Thermal Comfort Clustering; Climate Classification in Colombia

Roland Hudson
Universidad de Los Andes | Colombia | r.hudson@uniandes.edu.co

Rodrigo Velasco
Universidad Piloto | Colombia | rodrigo-velasco@unipiloto.edu.co

Abstract

Our goal is to develop a climatic classification system that extends understanding of human comfort and guides the design of buildings to provide greater thermal comfort to occupants. We propose that using k-means clustering with multivariate climate data a classification system can be defined to objectively represent comfort zones in the tropics. Our study focuses on Colombia, but the approach extends to other countries located in the tropics.

Keywords Human comfort; climate classification; clustering

INTRODUCTION

The climate of a region is a key determinant of the functional requirements to be considered in the design of buildings. Understanding a site's climatic characteristics indicates design strategies that can define comfortable living conditions inside a building. Conditions in tropical regions are characterised by daily variations whereas at higher latitudes seasonal variation is the dominant characteristic. Understanding tropical climates is difficult due to a paucity of data but also because studies on tropical climates and design strategies in published literature are based on approximations developed for higher latitudes. The scarcity of data and appropriate design strategies means basic low-energy design approaches (Olgyay & Olgyay, 2015) (Givoni, 1981) are poorly understood and underutilized in the tropics.

Two independent approaches to understanding comfort and climate can be identified, comfort indices based on climate data and climate classification using geo-referenced data. Both originate from outside the domain of the built environment, comfort indices have been adapted for use in building design the latter has not been rigorously studied as a tool for building design.

The need to assess human response to thermal environment has led to the development of comfort indices (Jendritzky, de Dear, & Havenith, 2012). For the built environment, these indices have been extended and become part of the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standards (2013) . ASHRAE defines methods for predicting occupant thermal comfort while other indices are designed to reflect an individual's perception of temperature given a set of variables. The origin of these approaches is based on human-heat-balance models and empirical studies involving responses of human subjects exposed to controlled climatic conditions. By reducing climatic variables affecting comfort to a single-dimensional index, underlying causes are occluded. Empirical studies, use limited subject group that inherently introduce geographical, and demographic bias. Reducing

multivariate data to a single index highlights a data-processing arbitrariness (Rhee, Im, Carbone, & Jensen, 2008). Adapting published standards to equatorial climates (Toe & Kubota, 2013, Bravo & González, 2013) illustrates the bias in methodologies developed and applied in higher altitudes.

Three major climate classification approaches exist (Allaby, 2010); A - common vegetation boundaries, B - boundaries defined by moisture budget and potential evapotranspiration and C - atmospheric circulation of wind and major airmasses. Classification techniques tend to emphasise botanical zoning which is not equivalent to the spatial distribution of the potential for human comfort. Traditional methods are heuristically based which can prejudice lesser understood areas by diminishing the impact of certain data in classification (Zscheischler, Mahecha, & Harmeling, 2012).

K-means clustering is an unsupervised learning algorithm that works iteratively to identify *k* groups within the dataset, assigning each data point to one cluster, based on feature similarity. The technique have been used to classify climate, and successfully defined zones of similar multivariate characteristics (Bharath & Srinivas, 2015). Clustering methods have been applied to long term regional planning and developing design strategies for use of water resources, agriculture, food security and studying the impacts of localised climate change.

We propose a system to address the complexity of tropical climates and human comfort in the built environment. Using unsupervised learning on a climatic dataset of variables that determine comfort, we define a classification system designed for human rather than botanical purposes. Rather than extend an existing comfort index, we propose that objective data processing can minimise the geographical bias associated with empirically derived comfort indices developed in other latitudes.

