Cluster Analysis

Ruben Zamar

Department of Statistics, University of British Columbia

November 12, 2018

Cluster Analysis

A brief definition

Sorting objects in such a way that items in the same group are more similar to each other than to those in other groups.

Approaches to Cluster Analysis

- Partitioning algorithms
- Hierarchical algorithms
- Model based algorithms
- Mean shift algorithms

Partitioning Algorithms

Data: $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, $\mathbf{x}_i \in R^p$

Means: $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K$

Distances:

$$d_i = d\left(\mathbf{x}_i, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K\right) = \min_{1 \leq k \leq K} D\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right)$$

Partitioning Algorithms

Cluster Centers

$$(\widehat{\boldsymbol{\mu}}_1,, ...,\widehat{\boldsymbol{\mu}}_K) = \arg\min\, T\left[d\left(\mathbf{x}_1, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k\right), ..., d\left(\mathbf{x}_n, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k\right)\right]$$

Clusters

$$C_k = \left\{\mathbf{x}_i : D\left(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}_k\right) = \min_{1 \leq l \leq K} D\left(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}_l\right)\right\}$$

K-Means

$$D(\mathbf{a}, \boldsymbol{b}) = \left[\sum_{j=1}^{p} (a_i - b_i)^2 \right]^{1/2} = \|\mathbf{a} - \boldsymbol{b}\|_2 \qquad (\mathrm{L}_2 \text{ distance})$$

$$T(d_1, ..., d_n) = \frac{1}{n} \sum d_i^2$$

K-Mediods

$$D(\mathbf{a}, \boldsymbol{b}) = \sum_{j=1}^{p} |a_i - b_i| = \|\mathbf{a} - \boldsymbol{b}\|_1 \qquad \text{(Manhattan Distance)}$$

$$T(d_1, ..., d_n) = \frac{1}{n} \sum d_i$$

K-Tau

$$D(\mathbf{a}, \boldsymbol{b}) = \left[\sum_{j=1}^{p} (a_i - b_i)^2 \right]^{1/2} = \|\mathbf{a} - \boldsymbol{b}\|_2 \qquad (\mathrm{L}_2 \text{ distance})$$

$$T(d_1, ..., d_n) = s^2 \frac{1}{n} \sum \rho \left(\frac{d_i}{s} \right)$$

$$s = Med(d_i)$$

The Number of Clusters

- Estimating the number G of clusters is one of the most difficult problems in cluster analysis

- A common approach: try different values of G and choose the one that maximizes the clusters strength.
- Cluster strength can be measure in several ways:
  - Silhouette
  - Gap statistics
  - Graphical approach (elbow in the interclass sum of squares plot)

Mean Shift Algorithms

- Strategy: iteratively move the data points toward the cluster centers.
- The number of different limiting points is an estimate for G.
- Example: in *gravitational clustering* data points are viewed as particles of unit mass and zero velocity attracted toward cluster centers by gravitational forces.

$$\text{Number of Clusters} = \text{Number of Fix Points}$$

Attractors

- Peña, Viladomat and Zamar (2012) present an algorithm – **ATTRACTORS** – to move observations toward cluster centers

- Iteratively, observations move to the location of its **nearest-neighbors median** approaching several fix points

$$m_0 = x_i \quad \text{(data point)}$$

$$A_l = \{k \text{ nearest neighbors of } m_l\}$$

$$m_{l+1} = Median_{x_j \in A_l}\{x_j\}$$

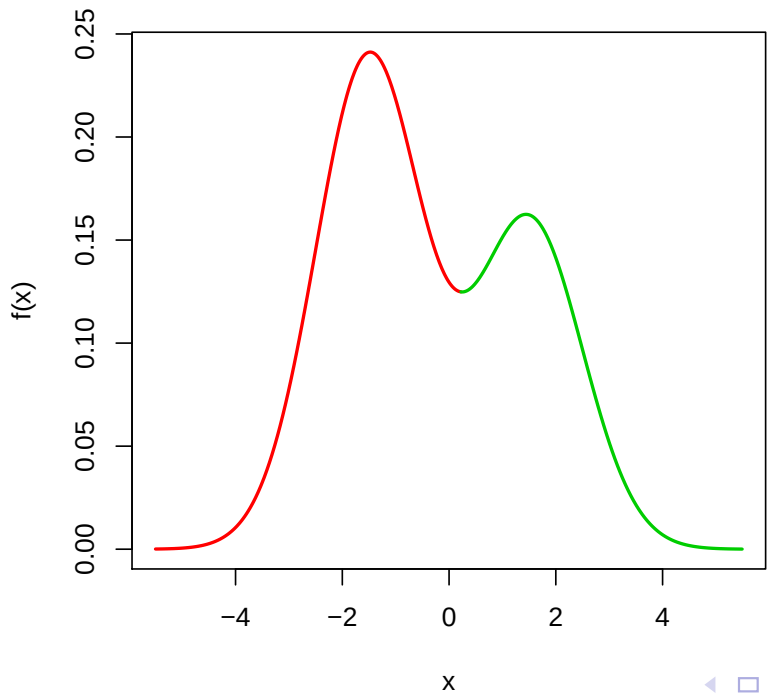
- The sequence of local medians m_l move toward the peaks and away from the valleys of the data density

