# DATA FUSION FOR LOCAL CLIMATE ZONES CLASSIFICATION BASED ON ENSEMBLE OF CLASSIFIERS

*Leidiane do Livramento Andrade[1], Camila Souza dos Anjos[2], Marielcio Gonçalves Lacerda[1], Roberto Neves Salles[1]*

[1] Institute for Advanced Studies -IEAv, Brazilian Air Force
Caixa Postal 6044 – 12.231-970 – São José dos Campos - SP, Brazil
{leidiane.andrade; marielcio; salles}@ieav.cta.br
[2] Federal Institute of Education, Science and Technology of South of Minas Gerais - IFSULDEMINAS,
Praça Tiradentes, 416, Centro - 37576-000 - Inconfidentes-MG, Brazil.
camila.lacerda@ifsuldeminas.edu.br

## ABSTRACT

Multisource remote sensing data provide information of high relevance for classification and climate studies in urban areas and are of particular interest for regional and global climate science. To classify the urban environment using predefined Local Climate Zones we show a methodology that uses feature extraction from the multisource data and image segmentation as input data to an ensemble of classifiers and verify which algorithm has the best accuracy. The algorithms used were AdaBoost, Random Forest, Multi-layer Perceptron and an ensemble of those classification methods. The multispectral images were from Landsat 8 and Sentinel 2 resampled to 100m. LCZs were generated for Paris and Sao Paulo and the visual analysis and quantitative testing of results show the ensemble of classifiers had the best result for both cities with OA of 87.7% and 83%, Kappa of 0.81 and 0.80, for Paris and Sao Paulo, respectively.

***Key words*** — Ensemble of Classifiers, Feature Extraction, Data Fusion.

## 1. INTRODUCTION

The number of people living on cities is growing year by year and perspectives show a substantial growth in near future. This change is not only about movement of people to urban centers but also about the transformation of natural land cover into artificial and impermeable materials [1] altering the climate of cities.

To understand the urban climate characteristics of each city, it is necessary to know the urban landscape. Fusion of multisource information is today considered to be a typical scenario in the exploitation of remote sensing data [2], allowing a detailed and precise characterization of cities, by improving classification performance, providing additional information and solving situations where conflicts might occur. However, the extreme heterogeneity of urban areas has been shown one of the main causes of confusion in the classification of land cover and land use [3].

For this reason, it is necessary to use classifiers that incorporate knowledge of urban structures by means of previously collected training data, e.g. random forest (RF) classifier proposed by [4].

The World Urban Database and Portal Tool (WUDAPT) was conceived as an international collaborative project for the acquisition, storage and dissemination of climate relevant data on the physical geographies of cities worldwide [5]. The fundamental data that are acquired represent information on the form and functions of cities.

The form of cities can be described based on the Local Climate Zone (LCZ) scheme [6], which represent a generic description of natural and urban landscapes into categories based on climate-relevant surface properties. In this paper, we tested an ensemble of three algorithms and their individual performance for LCZ classification of the cities of Paris and Sao Paulo using Landsat 8 (L8) and Sentinel 2 (S2) data resampled to 100m of spatial resolution.

## 2. MATERIAL AND METHODS

To improve the LCZ classification made with the Random Forest algorithm [7], we extend the methodology to test the use of an ensemble of classifiers.

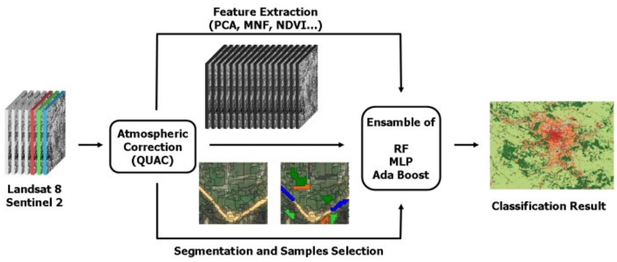


**Figure 1. Overview of the workflow.**

In this work, we used the images of Sao Paulo and Paris that were made available on the first stage of Data Fusion Contest (DFC) part of IGARSS 2017, together with their LCZs. In the new methodology, shown in figure 1, the previously consolidated procedures were adopted [7], plus the atmospheric correction and, after these steps, the images were classified with three different algorithms. The final result was obtained through majority voting.