METHODOLOGY

CLIMATE DATA

Our classification is based on ten years of multivariate, historical, monthly averages of gridded climate data from three sources. First, climatic data from 1901-2009 formatted as ESRI ASCII raster by CGIAR CSI (Cgiar-csi.org., 2012) based on original data from CRU (Jones & Harris, 2008). Secondly, wind speeds from the CCMP gridded surface vector winds (Wentz et al., 2015). Thirdly, elevation data is extracted from a hole-filled DEM of SRTM (Jarvis, Reuter, A., & Guevara, n.d.). The first phase of our process is the preparation the climate data.
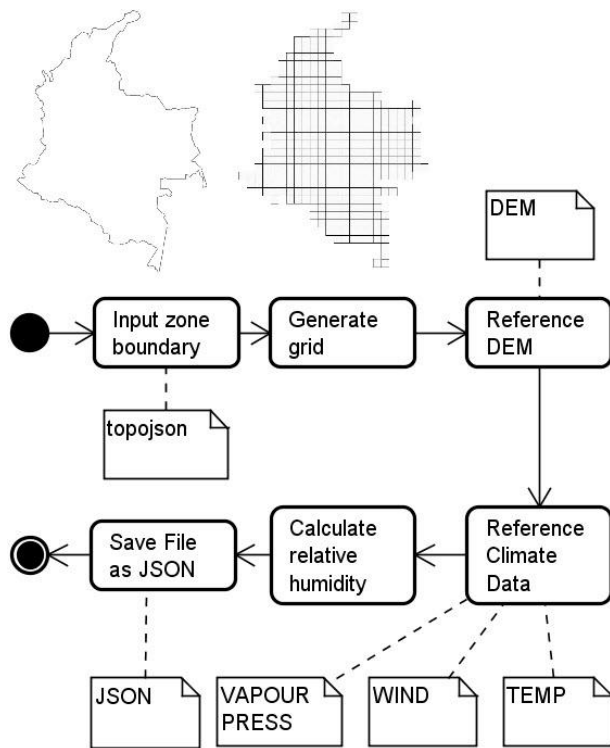

Figure 1. Flow diagram showing the stages of the data preparation process. Source: authors.

Figure 1 illustrates the data preparation steps. A C# dot net program was written that takes a topojson (Bostock, 2017b) format file as input, this describes the boundary (or collection of boundaries) that define the zone of interest. A point grid is generated at half degree latitude and longitude intervals filling the area(s) of study. Cross-referencing the grid to the DEM determines altitudes for each point.

The grid is used to extract climate data from a ten-year period (2000-2009) from the CRU and CCMP datasets. CRU data is formatted as ASCII ESRI raster format at the same resolution as the grid. CCMP data is in netCDF format, this was pre-processed with an independent Java program, written using the Unidata (2017) netCDF Java library. The netCDF data was converted to the ASCII ESRI raster format. Each ASCII raster file represents a single month of a year and contains data for earth's surface. Data points in the raster files that coincide with our grid vertices are found and stored with the georeferenced grid in arrays. Relative humidity derived using the ratio between vapour pressure and saturation

pressure. The data is averaged for the 2000-2009 period providing a typical year with monthly values. The prepared climate data can be saved to a Java Script Object Notation (JSON) file.

Phase two (figure 2) implements the k-means algorithm for a range of clusters. The JSON file containing the climate data from the previous stage is read. Three comfort indices are calculated for each climate data-point and used for evaluation later in the process. The climate data is scanned and maximum and minimums identified, and the data is normalised. Normalised data is used to define a set of tri-variate data-vectors (temperature, relative humidity and wind speed) that the clustering process learns from. Data-vectors are pre-processed using Principle Component Analysis (PCA) transformation to emphasise variation and make the normalised vector data more readily interpretable.
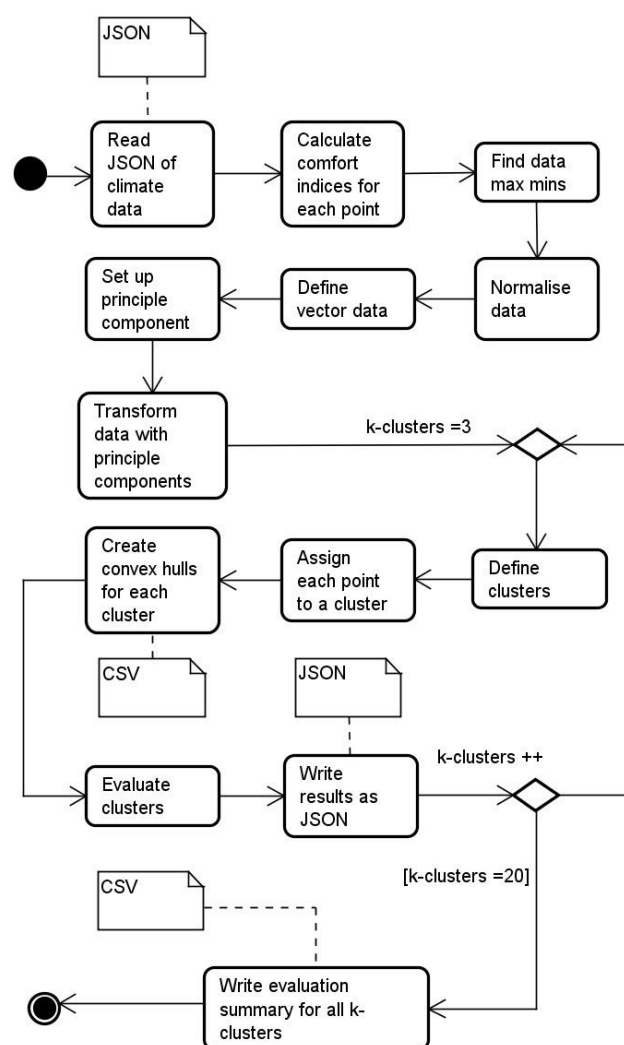