Attractors Domains

$$g(x) = F^{-1}\left[F(x-d) + \frac{\alpha}{2}\right]$$

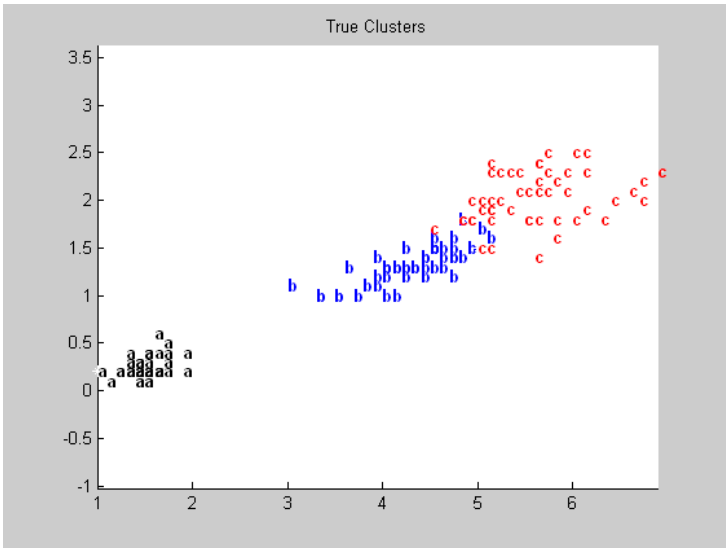
$$F(x+d) - F(x-d) = \alpha$$

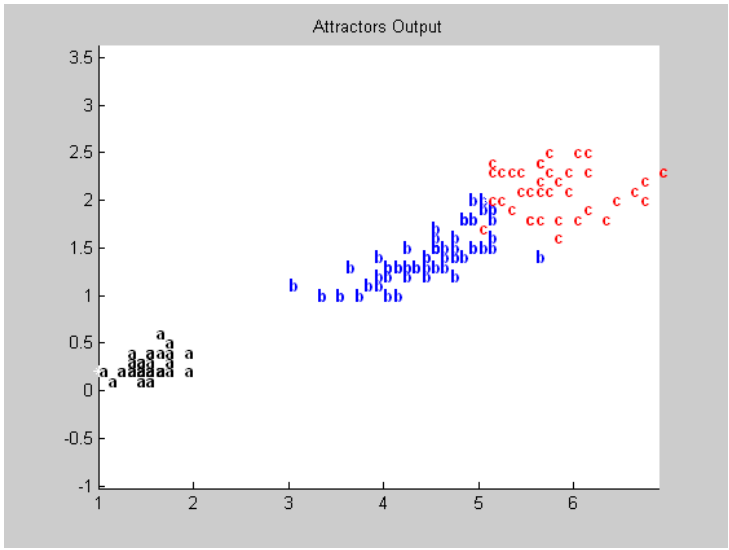
$$x_{(m+1)} = g(x_m)$$



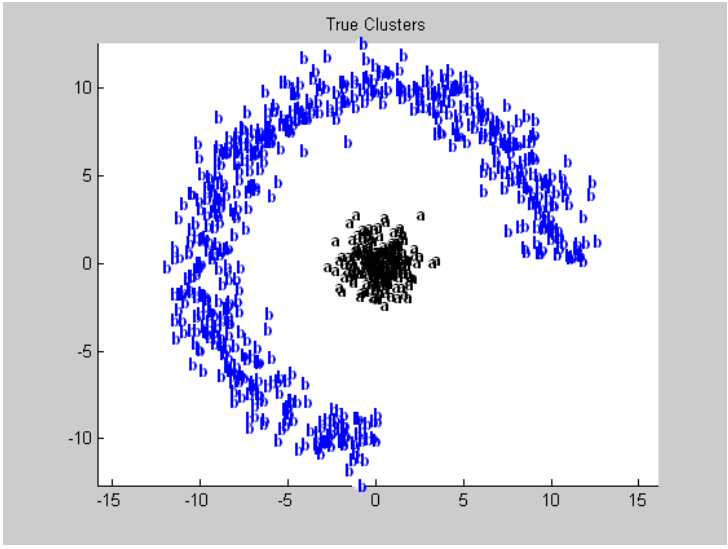
Fisher's Iris Datas

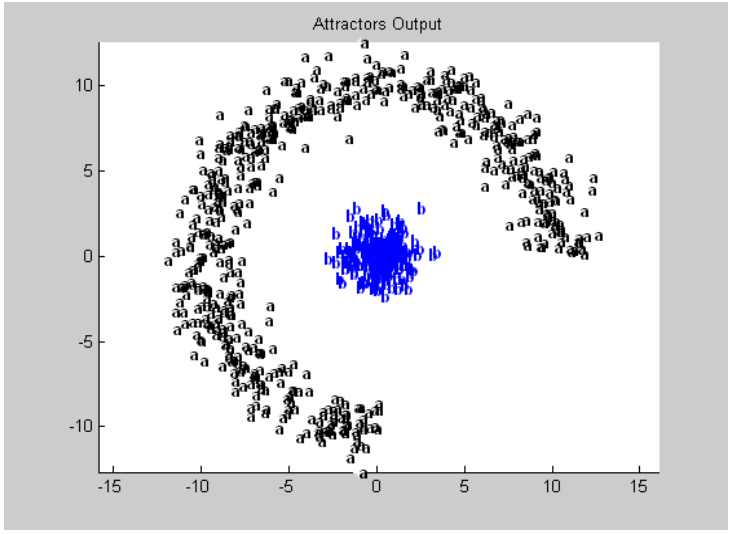
Projection on Setal Length and Width





Ring - Synthetic Data





Clusters Around Lower Dimensional Hyperplanes

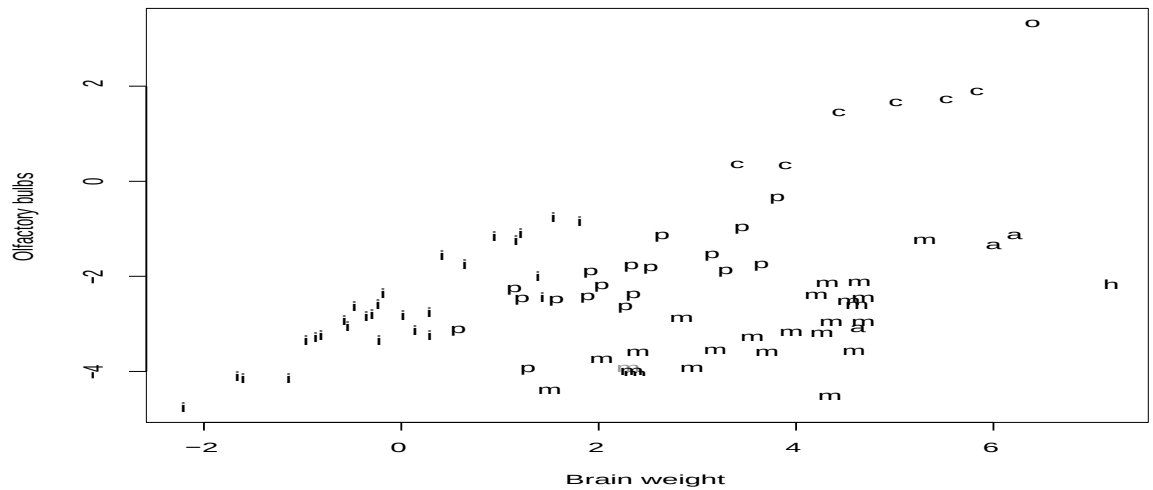
- Traditional cluster centers are points
- Points are zero-dimensional hyperplanes
- Van Aelst, Wang, Zamar and Zhu (2006) propose an algorithm to find clusters around low dimensional hyperplanes (points, lines, planes, etc)
- The proposed algorithm is called Linear Grouping Algorithm (LGA)
- Garcia-Escudero, Gordaliza, San Martin, Van Aelst and Zamar (2009) develop a robust version

Application to Allometry

- Biologists investigate the relationships between sizes of organs for different species.
- Generally, when the size of one organ is large the size of other organs is also large
- For example, a larger body also requires a larger brain. These relations are driven by the evolution process.

