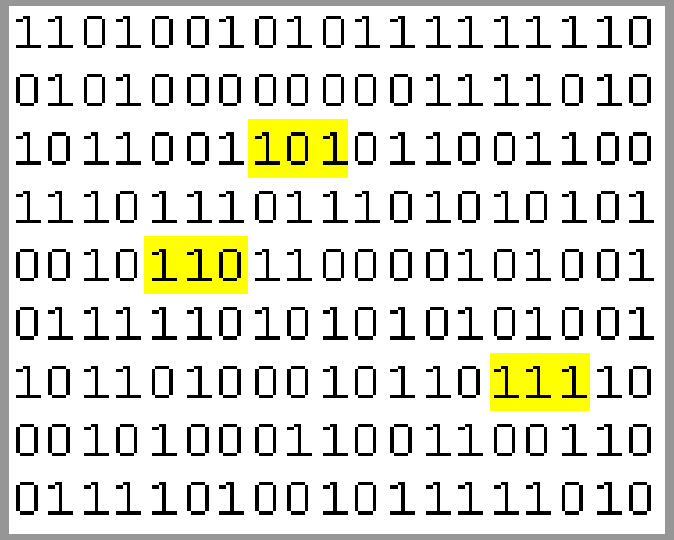
Automated Determination of Optimum Data Subsets for Robust Classification (overview)



Robert K. McConnell, PhD
WAY-2C
Arlington, Massachusetts
rkm@way2c.com
http://www.way2c.com

- Many fields of endeavor rely upon knowing the past and/or present state upon which to base decisions. Often this state is not completely known and must be inferred from some small subset of the available data.
- Analysts apply a variety of correlation methods to search for a relevant data subset from which to infer the state. For ease in computation, these methods usually implicitly or explicitly assume some “parametric” distribution (such as Gaussian) for the classes and data.

It is well known that:

- Parametric distributions are often poor matches for the actual data thus causing invalid conclusions regarding the state they predict.
- The traditional correlation functions used have been a necessary compromise to keep computation times reasonable.
- The traditional correlation methods are unable to determine optimum combinations of different data subsets to predict the state.

This presentation illustrates a newly patented* method of determining an optimum subset of the available data for classifying the state.

While the examples shown are images, the method is applicable to virtually any type of “big data” where one needs to find a subset whose combination best correlates with states of interest.

*U.S. Patent 8,918,347, others pending.

The new method, based on information theory:

- Does not rely on any invalid assumption of parametric distributions or associated traditional correlation methods.
- Can find optimum combinations of data subsets to predict the state.
- Typically finds optimum subset combination from hundreds of data sets, each with hundreds of thousands of individual data points, in a couple of minutes on a PC.
- Has been successfully applied to a variety of image and time-series data sets.

WAY-2C

Outline

Using the simple analogy of gray-scale, color and hyperspectral images, with user-defined classes of interest, we shown steps in:

- Finding a most relevant data subset upon which to base the classification.
- Using that most relevant subset to perform maximum likelihood classification.

The Analogy

Data consists of:

- Multiple separate views of the same scene viewed through a different filter. Each view is referred to as a “band”.

Objectives are:

- Find a most relevant data subset (band combination) upon which to differentiate several known classes.
- Using that most relevant subset, perform the classification for the test scene and other scenes containing the same classes.

Gray-scale, Color and Hyperspectral Images

- Gray-scale: one band.
- Color: three combined bands (e.g red, green and blue).
- Hyperspectral: tens to hundreds of bands.

More information for class identification in three-band (color) images than gray-scale images

Courtesy Marathon Sports, Inc.

WAY-2C

Hyperspectral Imagery: an example of data overload

- Single scene imaged simultaneously through tens to hundreds of different filters (bands).
- Immense amounts of data, some of it relevant for determination of user's defined classes of interest, much of it not.
- Problem is to determine relevant data subset to differentiate user-defined classes*.

*McConnell, R.K., Supervised method for optimum hyperspectral band selection, in Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIX, Sylvia S. Shen; Paul E. Lewis, Editors, Proceedings of SPIE Vol. 8743 (SPIE, Bellingham, WA 2013).

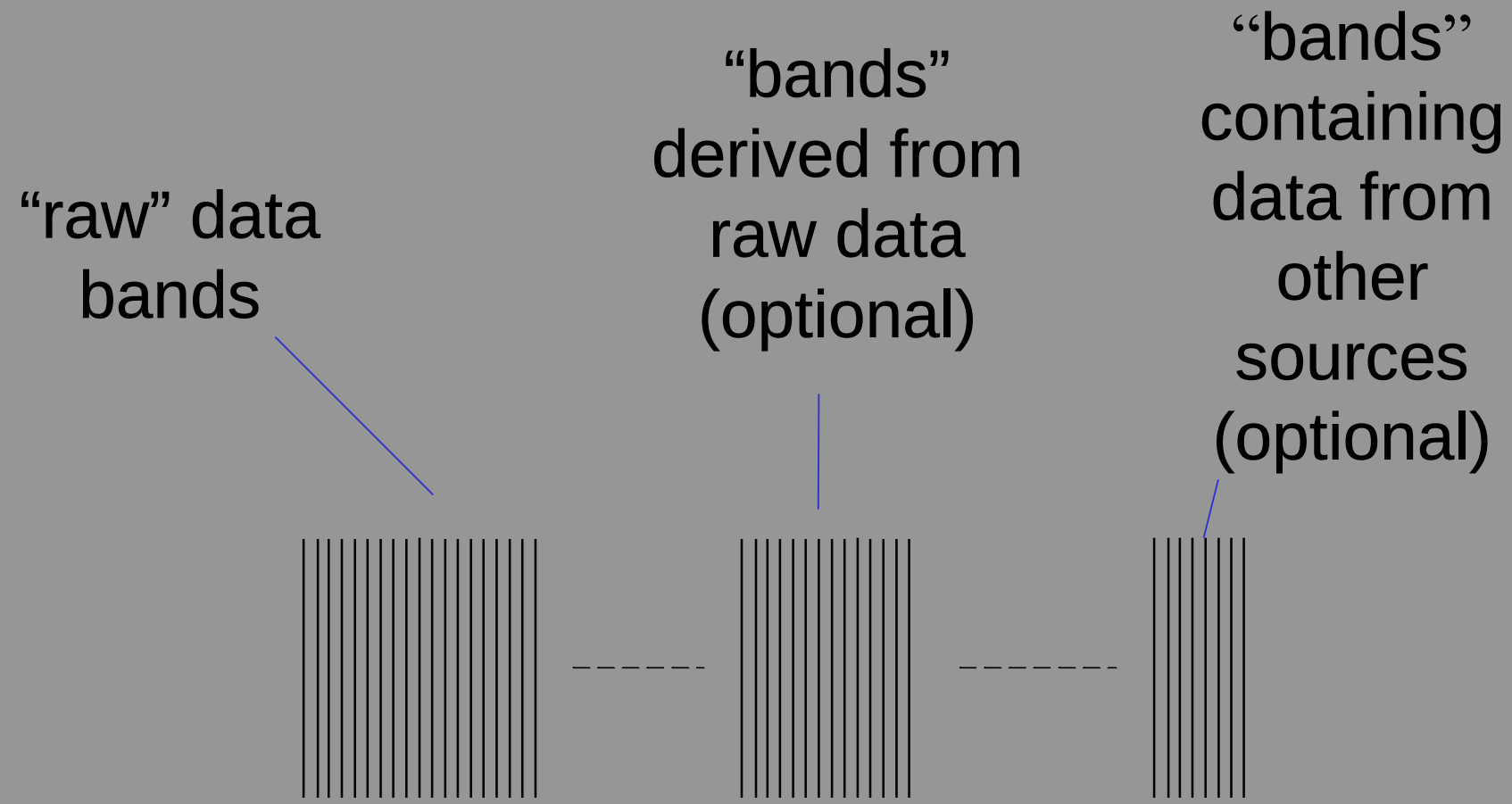
Hyperspectral Image
“Data Cube”

Graphic representation.

