

Unsupervised Classification of Multispectral Imagery for Landcover Type Analysis

Collaborators

Musfiq Rahman, Javed Tomal, Francesca Ramunno¹
Brady Holliday, Joanna Wand, Dana Hicks²
Ashlin Richardson³

Collaborators	1
Abstract	1
Introduction	2
Related Works	3
Data and Methods	4
Clustering Results	5
Summary and Recommendations for Future Work	7
References	8
Appendix -- Source Code	8

Abstract

The British Columbia Wildfire Service (BCWS) Fuel Type Layer is a geospatial data layer that's an important operational tool informing wildfire management activities in British Columbia. The BCWS FTL is also a valuable reference for wildland fire management research. Here we pursue improving the consistency of fuel type information available by research and development in AI/ML and Remote Sensing (RS) methods application, leveraging high-quality open data. Specifically, we evaluate a number of unsupervised classification methods for multispectral imagery with regards to their potential for discriminating among forestry land-cover types, with the objective of improving forest-attributing for the purpose of wildfire management decision-making support. Visualization for high-dimensional imagery and classification results is provided. Data-driven methods are used to assess the number of categories to be discriminated. Sensors investigated include Landsat-8 and Sentinel-2. Our experiments are conducted using non-proprietary tools for reproducible science, most notably the Python language.

¹ Thompson Rivers University (TRU)

² BC Wildfire Service (BCWS), Predictive Services Unit (PSU)

³ Digital Platforms and Data Division (DPDD), Office of BC Chief Information Officer (OCIO)

Introduction

Wildfires scorching vast tracts of vegetation and forcing mass evacuations seem to be a recurring phenomenon in British Columbia, Canada. In summer 2018 over 12,984 km² of the province burned, breaking all past historical records for the last century [1]. These trends may continue as global temperatures edge up due to climate change. To understand and manage extreme wildfire behaviour, technological innovation is required to better anticipate wildfire behaviour changes.

Biomass, wind speed, wind direction, humidity, temperature, and topography are all characteristics of the fire environment affecting fire behaviour. Moreover fallen trees, branches, forestry operations waste or other dead biomass may be the most important factor. Downed branches and leaves in the forest are categorized as readily-available fuel. When fuel moisture is high, fires don't ignite readily, or at all, since the fire's heat energy is spent evaporating the water. When fuel moisture content is low, flames develop: fires more easily start and spread.

In general, 'fuel' attributes include: vegetation and biomass structure, biomass loading, dominant species (especially for treed landscapes), characteristics of the forest floor and forest health issues (e.g. infestation by bark beetles or other insects) all of which affect the flammability and availability of biomass for combustion. Estimating fuel type is an important and challenging task which may be described qualitatively using discrete fuel type, or quantitatively using a number of variables relating to fuel structure or the amount of available fuel.

The Canadian Forest Fire Danger Rating System (CFFDRS) is a primary modelling system used across Canada by most operational fire management agencies[2], which uses a qualitative fuel type description in its Fire Behaviour Prediction (FBP) System [3]. A national Fuel Type Layer (FTL) and a provincial FTL both implement the FBP fuel types description which consists of 16 standard types of fuel [4]. Many wildfire management agencies including the British Columbia Wildfire Service (BCWS) rely on the CFFDRS system for many aspects of wildland fire management including: operational fire behaviour prediction, fire season preparedness, resource pre-location, and regulation of industrial or recreational activities.

One or more satellite image(s) of a fire-prone area overlayed with ground reference data including FTL categories is an intelligent starting point for fire behaviour forecasting, tactical planning research and investigating the possibility of more detailed fuel assessment, using Remote Sensing (RS) imagery to gather information inferentially, reducing the reliance on expensive ground-reference data collection studies. Image data corresponding to a given label in a ground reference map may in fact represent multiple distinguishable objects, hence sub-categories within a given ground-reference class may need to be determined to better understand the situation (e.g., using clustering). Moreover, clusters found by "unsupervised" Machine Learning (ML) methods may not be unique to an individual ground-reference class: an individual cluster can potentially relate to multiple classes in the ground-reference data.