Brain Weight and Olfactory Bulb Volume

Data for 83 mammal species (log scale) kindly provided by Dr. Jerison (UCLA)



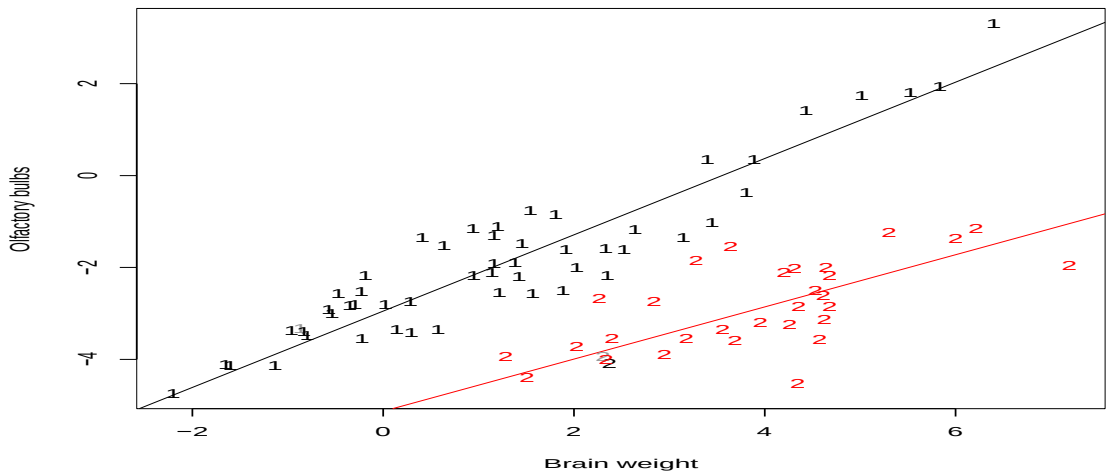
Application to Allometry

- Typically, there exists a linear association between the sizes of organs
- The linear associations are not the same across different classes of species because of different living habits, environment, food sources, etc.
- Different mammal species developed their smell senses according to their living environment, food searching, danger identifying needs, etc.
- Hence, clustering according to different linear patterns is necessary (**a job for LGA**).

Application to Allometry

- Automatic methods to decide the number of clusters yield 2 groups as the optimal number.
- Hence, without using external biological knowledge we would choose $k = 2$ groups

Two Linear Clusters



Two Linear Clusters

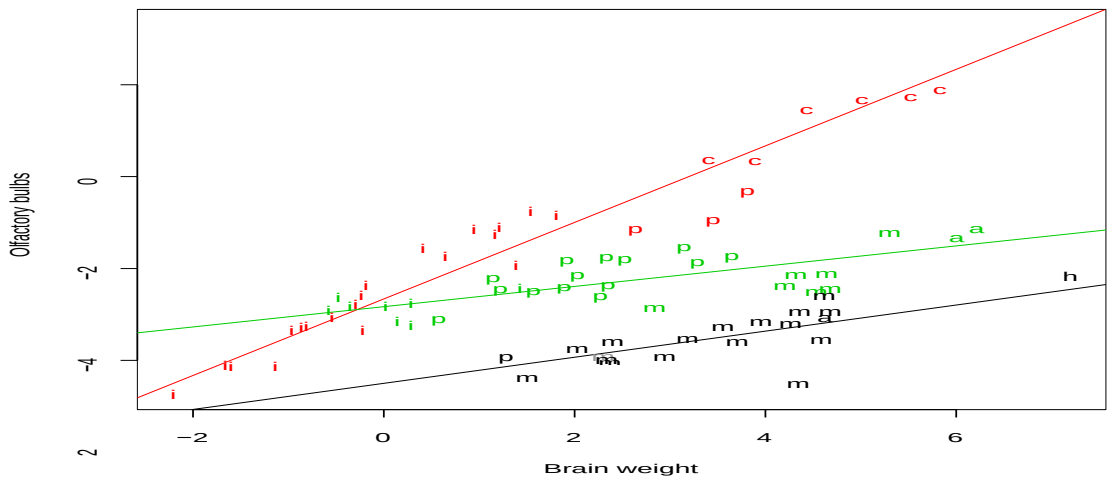
Group 1	Group 2
insectivores	monkeys
carnivores	apes
horses	human
prosimians	

Dr. Jerison Proposes Three Clusters

Group 1	Group 2	Group 3
insectivores	prosimians	monkeys
carnivores		apes
horses		human

Three Linear Clusters

LGA with $k = 3$ finds Dr. Jerison's linear clusters



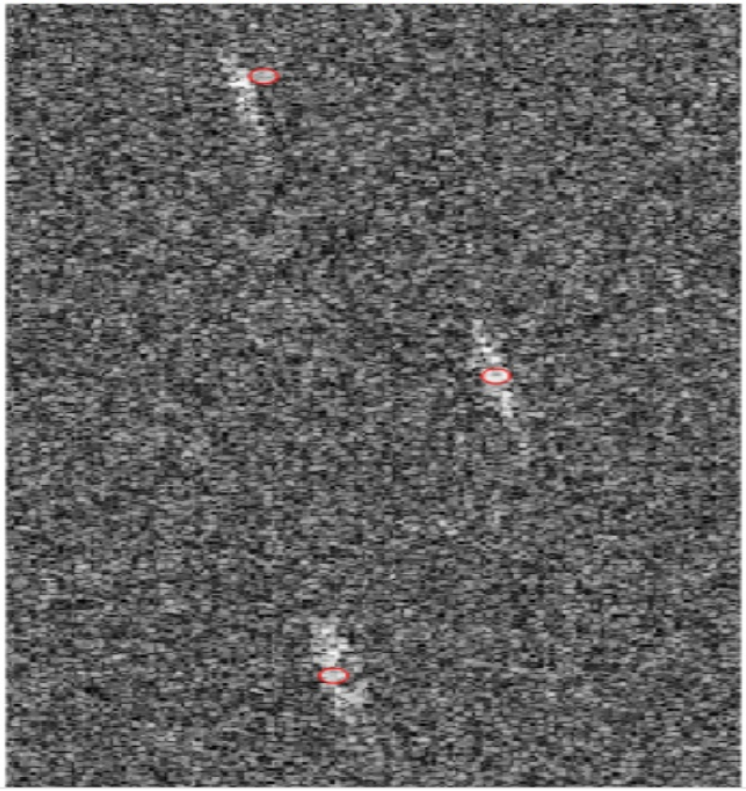
Digitized Image Processing

Potential fields of application for robust cluster analysis

- Computer vision
- Anomaly detection
- Inspection of industrial items
- Search for tumors in microscopic images
- Search for vessel in satellite images