Nicholas M. Short, Sr.
NASA

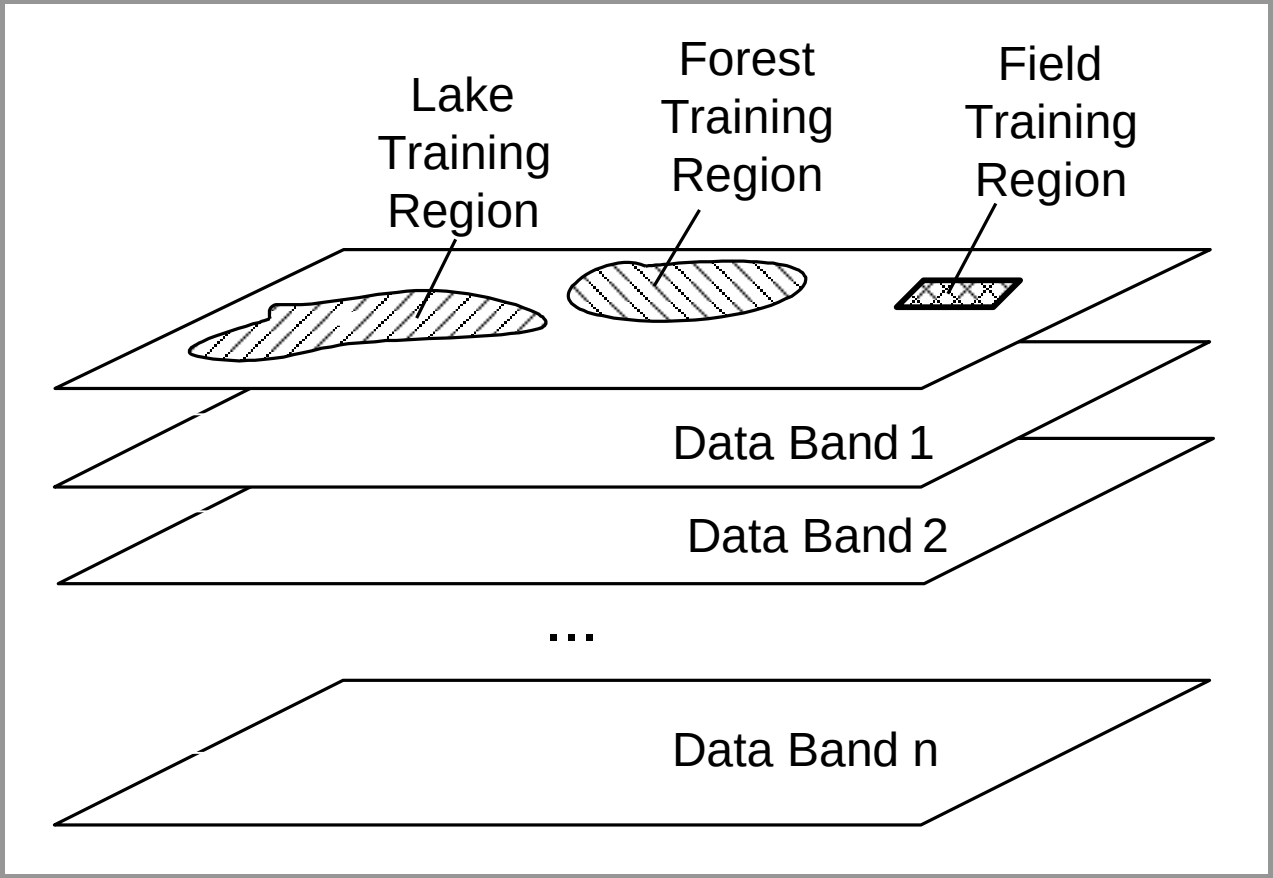
raw bands and more

(to be searched for relevance)

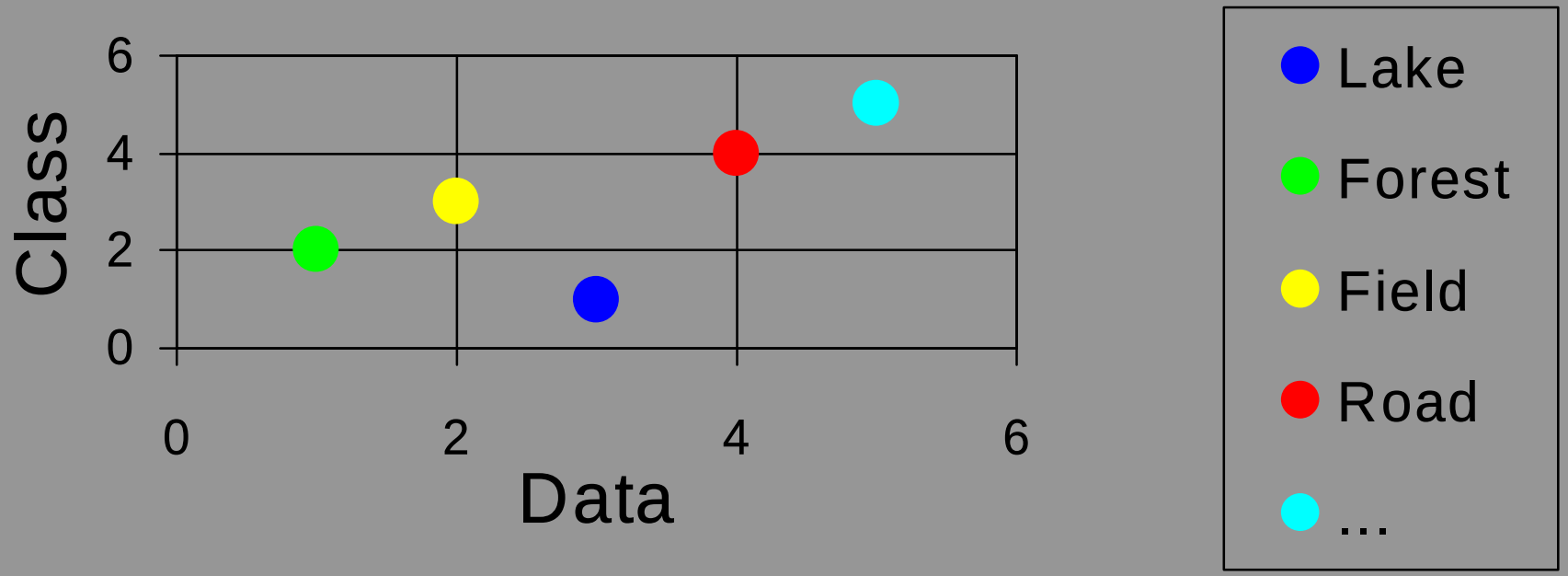


WAY-2C

Given some class training data ...