Therefore, identifying clusters and studying their relation with the ground reference classes is important. In this project we use high-definition satellite images overlaid with spatial ground-reference data over a fire-prone area, applying two different Machine Learning methods to find statistically meaningful clusters of natural objects (e.g., trees, terrain, rivers, urban areas, etc.) with the goal to assess the ability to characterize fuel types. In this case we compare ground-reference layers extracted from the BC Vegetation Resource Inventory (VRI), overlaid with imagery from modern sensors.

Related Works

In recent years wildfires in British Columbia (BC), Canada highly concerning. Given the right conditions, BC wildfires can burn millions of hectares costing millions of dollars and causing a provincial state of emergency [5] such as in 2017 and 2018. Problems induced by wildfires include infrastructure costs, business losses, evacuation costs [6], mental and physical health effects [7] and carbon-dioxide emissions exacerbating hotter weather and contributing to climate change [8].

Research in forest fire behaviour prediction dates to the 1920's in Canada [9] and has produced useful fire management tools. For example, the Fire Weather Index (FWI) which provides ratings of relative fire potential based upon weather information, but does not address fuels or topography considerations, is in use across Canada since 1970. The standardised FBP system which specifies sixteen variations in fuel types to model common vegetation structures across Canada, was first issued in 1984. The FBP fuel types are useful for fire managers predicting fire behaviour attributes such as rate of spread, fire intensity and potential fire growth with respect to a given time interval. Accurately mapping fuel type information is important for fire management and response agencies. Both FWI and FBP subsystems of the CFFDRS.

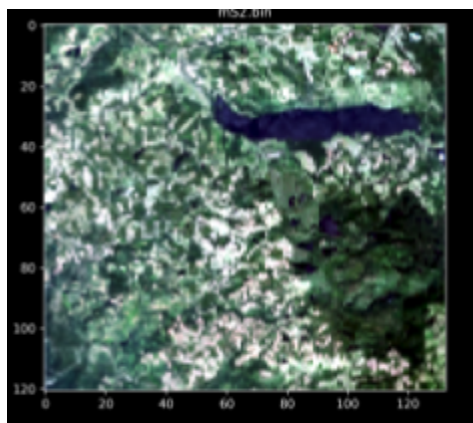
Predicting forest fire behaviour is an active research area. Artés et al. (2016) [10] proposed a genetic algorithm to predict forest fire propagation using wind as an input parameter, optimized to reduce a Fire Simulator System's execution time. Rodriguez-Aseretto et al. (2013) [11] proposed a wildfire behaviour model and dynamic data-driven application to predict the spread of large fires in Europe.

Predicting wildfire with Machine Learning (ML) is still a youthful area of research. Stojanova et al. (2006) [12] predicted Slovenian forest fires with decision-tree, random-forest and other models using forest-structure GIS, weather prediction and satellite imagery data. Downard et al. (2017) [13] used the K-Means algorithm to cluster forest fire incidents to help effectively place fire-fighting resources. Wijayanto et al. (2017) [14] proposed using the Adaptive Neuro-Fuzzy inference system (ANFIS) on forest fire hot-spot data to classify hot-spot occurrence in Central Kalimantan, Indonesia. In 2018 Kaggle [15] presented an ML-based competition for predicting forest fires. In this work we draw inspiration from the studies above, applying ML methods to address fire environment aspects that are unique to British Columbia.