Figure 2. Flow diagram showing the stages for the k-clustering study. Source: authors.

The k-clustering study iterates through each of the desired number of clusters. The transformed data-vectors and number of clusters are inputs to the k-means algorithm which finds the clusters and assigns each data-vector to a single cluster. Data-vectors are grouped by cluster and convex hulls are found containing all the data-vectors in

each of the clusters. Hull geometry is written to a csv file and used for visualisation of the results.

Two groups of evaluation metrics are generated to assess the quality of each of the k-clusters in the study. Similarity measures can be defined by considering the Euclidean distances between data-vectors and measures of conformity are defined by comparing results to existing comfort indices.

Similarity is assessed by calculating the silhouette coefficient (S) and cluster distortion. S is a measure of how cleanly the data is grouped this is calculated per data-vector by finding its cohesion a_i (the average distance between it and all other vectors in the same cluster) and separation b_i (the average distance between it and all other vectors in other clusters). Silhouette coefficient for a point (S_i) is calculated as:

$$S_{i=\frac{a_i-b_i}{\max(a_i,b_i)}}$$

By averaging S_i for all the data-vectors a cluster silhouette coefficient can be calculated and by averaging across all clusters an overall silhouette coefficient can be calculated for the study. S_i ranges from minus one to one, negative values indicate poor clustering with overlap and positive values suggest better clustering. Distortion is the average distance between all the data-vectors in a cluster and the cluster centroid, this is calculated for all cluster and averaged across the solution. Smaller distortion values indicate better formed clusters.

Conformity to three existing comfort indices is assessed. Universal Thermal Comfort Index (UTCI) (Jendritzky et al., 2012), apparent temperature (AT) (Steadman, 1994) and Colombia's IDEAM index (González, 1998) were calculated for each data-vector using temperature, relative humidity and wind speed. For each index, the standard deviation in each cluster is calculated and averaged across the solution.

For each clusters' centroid the temperature, relative humidity and wind speed can be found by reverting the PCA transformation. The three existing indices can be calculated for the centroid. Mean absolute deviation for each index in each cluster is found by the averaging the differences between the cluster's central index and the index of each data-vector in the cluster. This is averaged across the solution.

Evaluation metrics are stored, the results of the clustering iteration is written to a JSON and the number of k-clusters incremented, and the process is repeated. Finally, a summary of all the evaluation studies is written to a CSV file. Results are visualised in an online viewer (Clima-colombia.org, 2018) created with JavaScript and the D3.js library (Bostock, 2017a) for dynamically manipulating online graphics using standard web formats. Clusters are mapped geo-spatially maps and cluster convex hulls are modelled in a three-dimensional parameter space.

RESULTS

We compared the results of our classification study to three existing comfort indices to assess how closely our results conformed to published standards (table 1). The

international UTCI (Jendritzky et al., 2012), 2. AT (Steadman, 1994) and Colombia's IDEAM index (González, 1998). Mean absolute deviations were calculated between data-points and the index for their cluster's centroid. Standard deviations between points within the same cluster were calculated and averaged across each study. To evaluate the geometric similarity of the clustering Euclidean distances between points within cluster and distances between clusters were examined and silhouette coefficients and mean cluster distortion determined for each study.