### 2.1 Atmospheric Correction

Atmospheric correction of L8 and S2 images was done using an in-scene method called Quick Atmospheric Correction (QUAC) [8]. This is important considering the dataset contains different locations on different times. Even though it is an approximate method of atmosphere correction, research comparing QUAC and Fast Line-Of-Sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) [9], based on a model of radiance transfer, show similarities between resulting spectra of different materials of a scene using both atmospheric corrections [10] and also an absolute precision of ±15% on reflectance values obtained by FLAASH [11].

### 2.2 Feature Extraction, Image Segmentation, Local Climate Zones and Samples Selection

The first step of this methodology consists of the feature extraction that will be used as source data for the classifiers, i.e. the multispectral images go through transformations and arithmetical operations to generate new attributes that can effectively distinguish the classes of interest. Our attributes include the transformations Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF) and also band math attributes [7].

For the classification of Sao Paulo and Paris we used three scenes, being two scenes of Landsat 8 with 9 spectral bands (Coastal, Blue, Green, Red, NIR, SWIR1, SWIR2, TIR1 and TIR2) and one scene of Sentinel 2 with 10 spectral bands (Blue, Green, Red, RedEdge1, RedEdge2, RedEdge3, RedEdge4, NIR, SWIR1 and SWIR2), all resampled to 100m spatial resolution. In this case, there are 28 attributes from the spectral bands, 28 PCAs, 28 MNFs and 17 by band math generation, totalizing 101 attributes for each city. Those are the source data used for data mining by the all classification algorithms.

Image segmentation allows detection and separation of a set of pixels as objects of interest [12]. Thus an object is a region geographically defined by a segment and containing all layers, i.e. the spectral bands and the generated attributes. In this methodology, only the spectral bands were used to generate segments, were the parameters used were scale = 20, shape = 0.8 and compactness = 0.9.

Local climate zones are climate-based classifications of urban and rural sites that apply universally and relatively easily to local temperature studies using screen-level observations [6]. The 17 LCZ classes used are shown on Fig. 2. An available ground truth set was transformed into segments that resulted in 536 objects for Paris and 527 objects for Sao Paulo. Those selected samples represent less than 1% of the total area of the scenes. Average and standard deviation values of these sampled objects were split into two parts, 90% reserved for training the classifiers and 10% for classification validation [13]. There were no occurrences of class 7 Lightweight low-rise and class 13 Bush and Scrub on any of the two cities. On Paris there were not samples for class 3 Compact low-rise, class 10 Heavy industry and class 16 Bare soil or sand.

| | | | |
|---|---|---|---|
| Urban Local Climate Zones | 1 Compact high-rise | 6 Open low-rise | |
| | 2 Compact midrise | 7 Lightweight low-rise | |
| | 3 Compact low-rise | 8 Large low-rise | |
| | 4 Open high-rise | 9 Sparsely built | |
| | 5 Open midrise | 10 Heavy industry | |
| Rural Climate Zones | 11 A Dense trees | 15 E Bare rock or paved | |
| | 12 B Scattered trees | 16 F Bare soil or sand | |
| | 13 C Bush and Scrub | 17 G Water | |
| | 14 D Low plants | | |

**Figure 2. Local climate zone scheme [5] and legends.**

### 2.3 Classification Methods

In this work we investigated the use de two classifiers other than the RF, separately, and to further improve our results we combined the results through voting. The RF is a classification method introduced by [4] that uses a large collection of de-correlated decision trees.

Different random sets of samples are created from the original sampled objects. It uses fewer resources than conventional classification by decision trees because each tree uses only a fraction of the source input [14]. The results of classification from each tree for each object are called a class vote and the resulting classification is decided by majority of class votes [15]. The RF has three useful characteristics: internal error estimates, the ability to estimate variable importance, and the capacity to handle weak explanatory variables [16].