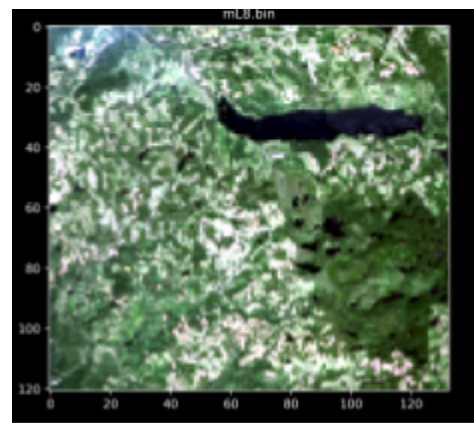
Data and Methods

This study uses the following multispectral data frames selected over an area of interest near Kamloops, British Columbia:

Sensor	Scene ID
Landsat-8	LC08_L1TP_046024_20170714_20170726_01_T1
Sentinel-2	S2A_MSIL1C_20170804T190921_N0205_R056_T10UFB_20170804T191650



Sentinel-2



Landsat-8

The data were extracted using ESA Snap [16] and converted to ENVI format IEEE 32-bit floating-point type with BSQ “band-sequential” data ordering, for convenient input for analysis. The data were overlaid onto the geo-coordinates of the Sentinel-2 scene using the `gdal_warp` utility[17], downsampled by a factor of four using the `gdal_translate` utility[18], and subset with QGIS[19] to a portion of the image intersection that was almost entirely cloud-free for a resulting image size of 399x363 pixels. Finally the data were down-sampled by a further factor of two in order to meet the memory requirements of the hierarchical clustering algorithm on a standard laptop with 16GB ram. The resulting Sentinel-2 and Landsat-8 images had 12 bands and 11 bands respectively. Fusing the two images resulted in 23 bands available for analysis.

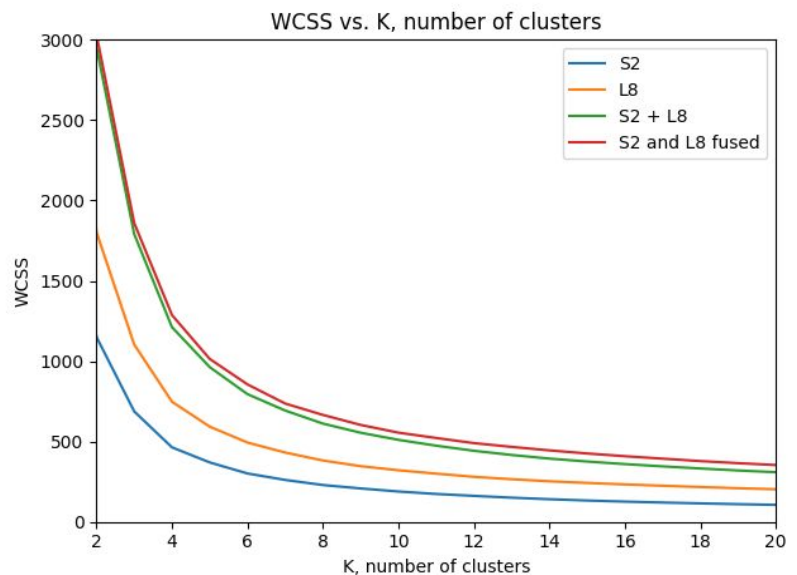
The images were read with Python using the Python interface for GDAL. NumPy was used to store the data and reshape the arrays. For visualization a band-selection was applied, the data were scaled in the range 0-1, with a 2% linear-stretch transformation applied to achieve an effective level of contrast in the visualization, and the result saved in png format. Images so inspected, as above, were then used for clustering.

Clustering Results

K-Means and Hierarchical Agglomerative Clustering (HAC) were used to create colormaps of the images, for visually representing clusters. Before running clustering algorithms, data are reshaped to match the format expected by the “scikit-learn” Python package we used. For consistent visualization, cluster label results from the initial run (Sentinel-2 data, KMeans algorithm) were identified with those from other runs using a simple combinatorial matching [20]. Moreover a high-contrast colouring scheme [21] assigned each label a colour.

K-Means

Prior to determining a number of clusters for overall study, we ran K-Means on Sentinel-2 data, Landsat-8 data and finally a fused set containing Sentinel-2 and Landsat-8 data. Below Within Cluster Sum of Squares (WCSS) is plotted for these cases (inertia property of scikit-learn package). The sum of the results for the two cases is also shown (red line).