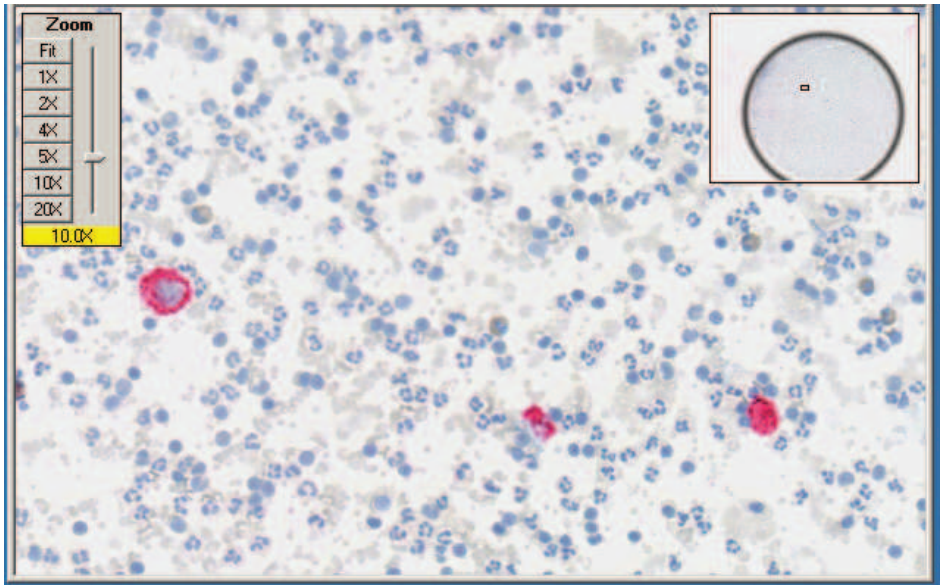
Satellite Imagery Used by Frontex to Detect and Rescue Migrant Boats

While the use by Frontex of satellite imagery is not new, Frontex released a copy of a satellite image used last week to detect and rescue 370 people on board three inflatable boats off the Libyan coast. (It is unclear whether the image made available by Frontex shows the actual spatial resolution available to Frontex.)



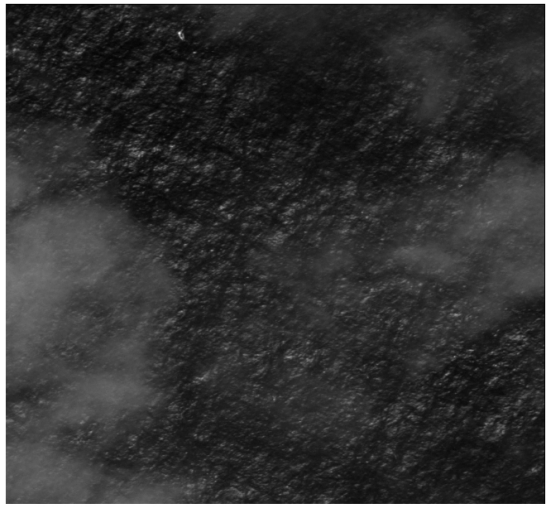
According to Frontex, the imagery is part of "Frontex's Eurosur Fusion Services ... made possible by the cooperation between experts at Frontex and the European Maritime Safety Agency (EMSA), Italian authorities and EUNAVFORMED. ... The Eurosur [fusion] services already include automated large vessel tracking and detection capabilities, software functionalities allowing complex calculations for predicting positions and detecting suspicious activities of vessels, as well as precise weather and oceanographic forecasts. Fusion Services use optical and radar

- **Micrometastasis of tumor cells in circulating blood**
- An algorithms tuned to find the rare events can process an entire slide quickly and reliably without fatigue

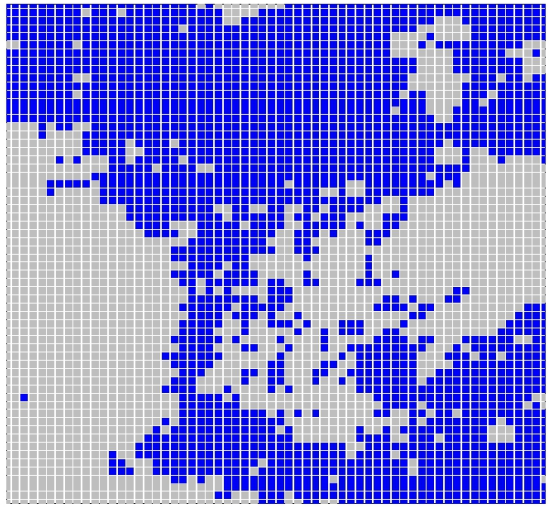


Analysis of a high Resolution Satellite Image

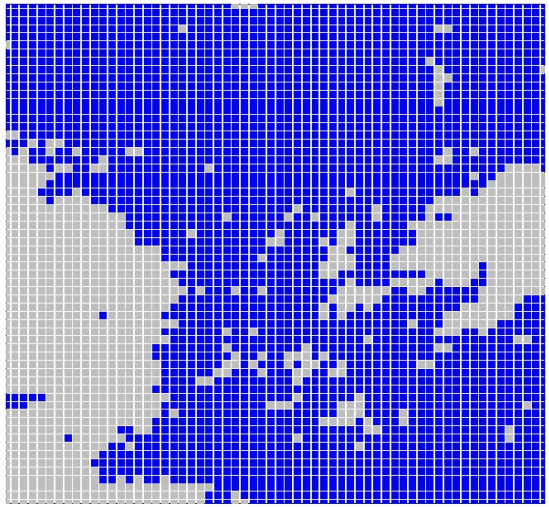
Image



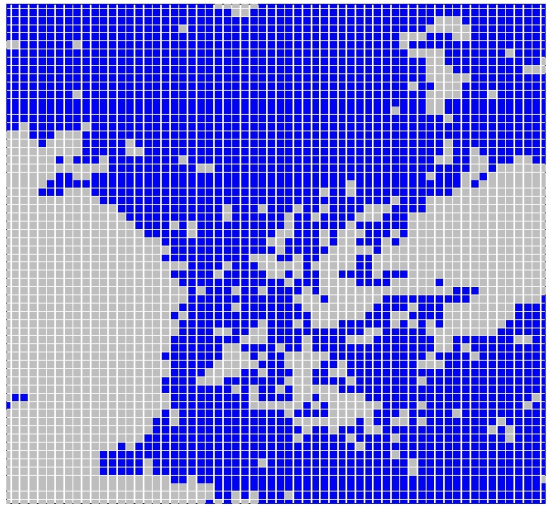
TAU



KMeans



TKMeans

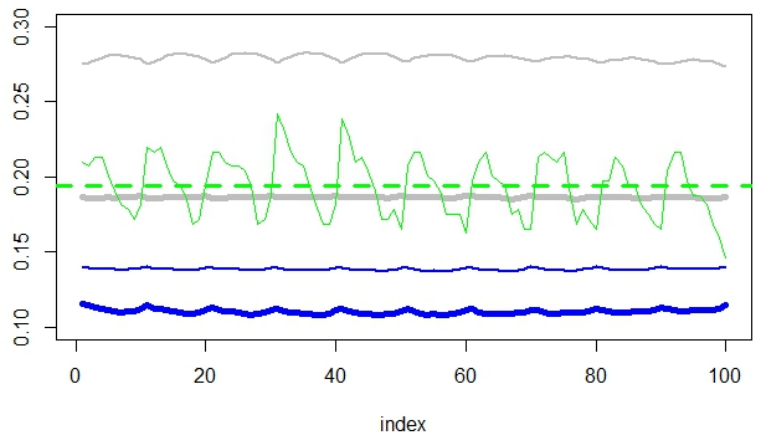


The Satellite Image Data

- A satellite image provided by INFOSAT, covering $500m^2$ of the ocean.
- The image has two components: clouds and water.
- The high resolution image (1 pixel $= 0.02m^2$) is divided into 3844 square cells, each packing 16×16 pixels.
- Each pixel conveys a gray-level intensity scaled between zero and one.
- our dataset consists of 3844 points in a 256-dimensional square.
- Goal: segment the image into two clusters: the **cloud cluster** and the **water cluster** using K-Tau