Table 1. Summary results of each k-cluster study in comparison to the indices (UTCI, IDEAM index and AT), cluster distortion and silhouette coefficient.

| n clusters	UTCI		IDEAM index		AT		DISTORTION	SILHOUETTE CF
	abs dev	SD	abs dev	SD	abs dev	SD		
3	4.50	5.29	1.53	2.80	5.20	4.95	0.06	0.36
4	3.26	3.58	1.40	1.80	3.74	3.09	0.05	0.41
5	2.92	3.07	1.54	2.01	3.52	2.90	0.05	0.37
6	2.76	2.92	1.46	1.88	3.36	2.75	0.04	0.35
7	2.67	2.84	1.40	1.75	3.36	2.68	0.03	0.34
8	2.55	2.78	1.43	1.67	2.96	2.34	0.03	0.35
9	2.52	2.71	1.37	1.59	2.95	2.30	0.03	0.34
10	2.40	2.59	1.52	1.68	2.75	2.16	0.03	0.35
11	2.39	2.61	1.50	1.66	2.70	2.17	0.03	0.34
12	2.37	2.54	1.46	1.59	2.75	2.15	0.03	0.32
13	2.31	2.47	1.44	1.60	2.73	2.11	0.02	0.31
14	2.26	2.42	1.40	1.53	2.74	2.08	0.02	0.31
15	2.26	2.38	1.37	1.47	2.78	2.04	0.02	0.30
16	2.27	2.36	1.35	1.41	2.82	2.02	0.02	0.30
17	2.25	2.32	1.33	1.37	2.81	2.01	0.02	0.30
18	2.27	2.29	1.34	1.32	2.94	1.99	0.02	0.28
19	2.24	2.27	1.37	1.35	2.86	1.96	0.02	0.28
20	2.20	2.25	1.41	1.40	2.82	1.95	0.02	0.28
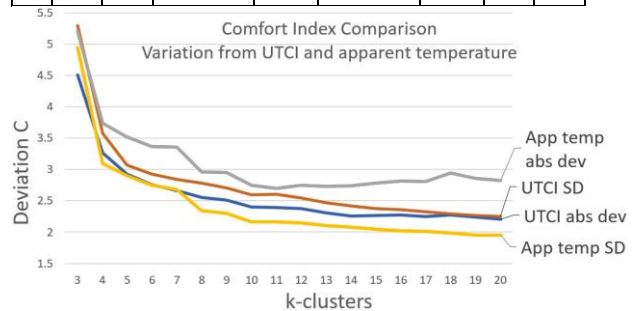
Figure 3. Standard deviation and Mean Absolute Deviation compared to UTCI and AT over the 3-20 k-cluster study.

Figure 3 shows how each of the clustering studies varies from UTCI and AT. Lower values for both types of measure indicate the clustering more closely represents the notion of comfort described by the indices. As the number of clusters increases these values drop until around ten clusters where the rates of reduction decreases.

IDEAM's index is developed specifically for Colombia and results in a range of values from 0 to 15 where 0 represents a very hot sensation. Unlike the other indices it

does not have units. Figure 4 shows the results of averaging the standard deviation and mean absolute deviation across all clusters for each of the studies. Lower values show closer alignment between clusters and the index, both rapidly drop in the first iteration of the study before slightly higher values at five clusters and an absolute minimum at 17-18 clusters and a second minimum at about nine clusters.
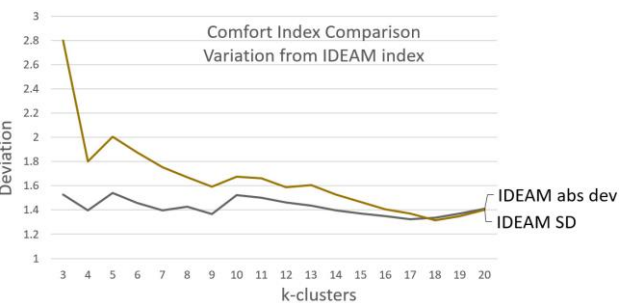
Figure 4. Standard deviation and Mean Absolute Deviation compared to IDEAM index over the 3-20 k-cluster study. Source: authors.