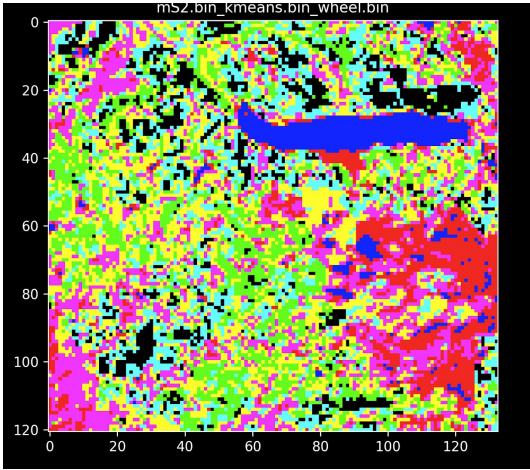
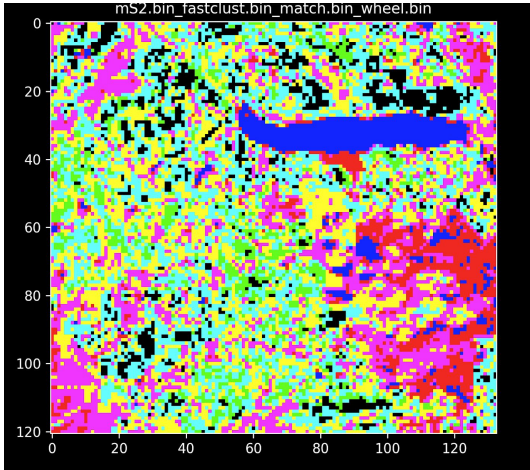
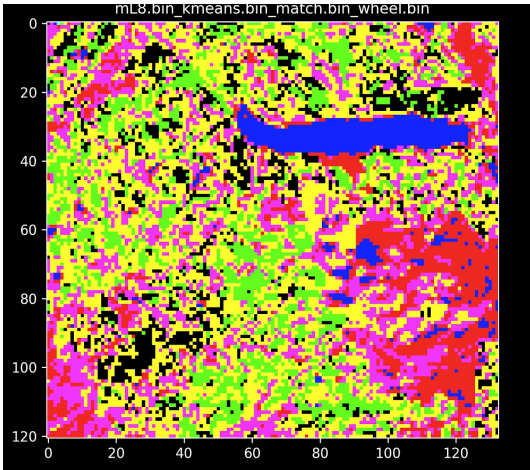
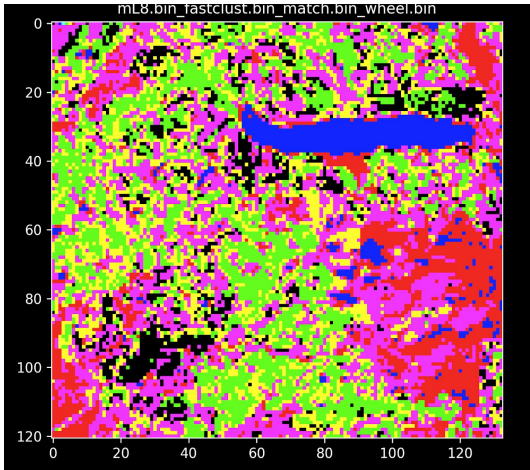
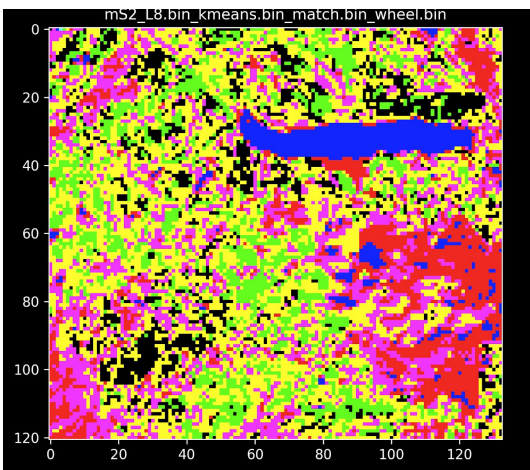
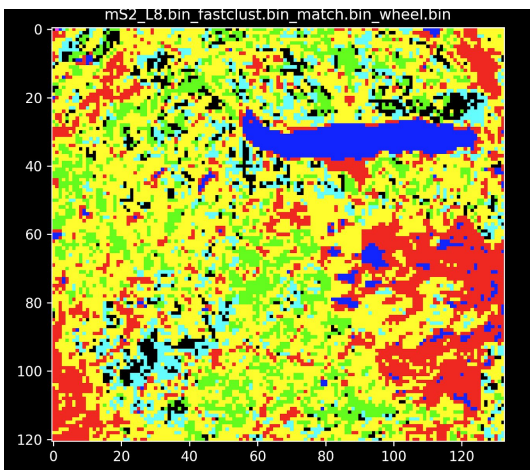