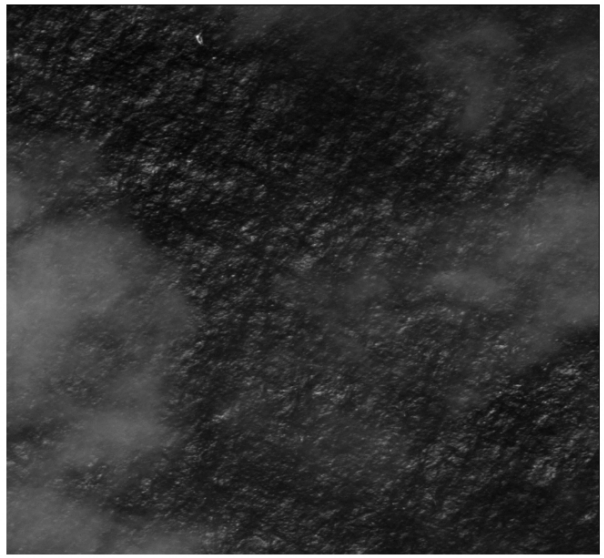
Outliers in the Satellite Image Data

- A patch of higher altitude clouds bearing large gray-level intensity levels
- The outliers brings up the level of the K-means clouds-cluster center.

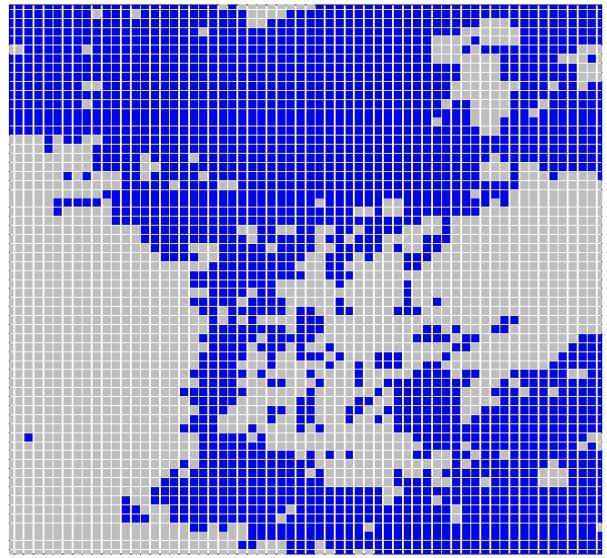


Water-Claud Segmentation

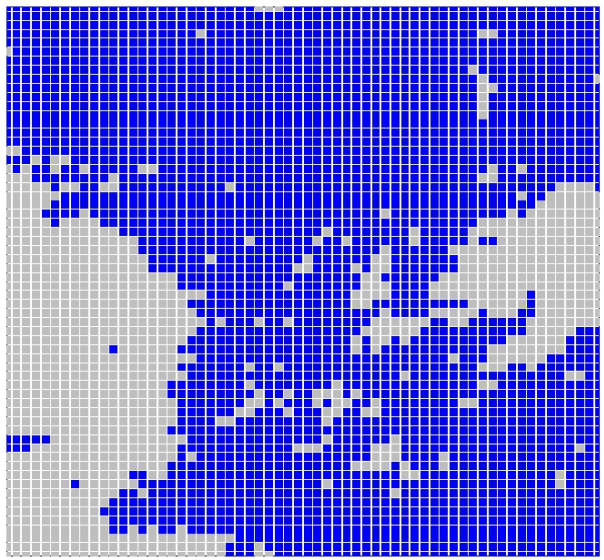
Original Image



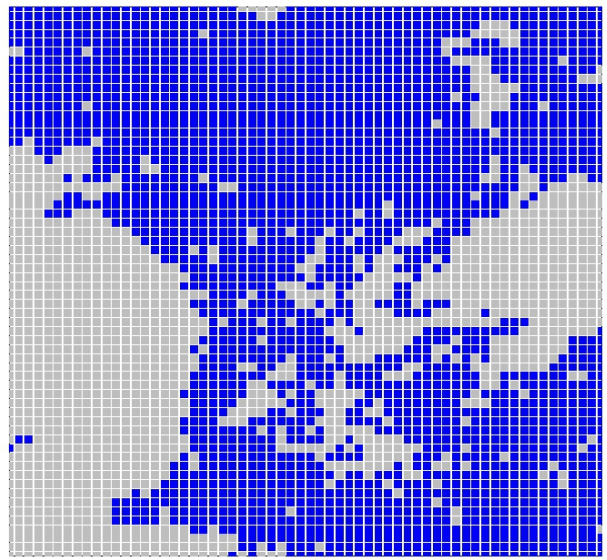
TAU



KMeans



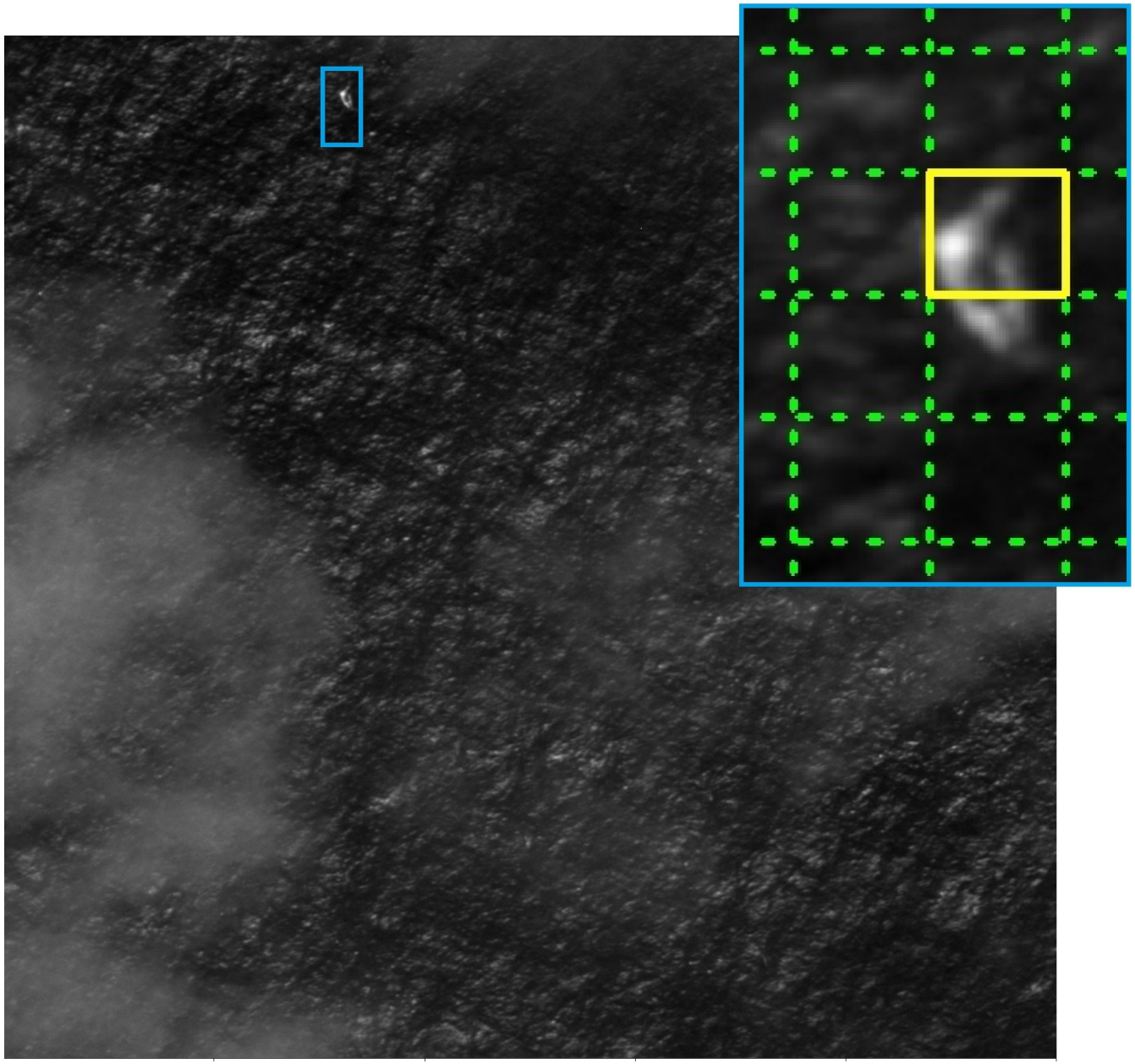
TKMeans



Cluster Outliers in the Satellite Image

- The small yacht reflects the signal different from the water and the clouds in the image.
- the cells containing the boat should appear as cluster outliers.
- We identify the largest outlier, that is, the cell lying furthest away from its cluster center

