Application of machine learning techniques for well pad identification in the Bakken oil field

Philip G. Brodrick and Jacob G. Englander
December 12, 2014

1 Introduction

There has been increased scrutiny in understanding the anthropogenic sources for methane emissions due to methane's potency as a greenhouse gas [2]. There are two approaches for studying of methane emissions, emissions inventories (or 'bottom up' studies), and remote measurements of methane fluxes (or 'top down' studies). Emissions inventories calculate the leakage rates for fugitive methane over a given area as:

$$E = \sum_{j}^{M} \sum_{i}^{N} A_{i,j} \times EF_i \tag{1}$$

where E is the emissions rate, $A_{i,j}$ is the activity (or the number of pieces of equipment of type i) at well pad j, EF_i is the emissions factor for equipment type i, N is the set of possible equipment types, and M is the set of wells in the location of interest. The inner sum provides the emissions rate for a specific well pad.

One of the significant problems with the 'bottom up' emissions inventory approach is that of scale. Even when very detailed measurements of the equipment and stages of operations are conducted, doing so comes at the price at looking at a small and not necessarily representative sample of the wells and operators. This is the most significant critique of Allen et al. which found lower leakage numbers when compared to other emissions inventories [3]. In order to better rectify the discrepancies between the 'top down' and 'bottom up' studies, two simultaneous development need to occur. There needs to be more proper accounting for the number of devices and pieces of equipment present on a particular site and there needs to be better sampling and estimation of emissions factors at the device level. The intention of this project, which is to apply machine learning techniques on high resolution spatial imagery to identify well pads, is meant to serve as an initial step in this process.

While there is uncertainty in estimating both the activity counts and the emissions factors, there are also significant disagreements in simply estimating the number of well pads in a given location. For example the SI to Brandt et al. report a 5x difference in the number of reported well completions in the United States between two different data sources (for 2010 the EPA reported 4871 wells, while IHS reported 18542) [2]. The purpose of this project is to identify well pads using high resolution spatial imagery, which is a first step towards estimating activity counts over an entire production field.

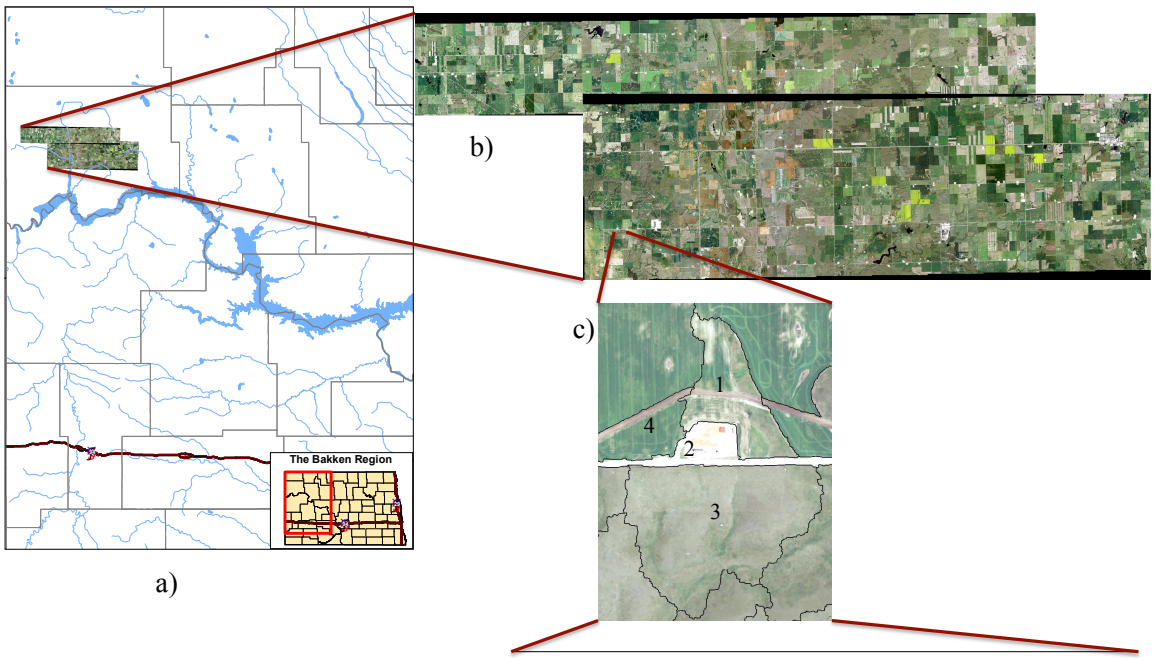
Machine learning classifiers are widely utilized for remote-sensing based classifications, and their application has been of particular interest in the literature of classification of high resolution spatial imagery. Common classifiers utilized in the literature include Naïve Bayes, k-Nearest Neighbor, Neural Network, Random Forests, and SVM. Mountrakis et al., in a review of the use of SVM for remote sensing applications, describe the ability for this classifier to better balance the bias-variance tradeoff as well as it ability to use small training sets. They also demonstrate that SVMs are disproportionally used in applications with high resolution imagery [5].