Elbow Method

The elbow method is a heuristic to determine ‘k’, an ideal number of clusters to study, by examining cluster dispersion measurements such as WCSS i.e. the variance. This involves increasing the number of clusters until it’s of marginal benefit to add more. Although the resulting number of clusters is hard to relate to the number of fuel types that can be discriminated from images, it’s a useful starting point for studying which land cover types we may distinguish. By inspection, seven was determined a reasonably appropriate number for cluster visualization.

HAC and Comparison

The images below are obtained by running K-Means and Hierarchical Agglomerative Clustering (HAC) all with seven resulting clusters as estimated with the elbow heuristic, re-coding the outputted label maps [20], and mapping the resulting labels onto the colour wheel [21]. Interestingly, HAC seems to have provided results which vary more clearly with choice of sensor.

	KMeans	HAC
Sentinel-2	 <p>Classification map for Sentinel-2 using KMeans. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mS2.bin_kmeans.bin_wheel.bin'.</p>	 <p>Classification map for Sentinel-2 using HAC. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mS2.bin_fastclust.bin_match.bin_wheel.bin'.</p>
Landsat-8	 <p>Classification map for Landsat-8 using KMeans. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mL8.bin_kmeans.bin_match.bin_wheel.bin'.</p>	 <p>Classification map for Landsat-8 using HAC. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mL8.bin_fastclust.bin_match.bin_wheel.bin'.</p>
Sentinel-2, Landsat-8 fused	 <p>Classification map for Sentinel-2, Landsat-8 fused using KMeans. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mS2_L8.bin_kmeans.bin_match.bin_wheel.bin'.</p>	 <p>Classification map for Sentinel-2, Landsat-8 fused using HAC. The map shows a landscape with a blue river feature. The axes range from 0 to 120 on both the x and y dimensions. The title is 'mS2_L8.bin_fastclust.bin_match.bin_wheel.bin'.</p>

The relative accuracy of the results remains to be confirmed by comparison with ground reference data. In the future, labels should be added to identify these clusters as fuel types, which can then be used to reason about how fast a fire will spread.

That K-Means clustering is single-level only, is a limitation that affects its potential accuracy. Moreover, k-means selects centroids randomly, so the final solution can be sensitive to the initial random selection of cluster centers. Despite using the `init='k-means++'` option to reduce the effect of this sensitivity, initialization sensitivity can entail a limitation on the potential shapes of clusters K-Means can find. Hierarchical clustering on the other hand, is a deterministic method that builds a cluster hierarchy using a tree structure called a dendrogram. Since HAC generates results unaffected by initialization sensitivity, and uses a greater number of distance comparisons and substantially more data to represent the structure of clusters, it's considered a more detailed method with potential for much higher accuracy.