None of the evaluation metrics provide absolute support for an optimum number of clusters although each metric (except for distortion) indicates a significant change at around ten clusters (nine in the case of IDEAM's index). Beyond ten clusters the gains diminish rapidly. At four clusters both the comparison with IDEAM's index and the silhouette coefficient indicate low values suggesting good clustering performance.

Silhouette coefficient is the ratio of cohesion to separation and is in a range of minus one to one (Janert, 2010). Negative values indicate poor clustering where cluster radii are greater than inter-cluster distances, positive values suggest better clustering. Figure 5 shows all values are positive with a clear peak at four clusters and a second peak at ten clusters.
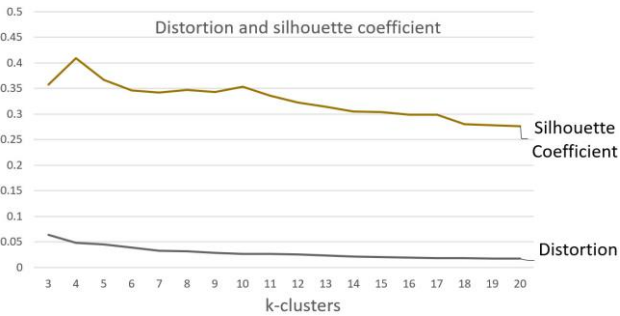
Figure 5. Distortion and silhouette coefficients over the 3-20 k-cluster study. Source: authors.

Distortion is measured per cluster as the average distance between the cluster points and its centroid. Lower distortion values represent better clustering as the more points are aggregated around cluster centroids the less the average distance. The lower line in figure 5 shows the average distortion for all clusters for each study showing low values reducing as the total clusters increases.

The silhouette coefficient and distortion values generally show good clustering, very small reductions can be observed in both. Very clearly defined and separated clusters would give silhouette coefficients closer to one here the values are around 0.3 which can be attributed to the smooth distribution of climate data.

Visualisation is the third approach to evaluation clustering. The online result viewer (figure 6) provides an interactive means of visual evaluation principally intended for the thermal comfort or meteorological specialist, but ultimately a prototype for dissemination to a wider range of users. Controls allow viewing different number of clusters at different months of the year. The scale indicates the three parameters associated with each cluster type and the UTCI calculated for that cluster.

Convex hulls of each cluster are modelled to show the distribution of clusters in three-dimensional parameter space (figure 6, top-left). The model can be rotated to view how clusters subdivide the parameter space supports choice of number of clusters.

DISCUSSION
OVERVIEW

Our goal is to develop a climate classification system that describes comfort. While climate can be described with classifications, current systems are problematic for representing human comfort:

a. Parameters used do not directly relate to comfort. The potential for the evaporation of sweat is key to understanding comfort. For this humidity and wind speed must both be represented. Current systems do not combine both, humidity is deduced from rainfall and temperature.

b. Seasonal variations captured in the zoning of current classifications do not represent tropical climate patterns typified by daily variations.

c. Boundaries between zones are arbitrarily defined.
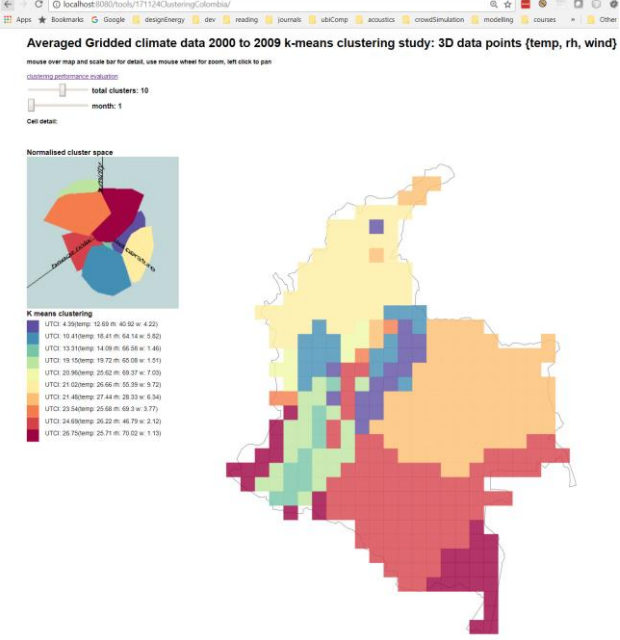
Figure 6. Online viewer for the Human Comfort Climate Classification for Colombia with ten classes (Clima-colombia.org, 2018). Source: authors.