... ideally find data subset that correlates one-to-one with class



WAY-2C

Finding most relevant data subset

is treated strictly as a black box operation to automatically discard all data except that most relevant for differentiating the specified classes of interest.

Data is discarded because the information it contains is either irrelevant or redundant.

“Green Chip” Example

Data set is:

- Scene, showing seven “paint chips” glued to a background, viewed through 80 different filters. Each view is referred to as a “band”. Some bands have more relevant information than others

Objectives are:

- Find a most relevant data subset (band combination and bits within those bands) to differentiate the 8 classes.
- Using that most relevant subset, perform a maximum likelihood classification for the training scene and other scenes containing the same classes.

“True Color” Image From Cube

	Band	nm
R	40	642
G	25	550
B	10	458

“Green Chips”

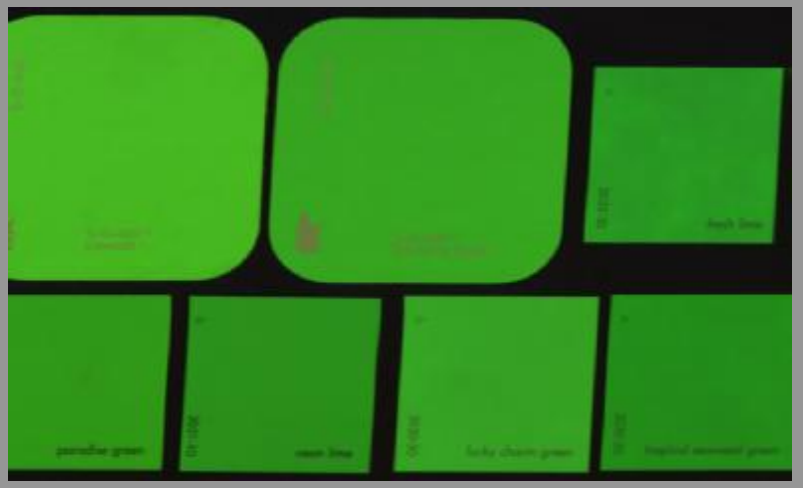
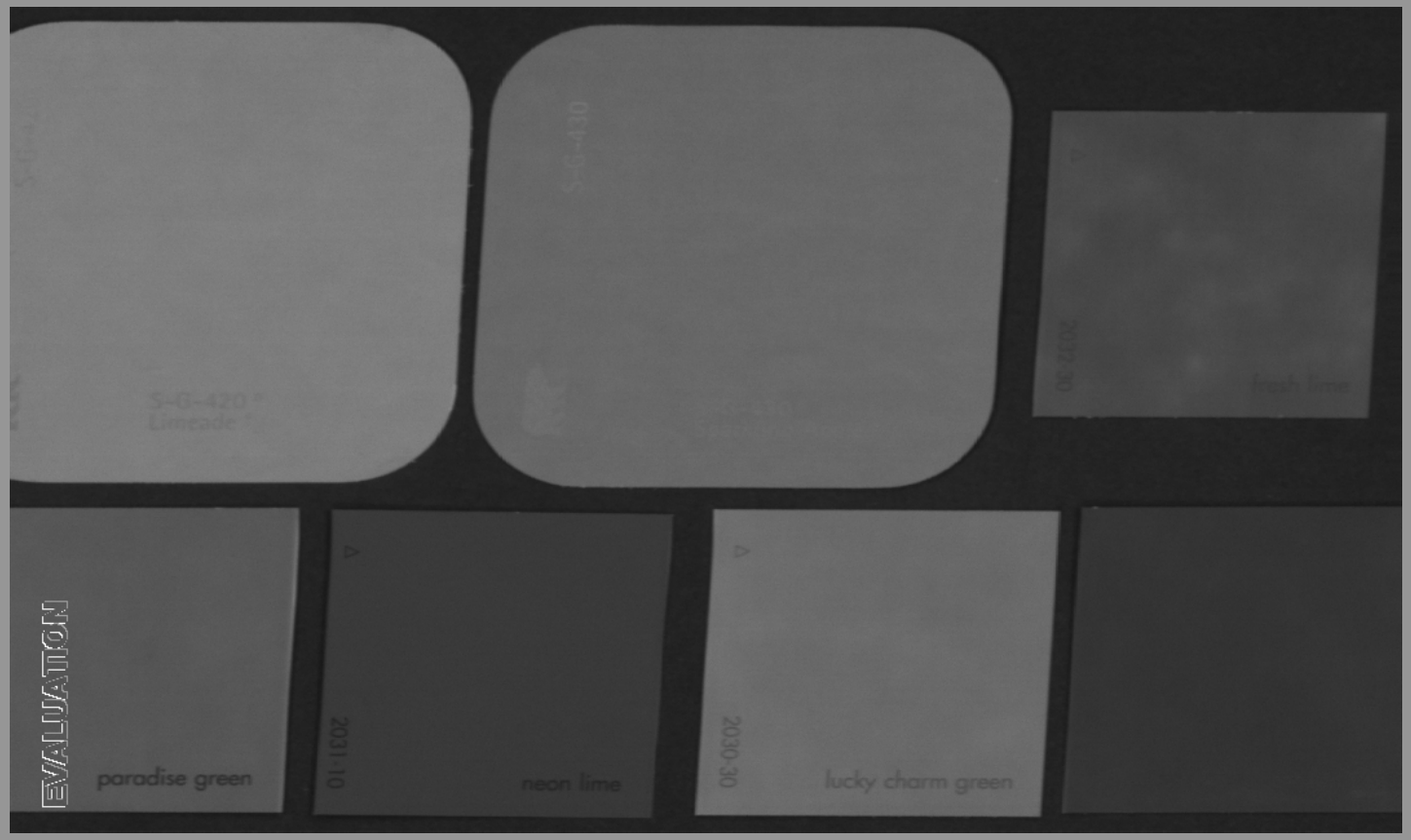


Image Courtesy Resonon, Inc.

When bands from the middle of the red, green and blue portions of the spectrum are combined to form a digital image, the green chips are almost indistinguishable.