Summary and Recommendations for Future Work

Further work is required to assess potential improvements to classification accuracy with respect to the ground reference data available. Moreover clustering is already a well-established method for distinguishing among forest land-cover types using satellite imagery, with precedents including national-level programs such as EOSD [22] which developed national forest land-cover maps using the K-Means algorithm. Moreover, similar projects e.g. [23] operating at larger scales will be important to refer to while developing regionally-validated high-resolution products.

For best results object-based (polygonal) approaches for improving Signal to Noise Ratio (SNR) should be implemented. Finally, more types of Remote Sensing (RS) data and combinations thereof must be explored, along with more sophisticated or detailed methods such as time-series analysis.

References

- [1] <https://www.cbc.ca/news/canada/british-columbia/state-emergency-bc-wildfires-1.4803546>
- [2] <https://cwffis.cfs.nrcan.gc.ca/background/summary/fdr>
- [3] <https://cwffis.cfs.nrcan.gc.ca/background/summary/fbp>
- [4] Daniel D.B. Perrakis, George Eade, Dana Hicks, "British Columbia Wildfire Fuel Typing and Fuel Type Layer Description"
- [5] https://en.wikipedia.org/wiki/2017_British_Columbia_wildfires
- [6] <https://globalnews.ca/news/4390618/bc-wildfire-new-normal/>
- [7] <https://www.cbc.ca/news/canada/british-columbia/forest-fires-smoke-mental-health-1.4792195>
- [8] <https://globalnews.ca/news/4406900/forest-fires-forest-management/>
- [9] R.S. McAlpine, B.J. Stocks, C.E. Van Wagner, B.D. Lawson, M.E. Alexander, T.J. Lynham, Forest Fire Behavior Research in Canada, Proceedings of International Conference of Forest Fire Research, 1990, A.02, 1-12
- [10] Tomàs Artés, Ana Cortés, Tomàs Margalef, Large Forest Fire Spread Prediction: Data and Computational Science, Procedia Computer Science, Volume 80, 2016, Pages 909-918.
- [11] Dario Rodriguez-Aseretto, Daniele de Rigo, Margherita Di Leo, Ana Cortés, Jesús San-Miguel-Ayanz, A Data-driven Model for Large Wildfire Behaviour Prediction in Europe, Procedia Computer Science, Volume 18, 2013, Pages 1861-1870.
- [12] Stojanova, Daniela & Panov, Pance & Kobler, Andrej & Džeroski, Sašo & Taškova, Katerina. (2006). Learning to predict forest fires with different data mining techniques.
- [13] <https://mapr.com/blog/predicting-forest-fires-with-spark-machine-learning/>
- [14] AK Wijayanto, O Sani, ND Kartika, Y Herdiyeni, Classification Model for Forest Fire Hotspot Occurrences Prediction Using ANFIS Algorithm, IOP Conference Series: Earth and Environmental Science, 54 (2017) 012059.
- [15] <https://www.kaggle.com/surya635/forest-fire-prediction>
- [16] <https://step.esa.int/main/toolboxes/snap/>
- [17] <https://gdal.org/programs/gdalwarp.html>
- [18] https://gdal.org/programs/gdal_translate.html
- [19] <https://qgis.org/en/site/>
- [20] https://github.com/bcgov/bcws-psu-research/blob/master/cpp/class_match_onto.cpp
- [21] https://github.com/bcgov/bcws-psu-research/blob/master/cpp/class_wheel.cpp
- [22] Satellite land cover mapping of Canada's forests: the EOSD land cover project. 2008. Wulder, M.A.; Cranny, M.M.; Hall, R.J.; Luther, J.E.; Beaudoin, A.; White, J.C.; Goodenough, D.G.; Dechka, J.A. Pages 21-30 (Chapter 3) in J.C. Campbell, K.B. Jones, J.H. Smith, and M.T. Koeppe, editors. North America Land Cover Summit. American Association of Geographers, Washington, DC, USA.
- [23] <https://www.esa-landcover-cci.org/?q=project%20team>

Appendix -- Source Code

<https://github.com/franarama/satellite-clustering>