These issues are addressed in our classification by:

a. Using parameters directly related to comfort (temperature, relative humidity and wind speed).

b.   Monthly data is used.
c.   Clustering is an objective procedure.

Comfort indices exist to describe human response to climate variables, but they are problematic because:

a.   They were created using human subjects in higher latitudes and do not represent the tropics.
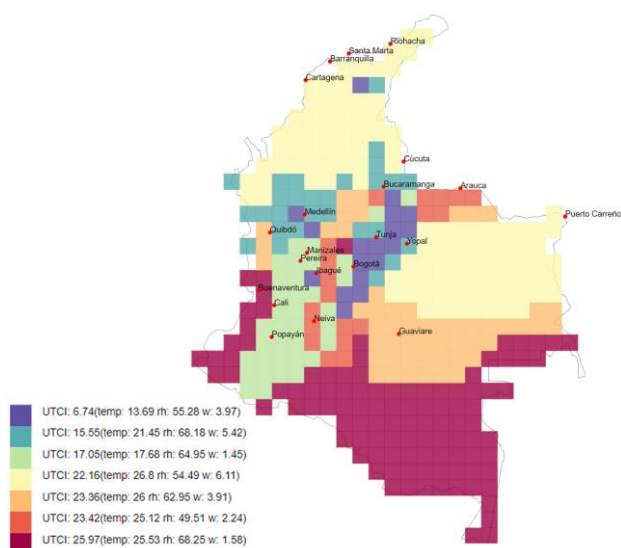b.   They seek to reduce comfort to a single value and occlude underlying causal data.

Shortcomings of comfort indices are avoided through:

a.   Use of data rather than sampling human perception.
b.   Actual parameter values are exposed.

IMPLEMENTATION FOR COLOMBIA

Statistical evaluation indicated four and ten clusters performed well, however, the selection of the number of clusters cannot be undertaken solely with metrics and must include domain knowledge to make informed decisions. Increasing cluster numbers serves to become more specific but at the cost of generality (Fovell & Fovell, 1993), lower cluster numbers represent a loss of detail but, they can enhance interpretation and generality. Climate data varies smoothly and hard edges between clusters do not exist, the choice is partly subjective and based on an adequate subdivision. We required sufficient clusters to represent the climatic diversity of Colombia but a quantity that was practically useful to apply knowledge extracted from the study.

Examining the clustering results using expert knowledge and known characteristics of Colombia suggested seven clusters. With seven clusters detail emerged describing climatic zones known not influenced by altitude. Our study classified monthly data, manual evaluation at this scale was cumbersome, to facilitate expert interrogation monthly parameters were averaged to produce an annual clustered classification (figure 7).


UTCI: 6.74(temp: 13.69 rh: 55.28 w: 3.97)
UTCI: 15.55(temp: 21.45 rh: 68.18 w: 5.42)
UTCI: 17.05(temp: 17.68 rh: 64.95 w: 1.45)
UTCI: 22.16(temp: 26.8 rh: 54.49 w: 6.11)
UTCI: 23.36(temp: 26 rh: 62.95 w: 3.91)
UTCI: 23.42(temp: 25.12 rh: 49.51 w: 2.24)
UTCI: 25.97(temp: 25.53 rh: 68.25 w: 1.58)

Figure 7. Human Comfort Climate Classification for Colombia with seven classes (Clima-colombia.org, 2018). Source: authors.