2 Methods and data

2.1 Data and preprocessing

The region of study for this project is the fast developing Bakken oil field in Western North Dakota. This area is of particular interest as oil production has increased from ≈2000 bbl/day in 2005 to $\approx 1,000,000$ bbl/day in 2014. Specifically, the study area utilized was a ≈1150 km^2 area of 1m^2 pixel aerial photographs acquired in August 2012

from the National Aerial Imagery Program at the USDA. A map of the study area within the context of the Bakken can be found in Figure 1.



d)

Segment	NDVI mean	Brightness	Area (pixels)	Roundness	GLCM Entropy
1	0.10	151.01	51652	1.16	6.91
2	-0.08	196.97	31473	2.05	6.67
3	0.06	143.45	137819	0.71	5.60
4	0.24	134.44	27498	0.70	6.11

Figure 1: Schematic for input data: a) project study area within the context of the Bakken region and North Dakota, b) Close up image of study area, c) example of segmented image, d) example of data exported from segmentation

.

For most remote sensing applications, analysis and classification is performed at the pixel level. However, this has been found to be not particularly effective for high resolution imagery due to the the heterogeneity of adjacent pixels [1]. The widely accepted technique for extracting data out of high resolution imagery has been to aggregate pixels into image objects (or segments) [6]. In order to isolate these image objects, the *eCognition* software platform was utilized using the multi-resolution segmentation function with the following parameters which were utilized to best identify well pads: scale (which roughly corresponds to size) = 300, shape (or the influence of color on segment) =0.6, and compactness (which corresponds to the ability of segments to have varying size) = 0.9. The result of this process divided the image into 9485 segments, which serve as the input rows for the classifier. This set was later trimmed down to 7339 segments to avoid edge effects around the perimeter of the region of interest. A visual example of the result of segmentation can be found in Figure 1. Each of the segments has 175 features (or columns) derived from three overarching categories: spectral reflectance (RGB and near IR) and brightness, segment geometry (shape and size), and texture (which is based on the grey-level co-occurance matrix from Haralick et al.) [4].

A matrix representation with the segments as the rows and the features as the columns is used as the input data for the classifier. The output is a binary indicator for whether the test segment is or is not a well pad. The well pads were verified with GIS layers of known oil and gas facilities in North Dakota [7], and further refined by hand. A schematic of the model function can be found in Figure 2.

