Climatology 23*, 1 (2003), 1–26. URL: http://dx.doi.org/10.1002/joc.859, doi:10.1002/joc.859. 1

[BF98] BRUSE M., FLEER H.: Simulating surface-plant-air interactions inside urban environments with a three dimensional numerical model. *Environmental Modelling & Software 13*, 3 (1998), 373 – 384. URL: http://www.sciencedirect.com/science/article/pii/S1364815298000425, doi:10.1016/S1364-8152(98)00042-5. 1

[Bru13] BRUSE M.: Envi-met version 3.1 beta v, 2013. retrieved February 01, 2013. URL: http://www.envi-met.com/. 1, 2

[CWS*08] CASATI B., WILSON L. J., STEPHENSON D. B., NURMI P., GHELLI A., POCERNICH M., DAMRATH U., EBERT E. E., BROWN B. G., MASON S.: Forecast verification: current status and future directions. *Meteorological Applications 15*, 1 (2008), 3–18. URL: http://dx.doi.org/10.1002/met.52, doi:10.1002/met.52. 1, 2

[GAB*09] GILLELAND E., AHIJEVYCH D., BROWN B. G., CASATI B., EBERT E. E.: Intercomparison of spatial forecast verification methods. *Weather and Forecasting 24*, 5 (Oct. 2009), 1416–1430. URL: http://dx.doi.org/10.1175/2009WAF2222269.1, doi:10.1175/2009WAF2222269.1. 1

[HLT*04] HE J., LAN M., TAN C.-L., SUNG S.-Y., LOW H.-B.: Initialization of cluster refinement algorithms: a review and comparative study. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on Neural Networks* (2004), vol. 1, pp. 297–302. doi:10.1109/IJCNN.2004.1379917. 3

[LK10] LAKSHMANAN V., KAIN J. S.: A Gaussian mixture model approach to forecast verification. *Weather and Forecasting 25*, 3 (Feb. 2010), 908–920. URL: http://dx.doi.org/10.1175/2010WAF2222355.1, doi:10.1175/2010WAF2222355.1. 2, 4

[MH87] MAYER H., HÖPPE P.: Thermal comfort of man in different urban environments. *Theoretical and Applied Climatology 38* (1987), 43–49. URL: http://dx.doi.org/10.1007/BF00866252, doi:10.1007/BF00866252. 1

[MHB*12] MIDDEL A., HÄB K., BRAZEL A., MARTIN C., GUHATHAKURTA S.: Urban form, landscape design, and microclimate in Phoenix, Arizona. In *Proceedings ICUC8 - 8th International Conference on Urban Climate (ICUC8), August 2012, Dublin, Ireland* (2012). 2

[MRM07] MATZARAKIS A., RUTZ F., MAYER H.: Modelling radiation fluxes in simple and complex environments - application of the RayMan model. *International Journal of Biometeorology 51* (2007), 323–334. URL: http://dx.doi.org/10.1007/s00484-006-0061-8, doi:10.1007/s00484-006-0061-8. 1

[SWMW09] SUKHAREV J., WANG C., MA K.-L., WITTENBERG A.: Correlation study of time-varying multivariate climate data sets. In *Visualization Symposium, 2009. PacificVis '09. IEEE Pacific* (2009), pp. 161–168. doi:10.1109/PACIFICVIS.2009.4906852. 4

[WFH11] WITTEN I. H., FRANK E., HALL M. A.: *Data Mining: Practical Machine Learning Tools and Techniques*, 3 ed. Morgan Kaufmann, Amsterdam, 2011. URL: http://www.sciencedirect.com/science/book/9780123748560. 2, 3

[WS09] WOODRING J., SHEN H.-W.: Multiscale time activity data exploration via temporal clustering visualization spreadsheet. *Visualization and Computer Graphics, IEEE Transactions on 15*, 1 (2009), 123–137. doi:10.1109/TVCG.2008.69. 4