The first three clusters define altitude associated conditions. The first includes peaks on the eastern and western mountain ranges of the Andes (Bogotá and Tunja). The second cluster is located around the first and includes Medellín, Bucaramanga and Yopal, it has a

higher temperature and higher wind speed than the third cluster providing an over cooler zone. The third cluster covers the central southern part of the Andes (Cali, Popayán, Ibagué, Manizales and Pereira). The next four clusters define diverse combinations of the three variables at low altitudes, the fourth cluster covers the Caribbean region and central Orinoco, characterised by highest wind speeds and includes Santa Marta, Barranquilla, Cartagena, Riohacha, Cúcuta and Puerto Carreño. Cluster five is at the limits of the fourth (Guaviare and Quibdó). The sixth includes low altitude areas between the mountain ranges (Neiva and Arauca). Cluster seven has the highest humidity and lowest wind speed and covers the Amazon and lower part of the Pacific coast (Leticia and Buenaventura).

Annual averaged clustering permits comparison to the KG (figure 8a) and Holdridge (figure 8b) classifications created for Colombia with the same dataset. KG classification defines five classes in Colombia, that partly capture the dryness of the Caribbean region and demarcate higher altitudes of the Andes, but 85% of the Colombian territory is classed 'Equatorial with dry summer', including most cities below 1000 meters altitude. The Holdridge classification defines ten classes suggesting more detail than KG and identifying Caribbean dryness and higher humidity in the Amazon and Pacific, 60% of the country is classed 'Subtropical moist forest'. That both KG and Holdridge place most of the country in a single class, does not align well with our comfort clustering or our knowledge of the actual diversity.
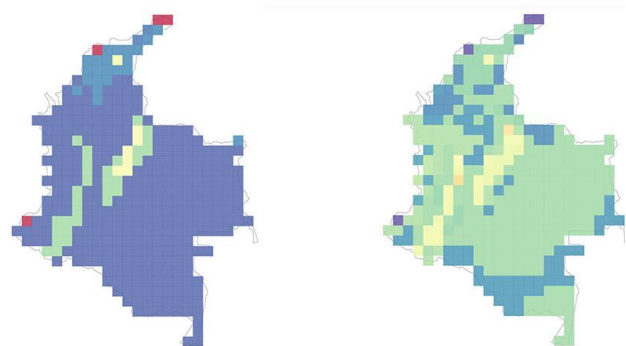
Figure 8. Köppen-Geiger Classification (a) and Holdridge Zones (b) for Colombia. Source: authors.

To further support to expert interrogation of results we clustered and mapped the data according to UTCI, AT and IDEAM index (figure 10). Visually, general similarities to our clustered classification can be observed. The UTCI and AT indices show smoother spatial variations (due to the lower dimensionality dataset) and higher detail in the Caribbean region, neither of which are captured in our classification. The IDEAM index leaves most of Colombia below 1000 meters within the same category, but our cluster classification captures a greater level of detail.

CONCLUSION

We present a climate classification that describes homogeneous zones of climatic conditions contributing to variations in human comfort. The development and application of the system in Colombia is described, where the goal is to support bio-climatic, low-energy design strategies aiming to improve comfort in the built

environment. The classification provides a geographical view that focuses on human habitation factors and includes temporal variation at monthly granularity. To develop this system of classification we used a k-means clustering algorithm, an unsupervised machine learning process capable of identification of similar groups within multivariate data-sets. We evaluated our results according to a statistical review of a range cluster numbers and by expert interrogation of visual output. Statistics enabled comparison to existing comfort indices and measured the similarity of clusters. The visualisation of our results allows manual interrogation with domain expertise and assessment driven by knowledge of the Colombian climate. The two modes of evaluation indicated that the applied approach could produce valid results with different numbers of comfort classes.

REFERENCES

Allaby, M. (2010). Climate classification. In *A Dictionary of Ecology*. Oxford University Press. Retrieved from http://www.oxfordreference.com.liverpool.idm.oclc.org/view/10.1093/acref/9780199567669.001.0001/acref-9780199567669-e-1137

ASHRAE. (2013). *2013 ASHRAE Handbook: Fundamentals*. *ASHRAE*. https://doi.org/10.1163/ej.9789004155947.i-937.23

Bharath, R., & Srinivas, V. V. (2015). Delineation of homogeneous hydrometeorological regions using wavelet-based global fuzzy cluster analysis. *International Journal of Climatology*, *35*(15), 4707–4727. https://doi.org/10.1002/joc.4318