2.2 Models

To perform the classification we used the following models and techniques which are commonly used in the literature, utilizing default MATLAB functions:

1. Support Vector Machine (Gaussian Kernel)

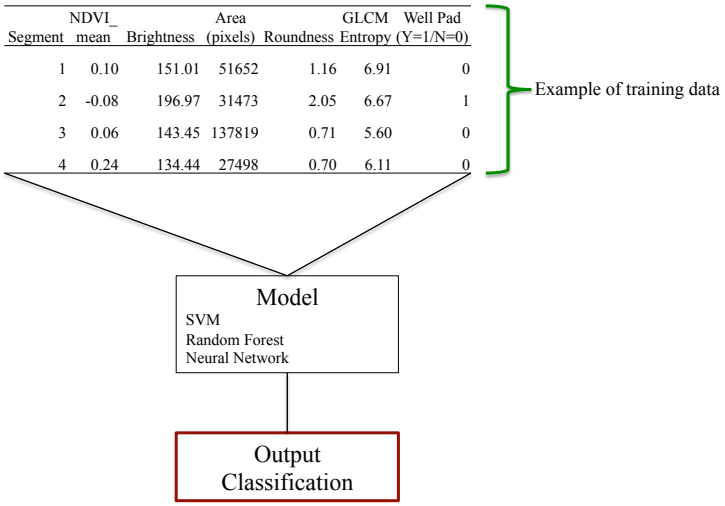


Figure 2: Schematic of model function

2. Random Forest
3. Neural Network
4. Feature Evaluation: Principal Component Analysis

3 Results

Well pads are geographically scarce throughout the Bakken oil field, and since size and shape are important segmentation characteristics utilized in our application of *eCognition*, are also scarce amongst the generated segments. Of the 7339 segments used (post-processing), only 195 are well pads (2.66%). Consequently, a classification could have a 95% success rate while successfully identifying every well pad in the area, and still have falsely identified almost as many wells as actually exist. In the context of using this classification to estimate fugitive emissions, this degree of error is well above an acceptable limit. Figure 3 demonstrates the scarcity of well pads in the data set, highlighting the importance of highly accurate classification.

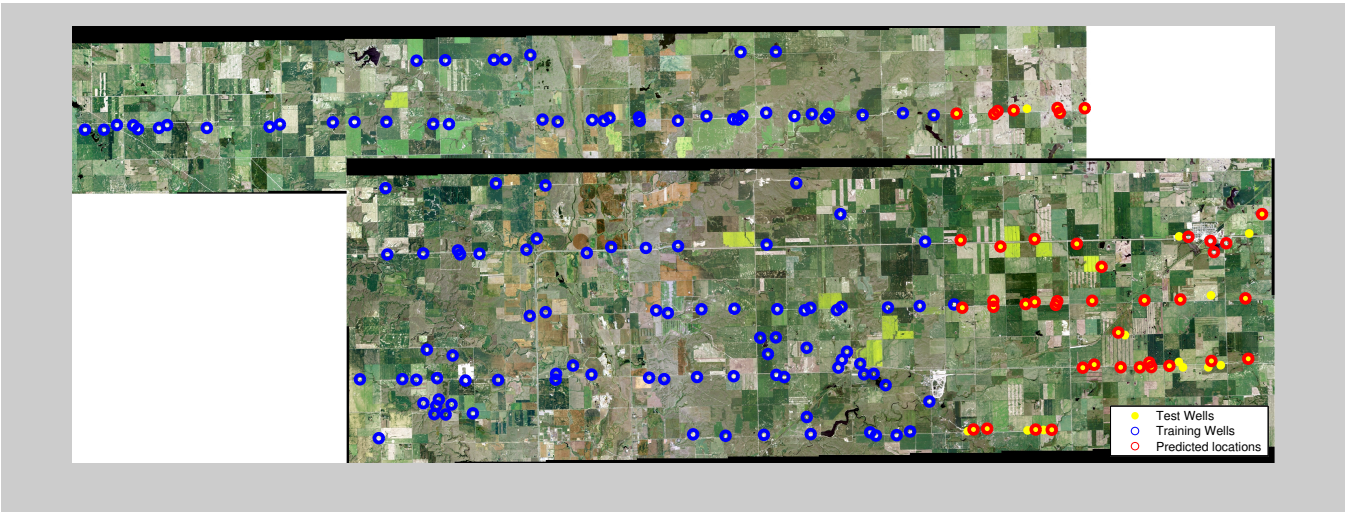


Figure 3: Demonstration of well pad classification using a Random Forest with 75% of the data as a training set.