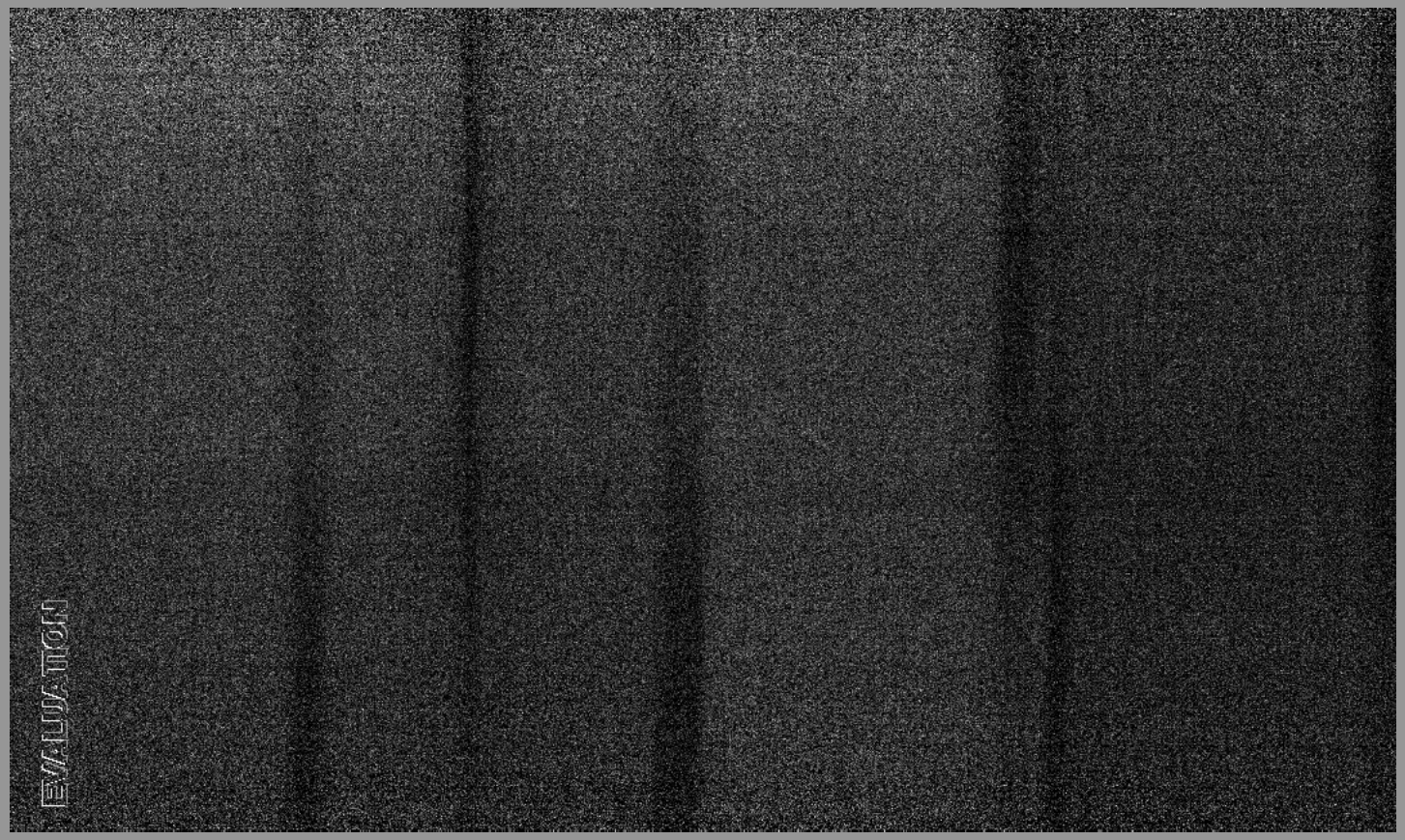
Band 60: much relevant data



WAY-2C

Base Image Courtesy Resonon, Inc.

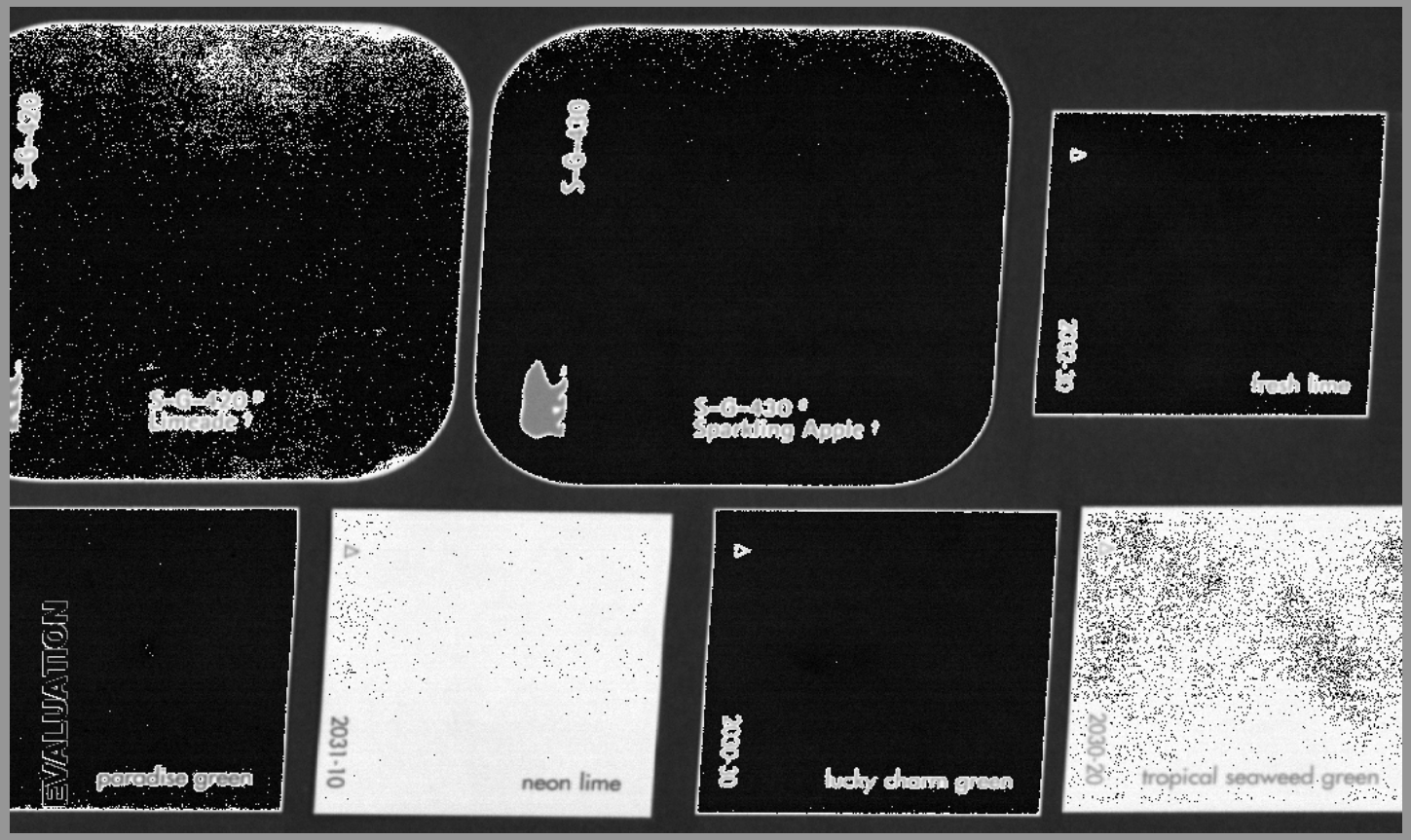
Band 0: no relevant data



WAY-2C

Base Image Courtesy Resonon, Inc.