Were exported from eCognition all objects, including previously used sampled objects, this time with no associated class information. All 62988 objects for Paris and 38364 for Sao Paulo, were classified by the open source software WEKA 3.7 (Waikato Environment for Knowledge Analysis) with 100 trees and the resulting classification visualized on QGIS Desktop 2.12.3.

Another classifier chosen for this work is the Adaptive Boost ou AdaBoost [17], which is used together with a base algorithm, in our case the Random Forest. The training sample set was used for training. Later the RF is repeatedly invoked and for each call the AdaBoost gives the RF a different distribution of weights for each of the samples. The classification begins with associated weights evenly distributed. For each cycle of learning, the RF generates a hypothesis based on the current weights prioritizing the correct classification of data that has the largest associated weights. Iteratively those weights are reexamed so as to change the ones that are related to incorrectly classifications and a new round begins. After all rounds previously set, the AdaBoost combines the entire intermediate hypothesis to generate a final hypothesis with the least classification errors.

Multi-layer Perceptron (MLP) [18] is a neural network with multiple intermediate layers between input and output. The input layer is responsible for reception and propagation of the input information to the following layer, while the output layer receives information from the middle layers delivering the resulting classification. In this neural network architecture, the intermediate layers are composed of neurons bounded by weighted synapses. The learning is done by error backpropagation, when errors from the exit values are calculated and are propagated back until the beginning while adjusting the weights and the output values are again calculated. Lastly we did an ensemble of classifiers using Majority Voting [19], which consists of classifying an object as the class chosen by the majority of the individual classifiers.

## 3. RESULTS AND DISCUSSION

Source data for our method were multi temporal and multisource remote sensing images. The resulting four classifications of Paris and of Sao Paulo were quantitatively evaluated. For Paris, RF had the worst result compared to the other classifiers with an OA of 84.2%, while the MLP was the best classifier with OA of 85.9%. For Sao Paulo, the RF was also the worst performing algorithm with OA of 75%, while MLP again achieved the best results with OA of 77.4%. For both cities the Majority Voting achieved a better overall result, as can be seen on Table 2.

| | Paris | | Sao Paulo | |
|---|---|---|---|---|
| | OA (%) | Kappa | OA (%) | Kappa |
| AdaBoost | 85.9 | 0.78 | 75.5 | 0.72 |
| MLP | 85.9 | 0.79 | 77.4 | 0.74 |
| RF | 84.2 | 0.75 | 75.0 | 0.71 |
| Voting | 87.7 | 0.81 | 83.0 | 0.80 |

**Table 2. Classifiers Results.**

It can be seen that, although AdaBoost was better than RF for Paris, the same didn't happen with Sao Paulo, where the results were similar. The reason might be the reduced set of training samples.

For a better evaluation of the results we made an analysis using the confusion matrix where the size of circles is proportional to correctness, if they are on the main diagonal, or confusions, if they are elsewhere (Figure 3). The colors used are similar to those of the LCZs. Analysis of the ensemble confusion matrices show that Paris had C2 (compact midrise) correctly classified while C5 (open midrise) was the most misclassified class, and the Rural Climate Zones had a very good classification. The remaining difficulties were in separating urban classes because different LCZs have the same materials on urban environments but not the same height, having similar interclass spectral signatures. The class C8 (large low-rise) was also correctly classified and this is probably because it is a well-defined class in terms of composition of materials.

For Sao Paulo we can see that C1 (compact high-rise) and C4 (open high-rise) were correctly classified while C3 (compact low-rise), C6 (open lowrise) and C8 (large low-rise) being slightly misclassified. The class C5 (open midrise) was also the most misclassified with only 33.3% accuracy, but in Sao Paulo it was misclassified as an industrial class (C10 – heavy industry) and also as C3 (compact low-rise). Finally, C9 (sparsely built) was completely misclassified as C6 (open low-rise) probably due to the similar materials present on those classes. Rural Climate Zone classes were all well classified except C16F that was half misclassified as C5 (open midrise). Vegetation classes also had confusion between themselves (C11A and C12B).