Finding the Tunante II in the Satellite Image



Analysis of Color Pictures

Argentina's No-Signal TV Image

$$\mathbf{40 \times 50 = 2000\ pixels}$$

Pixel Color

RGB-Color Coding

Each pixel corresponds to a 3-d vector:

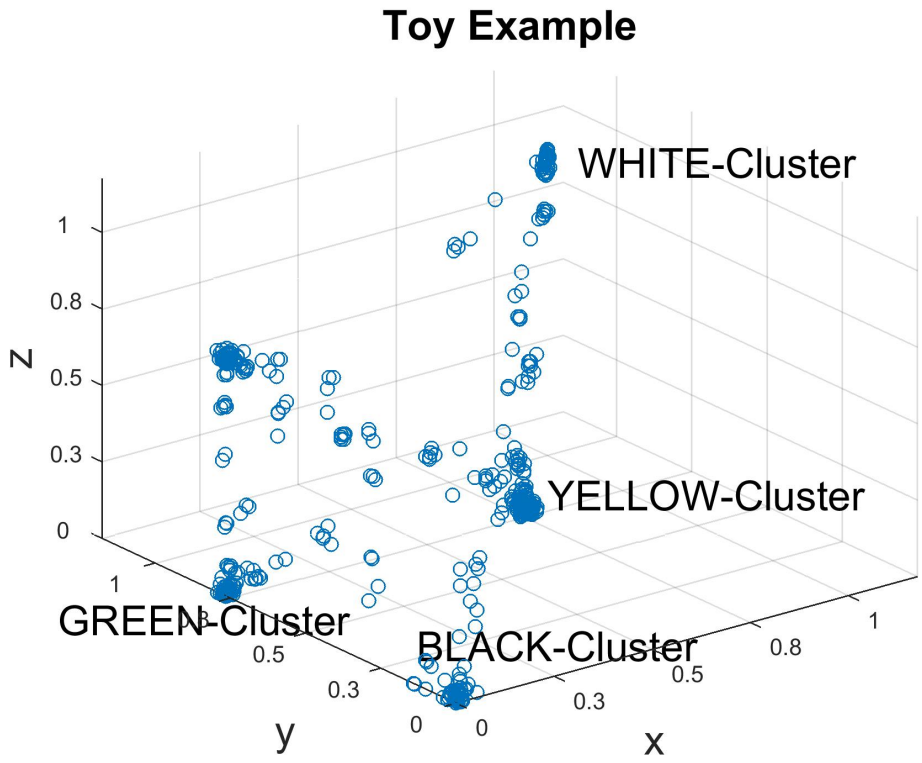
$$(\ R,\ G,\ B\ ),\quad 0\ \leq\ R, G, B\ \leq\ 1.$$

The vector $(\ R,\ G,\ B\ )$ gives the intensity of red, green and blue for the pixel.

For example:

$$(R, G, B) = (1, 0, 0) = \text{ Red}$$
$$(R, G, B) = (0, 1, 0) = \text{ Green}$$
$$(R, G, B) = (0, 0, 1) = \text{ Blue}$$
$$(R, G, B) = (1, 1, 1) = \text{ White}$$
$$(R, G, B) = (0, 0, 0) = \text{ Black}$$
$$(R, G, B) = (.5, .6, 0) = \text{ Yellow}$$

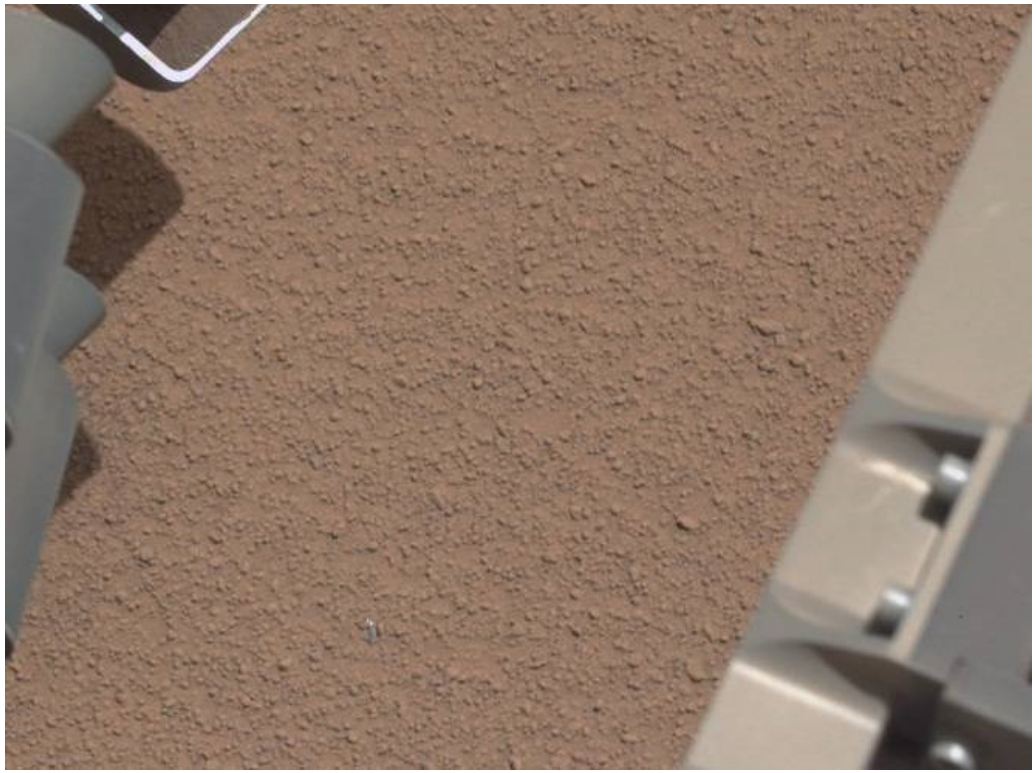
TELEFE Picture Again (left side only)