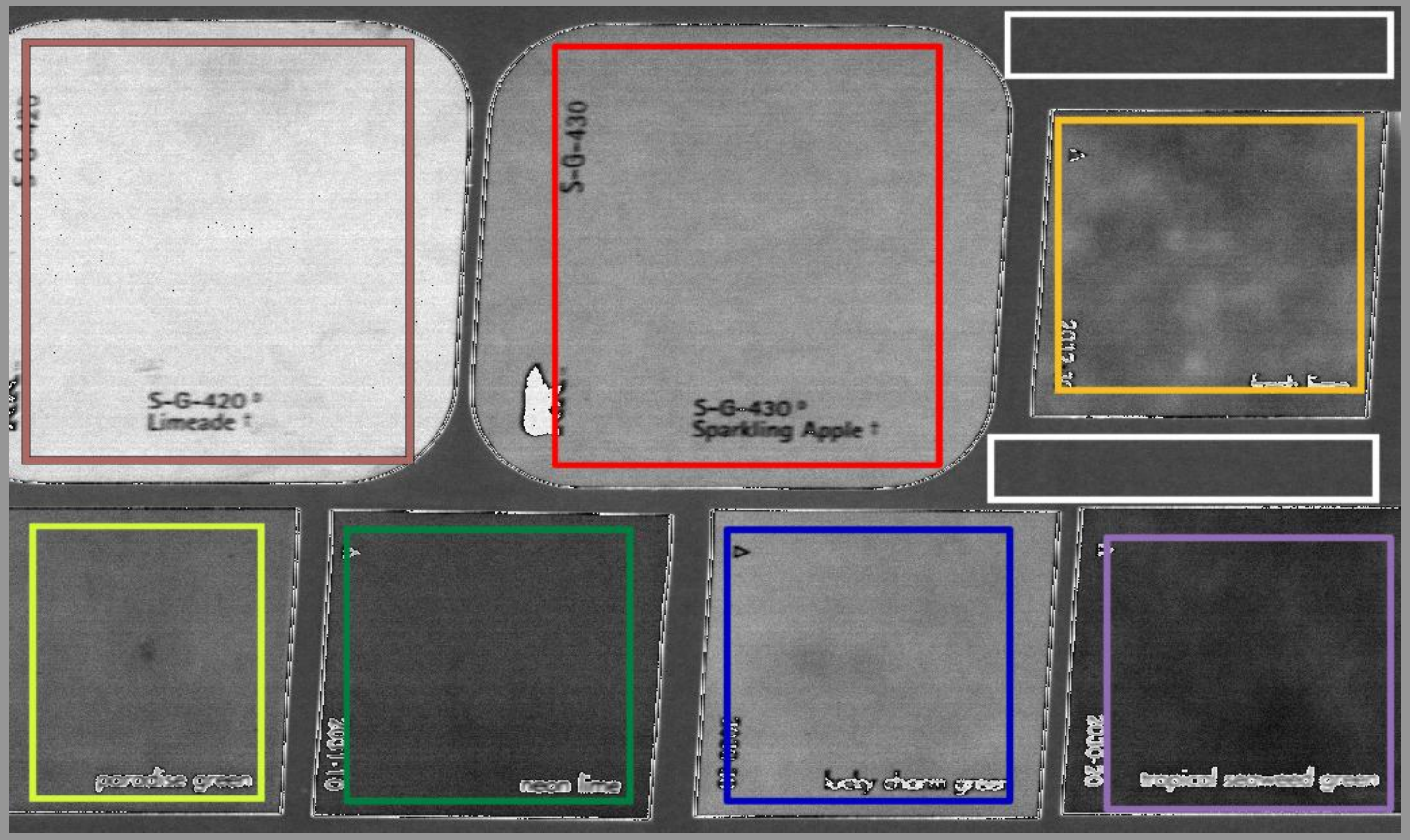
Band 79: some relevant data



WAY-2C

Base Image Courtesy Resonon, Inc.

“Green Chip” Class Training Regions



WAY-2C

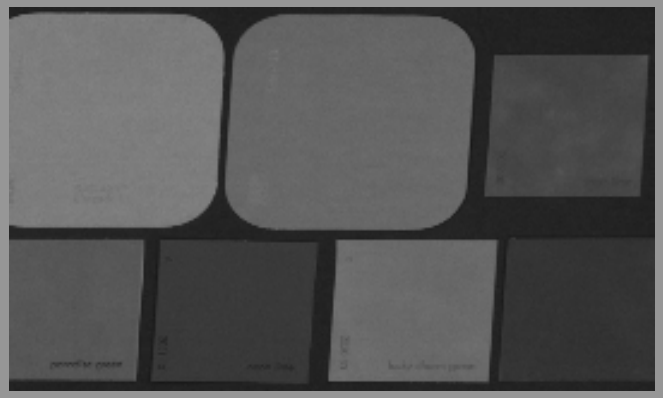
Base Image Courtesy Resonon, Inc.

Most relevant data band combination

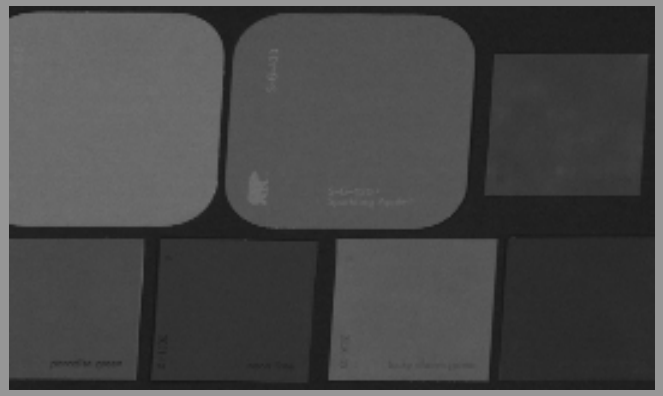
Optimum band combination to differentiate the 8 classes reliably as determined using methods described in U.S. Patent 8,918,347.

Band

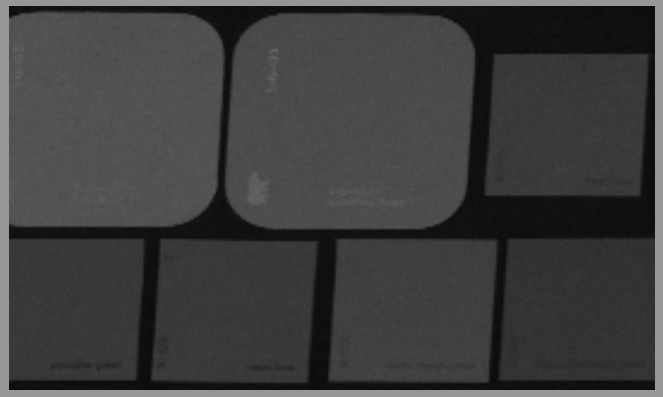
60



57



17



WAY-2C

“Color” image combining most relevant data band combination

Plane	Band	nm
R	60	764
G	57	745
B	17	501

Chips are now somewhat distinguishable.