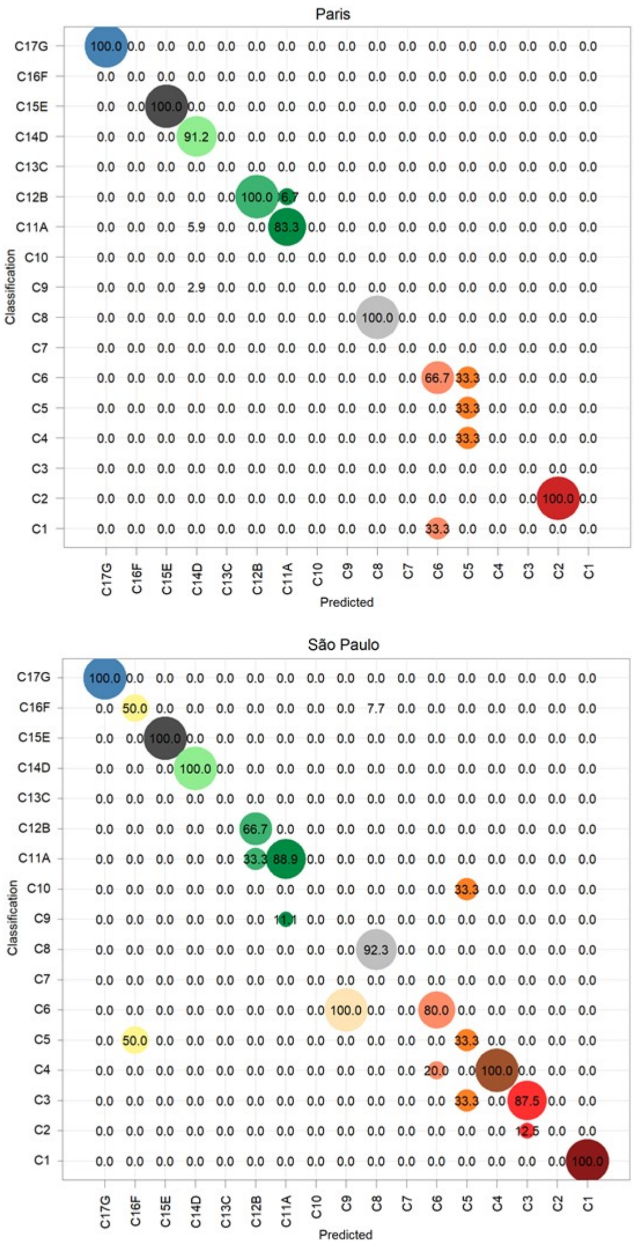


**Figure 3. Percentage analysis of the confusion matrix of the classification resulting from the ensemble of classifiers.**

## 4. CONCLUSIONS

This work presented a method to classify urban environments with low spatial resolution, more specifically, LCZ mapping with data from L8 and S2 resampled to 100m. The established workflow based on attribute generation from source images and use of three classifiers and their ensemble produce reliable results. The use of the ensemble of classifiers brings a better precision than using any single classifier by itself. Meanwhile, due to a small set of training and validation samples, the AdaBoost classification could not significantly improve the RF results as expected. It must be noted that restricted spectral separability of certain materials imposes limitations on the method. Further works could use data fusion with the inclusion of normalized digital surface models or laser scanning to reduce confusions between urban classes.

## 5. ACKNOWLEDGMENT

The authors would like to thank the WUDAPT (http://www.wudapt.org/) and GeoWIKI (http://geo-wiki.org/) initiatives for providing the data packages used in this study, the DASE benchmarking platform (http://dase.ticinumaerospace.com/), and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. Landsat 8 data available from the U.S. Geological Survey (https://www.usgs.gov/). Original Copernicus Sentinel Data 2016 available from the European Space Agency. (https://sentinel.esa.int).

## 6. REFERENCES

[1] Mills, G., “Cities as agents of global change”, *Int. J. Climatol.*, v.27, pp.1849–1857, 2007.