Mars Rover Curiosity

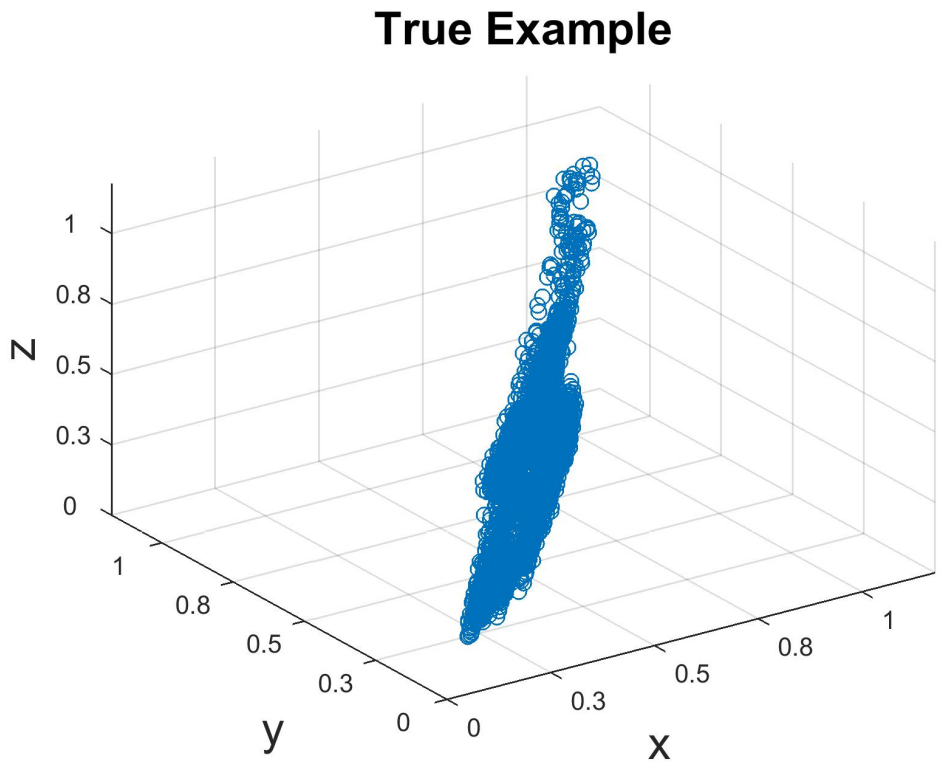
Part of a high resolution NASA's picture

$$\mathbf{495 \times 664 = 328{,}680 \text{ pixels}}$$



Mars Rover Curiosity

RGB - Representation



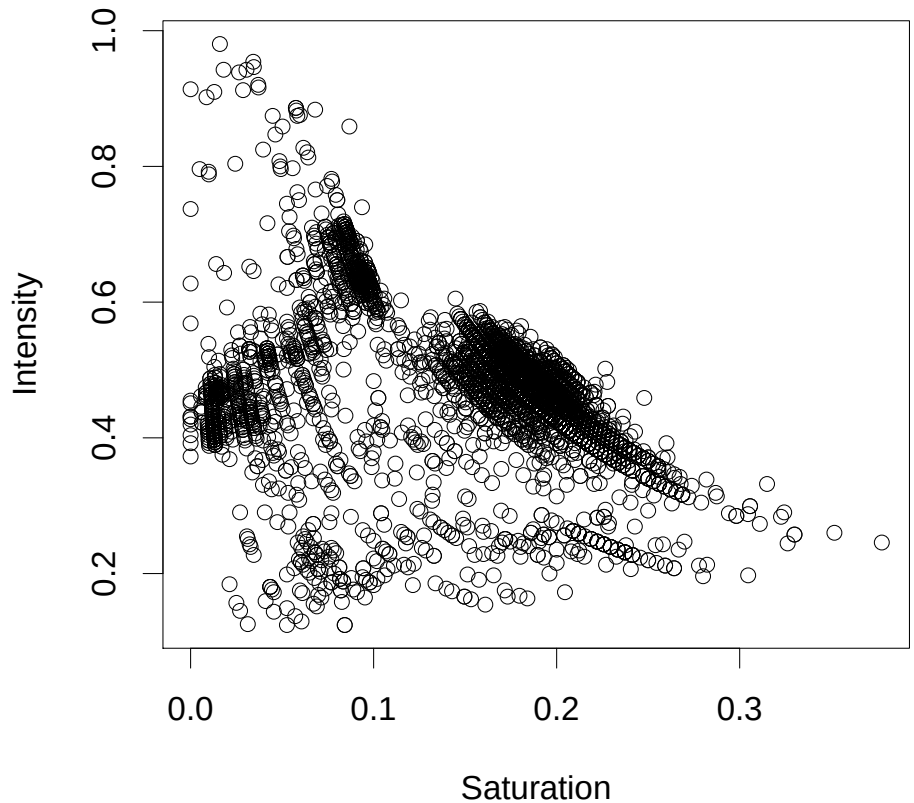
Color picture 2-d representation

$$I = \frac{R + G + B}{3} \qquad \text{Intensity}$$

$$S = 1 - \frac{3 \times \min\{R, G, B\}}{R + G + B} \qquad \text{Saturation}$$

Mars Rover Curiosity

SI - Representation



Cluster Analysis

- Image is divided into $n = 3234$ cells with 10×10 pixels
- Each pixel has two values (I, S)
- Each observation represents a 200-d vector

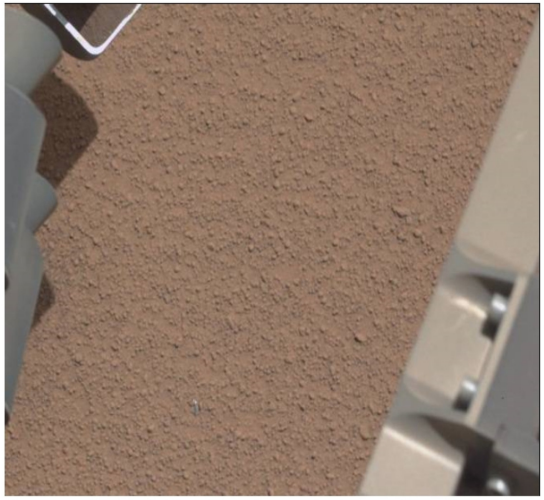
$$(I_1, ..., I_{100}, S_1, ..., S_{100})$$