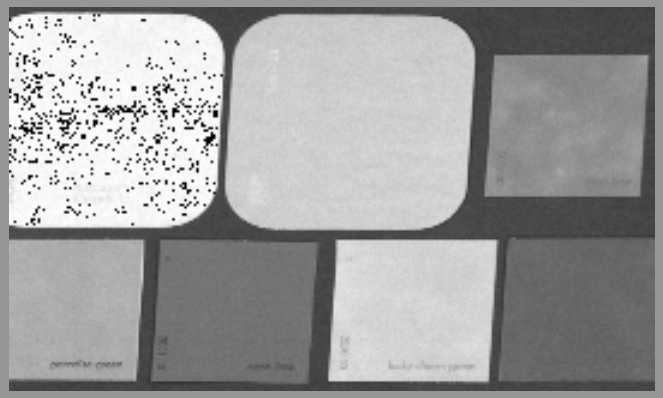
WAY-2C

Most relevant data subset combination

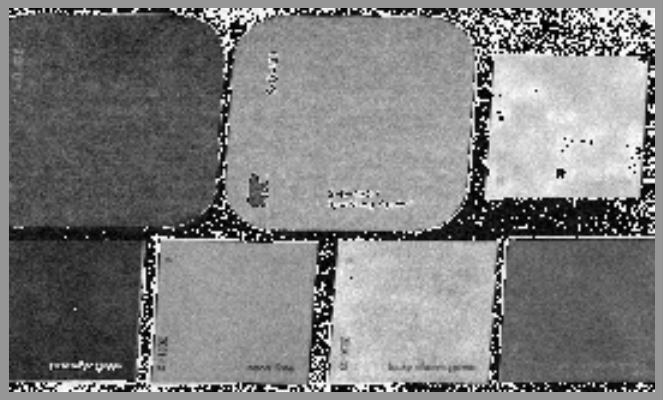
Optimum band data subset combination to differentiate the 8 classes reliably as determined using methods described in U.S. Patent 8,918,347.

Band

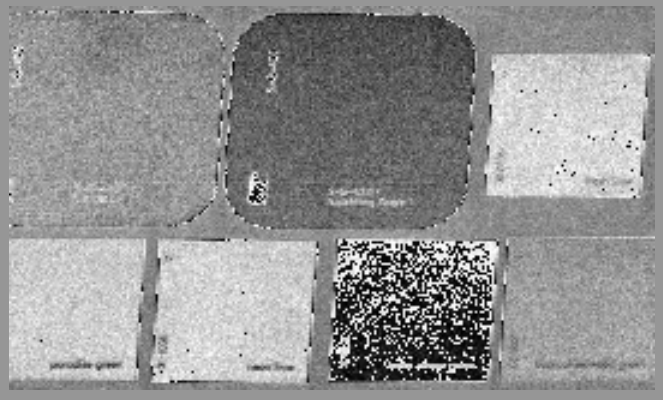
60



57



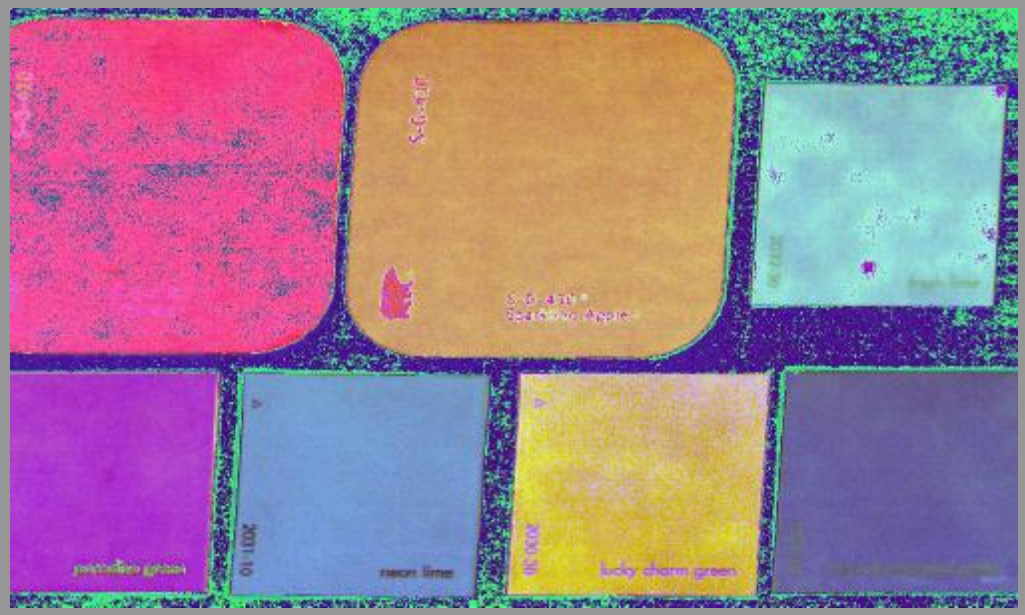
17



WAY-2C

“Color” image combining most relevant data subset

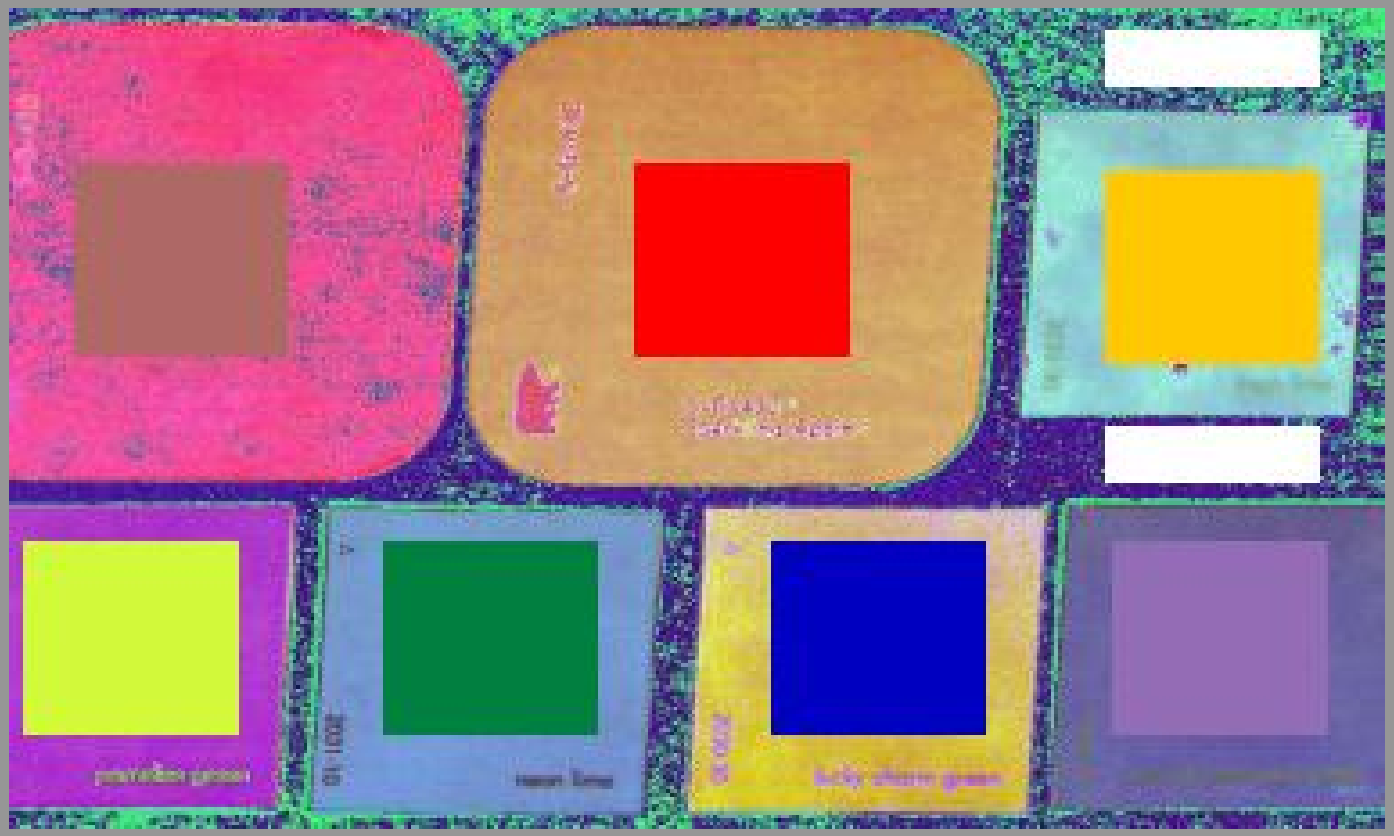
Plane	Band	nm
R	60	764
G	57	745
B	17	501