[2] Dalla Mura, M.; Prasad, S.; Pacifici, F.; Gamba, P.; Chanussot, J. and Benediktsson, J. A., “Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing”, Proc. IEEE Geosci. Remote Sens. Society, v. 103, n. 9, pp. 1585-1601, 2015.

[3] Anjos, C.S.; Almeida, C.M.; Galvão, L.S. and Garcia Fonseca, L. M., “Urban feature extraction on simulated WorldView-2 images”, Jt. Urban Remote Sens. Event 2013, JURSE 2013, v. 856, pp. 284–287, 2013.

[4] Breiman, L., “Random forests”, *Machine Learning*, v. 45, n. 1, pp. 5-32, 2001.

[5] Mills, G. J.; Ching, J.; See, L.; Bechtel, B. and Foley, M., “An Introduction to the WUDAPT project.” *Proc. ICUC9. Meteo Fr.,* n. February, pp. 6, 2015.

[6] Stewart, I. and Oke, T., “Local Climate Zones for Urban Temperature Studies”, *Bull. Am. Meteorol. Soc.*, v.93, pp.1879–1900, 2012.

[7] Anjos, C.S.; Lacerda, M.G.; Andrade, L.L. and Salles, R.N., “Classification of Urban Environments Using Feature Extraction and Random Forest”, In Proc IEEE International Geoscience and Remote Sensing Symposium (IGARSS) – Winners of Data Fusion Contest (DFC), Fort Worth - Texas, pp.1205-1208 , 2016.

[8] Bernstein, L.S.; Adler-Golden, S.M.; Sundberg, R.L.; Levine, R.Y.; Perkins, T.C.; Berk, A.; Ratkowski, A.J. and Hoke, M.L., "A New Method for Atmospheric Correction and Aerosol Optical Property Retrieval for Vis-SWIR Multi and Hyperspectral Imaging Sensors: QUAC (QUick Atmospheric Correction)," In Proc 3th JPL Airborne Earth Science Workshop, April 2004.

[9] Adler-Golden, S.; Berk, A.; Bernstein, L.S.; Richtsmeierl, S.; Acharyal, P.K.; Matthew, M.W.; Anderson, G.P.; Allred, C.L.; Jeong, L.S. and Chetwynd, J.H., “FLAASH, A MODTRAN4 atmospheric correction package for hyperspectral data retrievals and simulations”, In Proc 7th Ann JPL Airborne Earth Science Workshop, pp.9-14, 1998.

[10] Bernstein, L.S.; Adler-Golden, S.M.; Jin, X.; Gregor, B. and Sundberg, R.L., “Quick atmospheric correction (QUAC) code for VNIRSWIR spectral imagery: Algorithm details”, In Proc IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), China, pp.1–4, 2012.

[12] Costa, L. F. and Cesar Jr., R. M., “*Shape analysis and classification: theory and practice*”, CRC Press, Boca Raton, pp.680, 2000.

[13] McLachlan, G. J.; Kim-Anh, D. and Ambroise, C. “*Analyzing microarray gene expression data*”, Wiley, pp.328, 2004.

[14]Gislason, O.; Benediktsson J.A. and Sveinsson, J.R., “Random Forests for land cover classification”, *Pattern Recognition Letters*, v.27, n.4, pp.294-300, 2006.

[15]Hastie, T.J.; Tibshirani, R.J. and Friedman, J.H., “*The elements of statistical learning: data mining, inference, and prediction*”, Springer, New York, pp.536, 2009.

[16]Walton, J.T., “Subpixel urban land cover estimation: comparing cubist, random forests, and support vector regression.” *Photogrammetric Engineering and Remote Sensing*, v.74, n.10, pp.1213-1222, 2008.

[17] Freund, Y. and Schapire R.E., “Experiments with a new boosting algorithm” In Proc 13th International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

[18] Haykin, S., “*Neural Networks and Learning Machines*”, third ed., Pearson, New Jersey, pp.938, 2009.

[19] Kittler, J.; Hatef, M.; Duin, R. P.W. and Matas, J., “On combining classifiers”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.20, n.3, pp.226-239, 1998.