To better analyze classifications results (such as shown in Figure 3) we define both a true-positive and false-positive

rate. The first, true-positive rate (tpr), is the commonly used expression:

$$tpr = \frac{\sum_i^m 1\{y_i = y_i' | y_i' = 1\}}{\sum_i^m y_i} \tag{2}$$

where y_i is the true classification, y_i' is the hypothesized classification, and m is the size of the testing set. The second measure, false-positive rate (fpr) is defined in this context slightly more conservative than is often seen, as:

$$fpr = \frac{\sum_i^m 1\{1 - (y_i = y_i') | y_i' = 1\}}{\sum_i^m y_i}. \tag{3}$$

A receiver operating characteristic (ROC) curve is examined in order to explore the trade-off between tpr and fpr in different algorithms. To generate the curve, 75% of the available data is used as the training set, and 25% is used as the testing set. Each algorithm is applied ten times to (the same set of) random permutations of the data, and the ROC curves are shown in Figure 4.

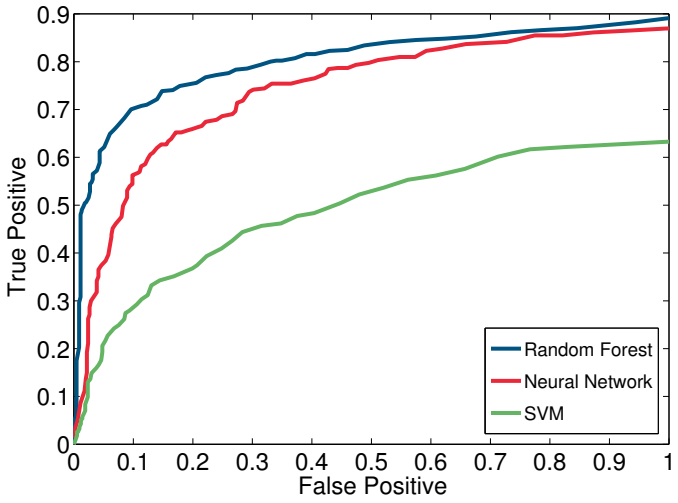


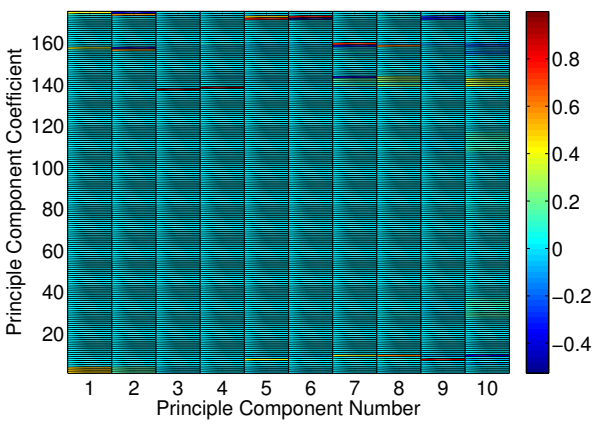
Figure 4: ROC curve showing the trade-off between true-positive and false-positive rates.

.

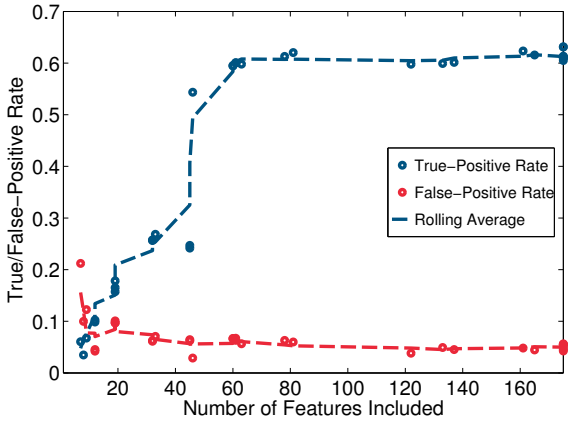
The data set required is small enough that the time required for classification, even for the entire Bakken oil field, is not prohibitively expensive. However, the computation time required to generate the features is; the data set used in the limited area surveyed in this work took roughly 40 hours to generate. In order to determine the order in which to include features, a principle component analysis (PCA) is performed first. Examining the principle components (the first 10 of which are shown in Figure 5(a)) it is clear that each component is heavily dominated be a small subset of features. An arbitrary (and conservative) threshold of 0.1 is set, and (in order) the features with coefficients whose absolute value exceeds the threshold are selected. This roughly ordered set of features is then used to explore the effect of reducing the feature space, by classifying the same random permutations of the data set used above with the best performing algorithm, the Random Forest, with an increasing number of features. The resulting true and false-positive rates are shown in Figure 5(b).