Chips are now easily distinguishable.

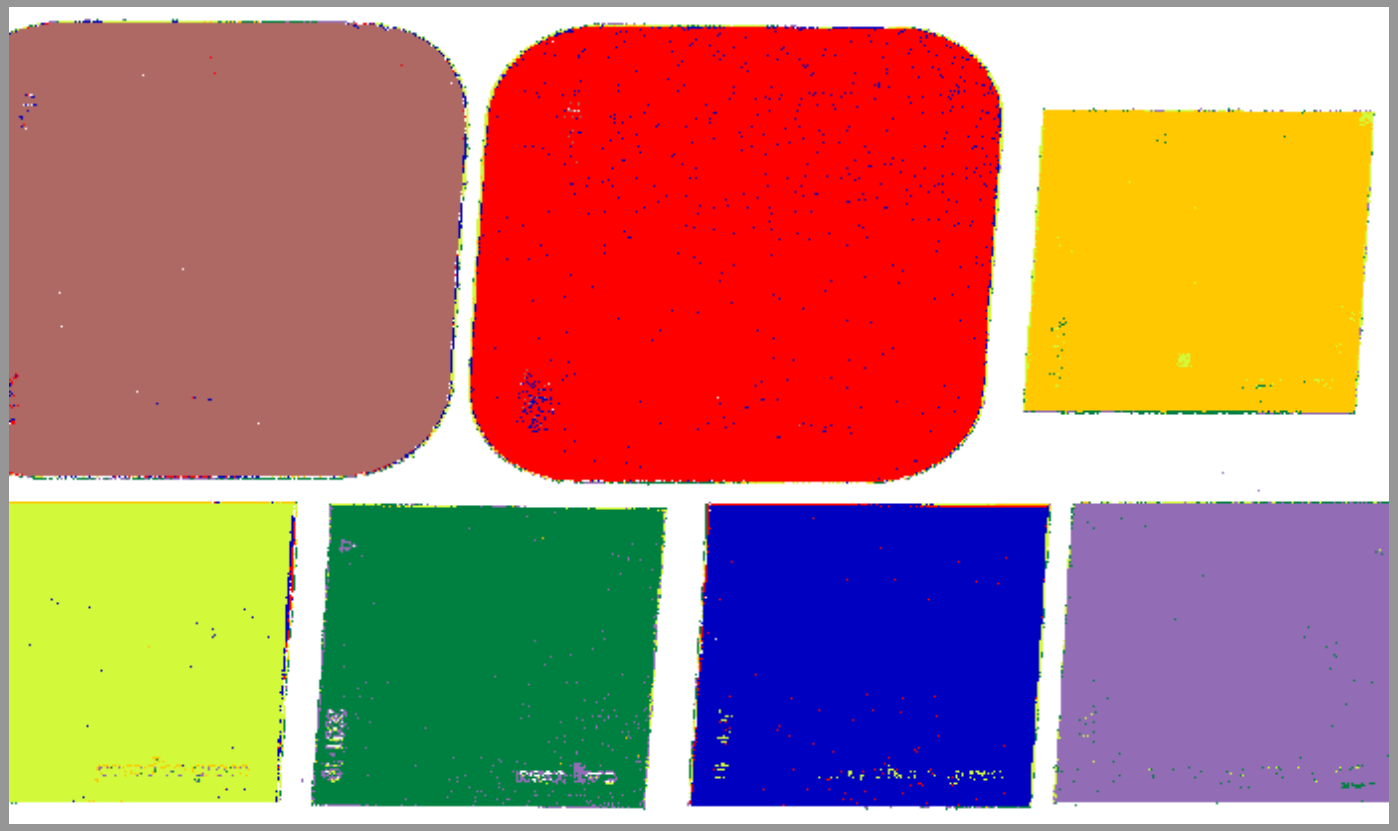
WAY-2C

Maximum likelihood classification based on most relevant band triplet combination.



WAY-2C

Point-by-point classification based on most relevant band triplet combination.



WAY-2C

A more complex example: optimum band set for terrain class differentiation

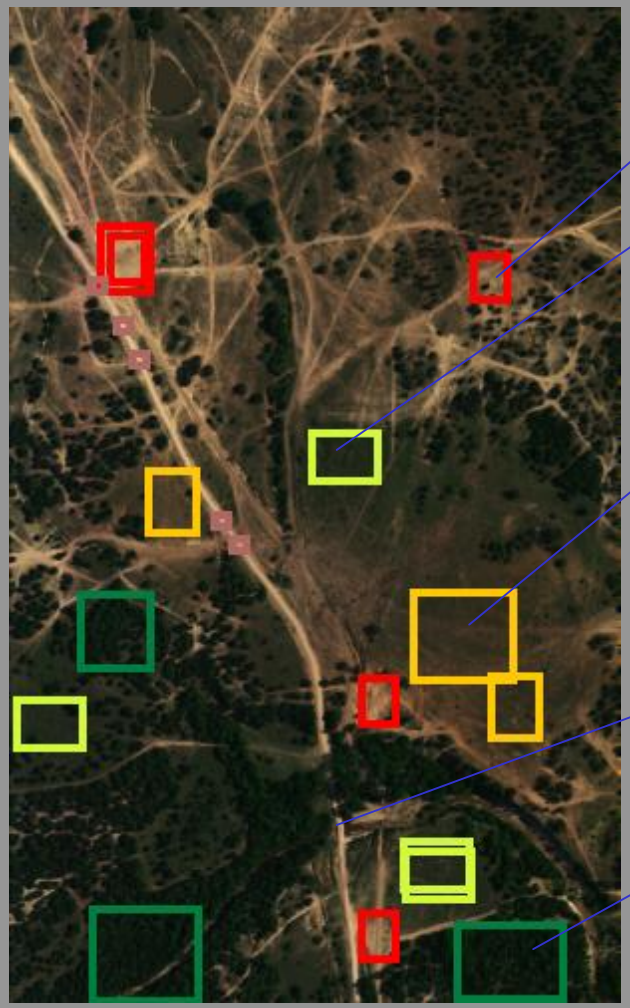
Data collected in 225 different spectral bands. Some bands have more relevant information than others. Objective is to find optimum combination of three bands to differentiate classes of interest.