Bostock, M. (2017a). D3.js - Data-Driven Documents. Retrieved 28 June 2018, from https://d3js.org/

Bostock, M. (2017b). TopoJSON. Retrieved 3 May 2017, from https://github.com/topojson/topojson

Bravo, G., & González, E. (2013). Thermal comfort in naturally ventilated spaces and under indirect evaporative passive cooling conditions in hot–humid climate. *Energy and Buildings*, (63), 79–86. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378778813001758

Cgiar-csi.org. (2012). CRU-TS v3.10.01 Historic Climate Database for GIS | CGIAR-CSI. Retrieved 5 November 2017, from http://www.cgiar-csi.org/data/uea-cru-ts-v3-10-01-historic-climate-database

Clima-colombia.org. (2018). Averaged Gridded climate data 2000 to 2009 k-means clustering study: 3D data points {temp, rh, wind}. Retrieved 16 September 2018, from http://www.clima-colombia.org/tools/kmeans/monthly3d/

Fovell, R. G., & Fovell, M. Y. C. (1993). Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate*. https://doi.org/10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2

Givoni, B. (Barukh). (1981). *Man, climate and architecture*. Van Nostrand Reinhold.

González, O. (1998). *Metodología para el Calculo del Confort Climático en Colombia: NOTA TECNICA DEL IDEAM*. Bogota. Retrieved from http://documentacion.ideam.gov.co/openbiblio/bvirtual/007574/Metodologiaconfort

Janert, P. (2010). *Data Analysis with Open Source Tools*. *A hands-on guide for programmers and data scientists*. Retrieved from http://books.google.co.kr/books?id=mTXnXCLXJYgC&printsec=frontcover&dq=data+analysis+with+open+source+tools&hl=&cd=1&source=gbs_api

Jarvis, A., Reuter, H. I., A., N., & Guevara, E. (n.d.). Hole-filled SRTM for the globe Version 4, Available at: CGIAR-CSI SRTM 90m Database. Retrieved 5 November 2017, from http://srtm.csi.cgiar.org

Jendritzky, G., de Dear, R., & Havenith, G. (2012). UTCI-Why another thermal index? *International Journal of Biometeorology*, *56*(3), 421–428. https://doi.org/10.1007/s00484-011-0513-7

Jones, P., & Harris, I. (2008). Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena. Retrieved 5 November 2017, from http://catalogue.ceda.ac.uk/uuid/3f8944800cc48e1cbc29a5ee12d8542d

Olgyay, V., & Olgyay, A. (2015). *Design With Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press.

Rhee, J., Im, J., Carbone, G. J., & Jensen, J. R. (2008). Delineation of climate regions using in-situ and remotely-sensed data for the Carolinas. *Remote Sensing of Environment*, *112*(6), 3099–3111. https://doi.org/10.1016/j.rse.2008.03.001

Steadman, R. G. (1994). Norms of apparent temperature in Australia. *Australian Meteorological Magazine*.

Toe, D. H. C., & Kubota, T. (2013). Development of an adaptive thermal comfort equation for naturally ventilated buildings in hot-humid climates using ASHRAE RP-884 database. *Frontiers of Architectural Research*, *2*(3), 278–291. https://doi.org/10.1016/j.foar.2013.06.003

Unidata. (2012). NetCDF-Java library and TDS version 4.6.9. Boulder. CO: UCAR/Unidata. https://doi.org/http://doi.org/10.5065/D6RN35XM

Wentz, F., Scott, J., Hoffman, R., Leidner, M., Atlas, R., & Ardizzone, J. (2015). Remote Sensing Systems Cross-Calibrated Multi-Platform (CCMP) 6-hourly ocean vector wind analysis product on 0.25 deg grid, Version 2.0, [subset: CCMP V2.0 Level-3.5]. Retrieved 5 November 2017, from www.remss.com/measurements/ccmp

Zscheischler, J., Mahecha, M. D., & Harmeling, S. (2012). Climate classifications: The value of unsupervised clustering. In *Procedia Computer Science* (Vol. 9, pp. 897–906). https://doi.org/10.1016/j.procs.2012.04.096

ACKNOWLEDGMENTS