4 Discussion and Conclusions

The results of this work demonstrate the viability of utilizing machine learning classification for well pad identification in the Bakken. As demonstrated in Figure 4, the overall accuracy of well pad identification is highly dependent on the acceptable rate of false positives. Though both the Neural Network and the Random Forest performed well overall, the Random Forest benefits the most with a give allowance of false positive error. The ROC curve shows a change in the relative increased true positive rate as a function of false positive rate at an approximate 10% false-positive rate, indicating that this point results in the best model performance for the intended application.



(a) Visualization of the first 10 principle components.



(b) Demonstration of the affect of data reduction on true-positive and false-positive rates.

Additionally, the result of the PCA allows for a dramatic reduction of the feature set, and consequently a significant reduction in computation effort. Figure 5(b) demonstrates that after the inclusion of ≈ 65 featuers, the accuracy of the classifiers do not improve considerably. We estimate that this could lead to a reduction in computation time of 50-65%.

In the process of this analysis, we added ≈ 725 km^2 of area to an initial area of ≈ 425 km^2. This additional data was meant to provide an analysis on the effect of classification accuracy as the training set became larger, however only a modest increase in classification accuracy was observed. The lack of improvement likely resulted from using the same segmentation parameters between data sets, which introduced some segmentation errors upon examining the segments. However, the computation time required for segmentation was prohibitive to experimenting widely on the effect of segmentation parameter selection. With the analysis of feature space reduction performed in this work, further analysis in this area can be performed. Additionally, the strict accuracy constraints for the calculation (low acceptable false positive rate) likely contributed to the modest gains in accuracy.

Future work will utilize the classification results of this work to isolate imagery in the immediate vicinity of the well pads in order to perform a second segmentation step in order to identify surface equipment such as tanks, pumpjacks, and compressors. This might test the limit of the spatial resolution of these images and as such higher resolution imagery sources will also be explored.

References

1. T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):2–16, Jan. 2010. doi: 10.1016/j.isprsjprs.2009.06.004.

2. A. R. Brandt, G. A. Heath, E. A. Kort, F. O'Sullivan, G. Petron, S. M. Jordaan, P. Tans, J. Wilcox, A. M. Gopstein, D. Arent, S. Wofsy, N. J. Brown, R. Bradley, G. D. Stucky, D. Eardley, and R. Harriss. Methane Leaks from North American Natural Gas Systems. *Science*, 343 (6172):733–735, Feb. 2014. doi: 10.1126/science.1247045.

3. EPA. Greenhouse Gas Reporting Program, subpart W, Petroleum and Natural Gas Systems. Technical report, Environmental Protection Agency, 2013. URL `http://www.epa.gov/ghgreporting/reporters/subpart/w-reported.html`.

4. R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, Nov. 1973. doi: 10.1109/TSMC.1973.4309314.

5. G. Mountrakis, J. Im, and C. Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, May 2011. doi: 10.1016/j.isprsjprs.2010.11.001.

6. S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing of Environment*, 115(5):1145–1161, May 2011. doi: 10.1016/j.rse.2010.12.017.

7. North Dakota Department of Mineral Resources. North Dakota Drilling and Production Statistics, 2013. URL `https://www.dmr.nd.gov/oilgas/stats/statisticsvw.asp`.