WAY-2C

“True Color” TERRAIN Image
(class training regions outlined)

HYDICE
bands 49,35,15

Note that user-defined example training regions, color-coded by class, are not particularly “pure”.

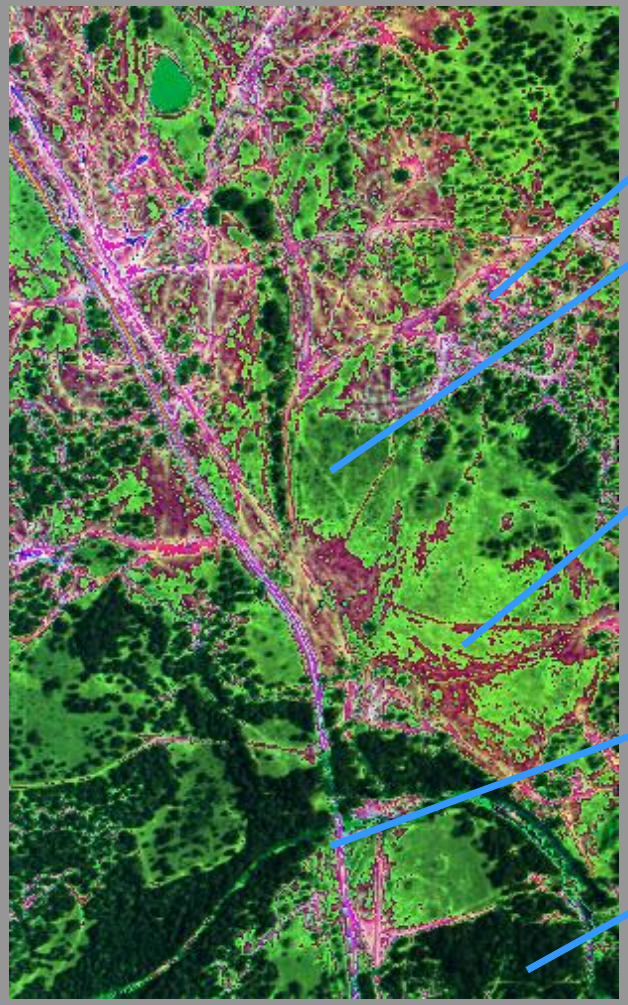


WAY-2C

“Most Relevant” TERRAIN data subset

HYDICE
bands 191,52,17

> 98% data reduction from original hyperspectral cube.

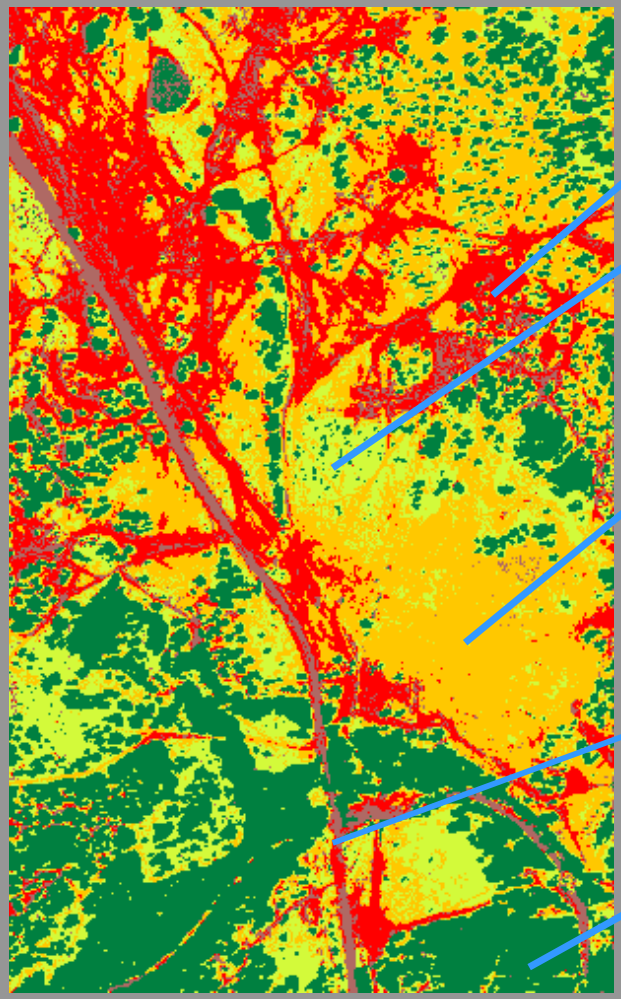


WAY-2C

Classified TERRAIN image

Based on
HYDICE
bands 191,52,17

Color coding
same as training
region image.



bare soil
thick grass
thin grass
road
trees

WAY-2C

Summary

Using the simple analogy of gray-scale, color and hyperspectral images, with user-defined classes of interest, we've shown steps in:

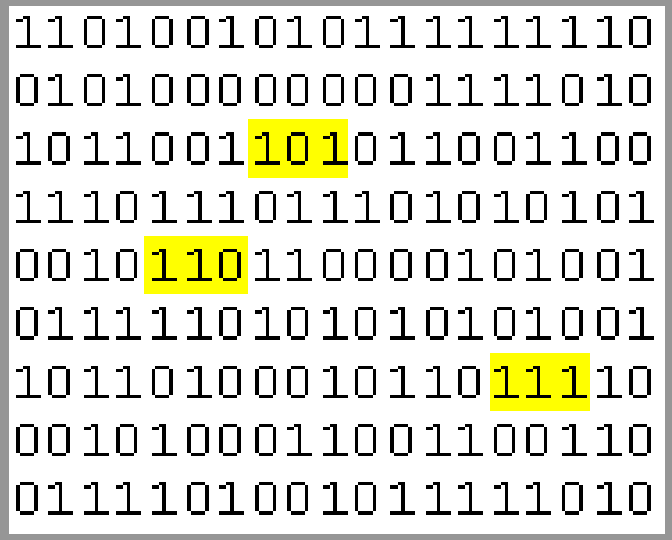
- Finding a most relevant data subset upon which to base a classification.
- Using that most relevant subset to perform maximum likelihood classification.

Advantages

- Method useful for wide variety of data sets.
- Quickly finds most relevant subset combination.
- Can dramatically reduce data transmission and storage bandwidth requirements.
- Eliminates errors caused by assumption of “parametric” data distributions.
- Eliminates need for manual outlier removal.
- Provides numerical estimate of accuracy.
- Speeds data interpretation.
- Provides maximum likelihood classification.

Automated Determination of Optimum Data Subsets for Robust Classification

End



Robert K. McConnell, PhD
WAY-2C
Arlington, Massachusetts
rkm@way2c.com
http://www.way2c.com