- The picture has three components (clusters):
  - *shinning metal (SHM)*
  - *opaque metal (OPM)*
  - *sand (SND)*

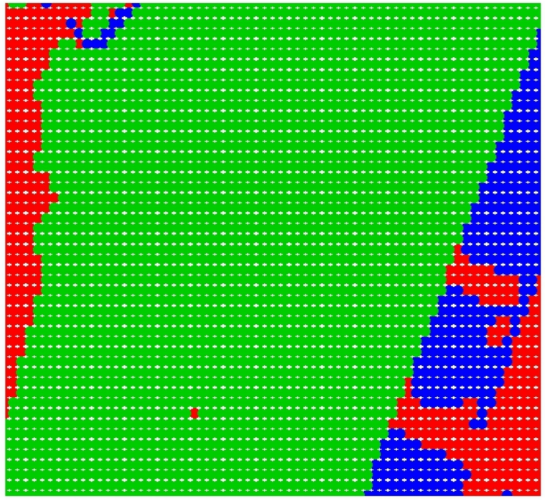
Picture Segmentation

From robust and non-robust clustering

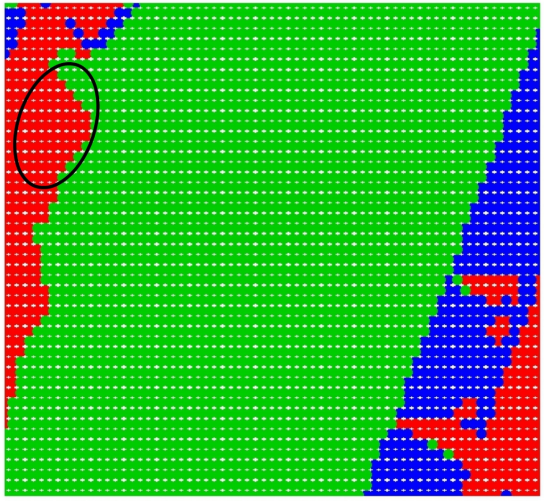
(a) Original Image



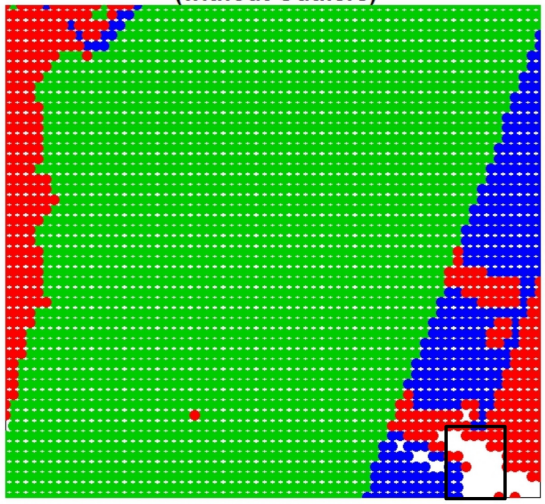
(b) TAU



(c) K Means

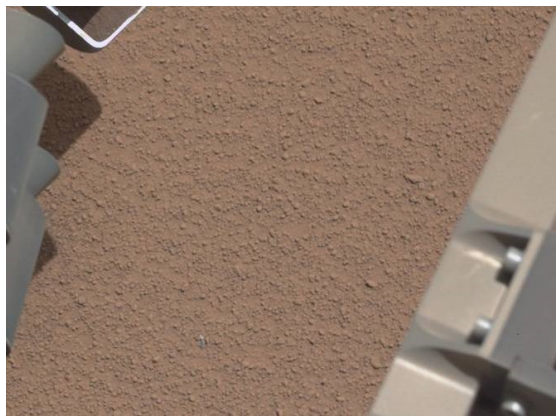


(d) K Means
(without Outliers)

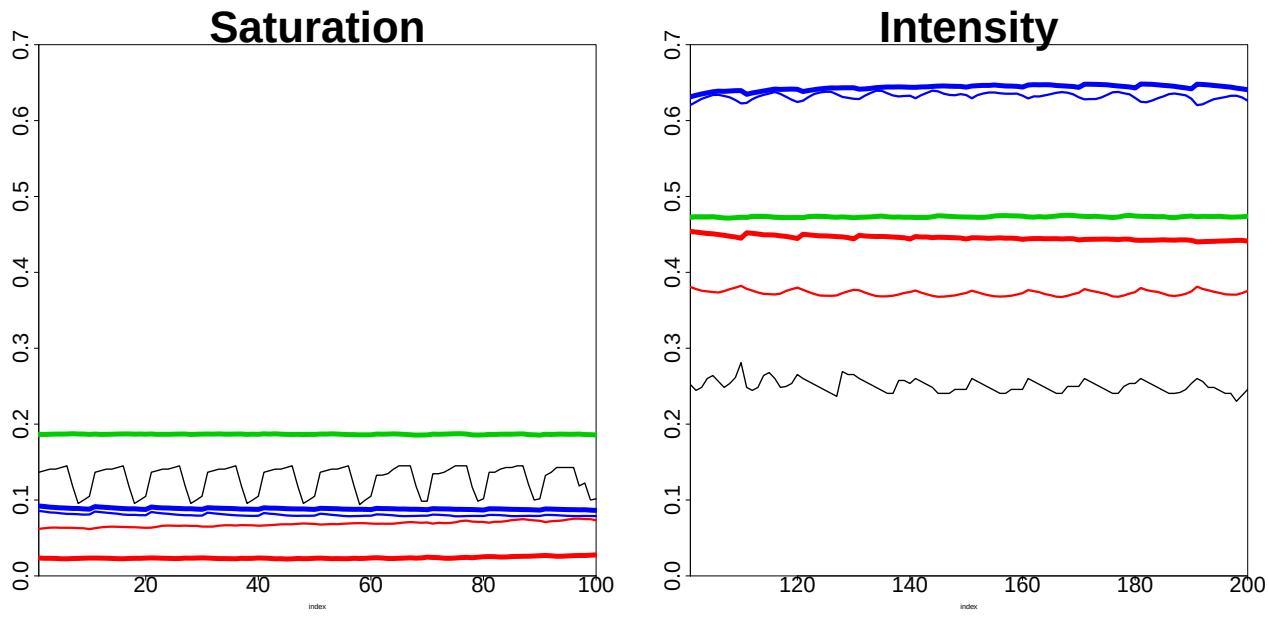


Ouliers

- 15 % of the OPM–cluster cells are very opaque (right lower corner in the picture)
- These cells have unusually low I–levels and high S–levels.
- These outliers upset the K-Means OPM–cluster center.

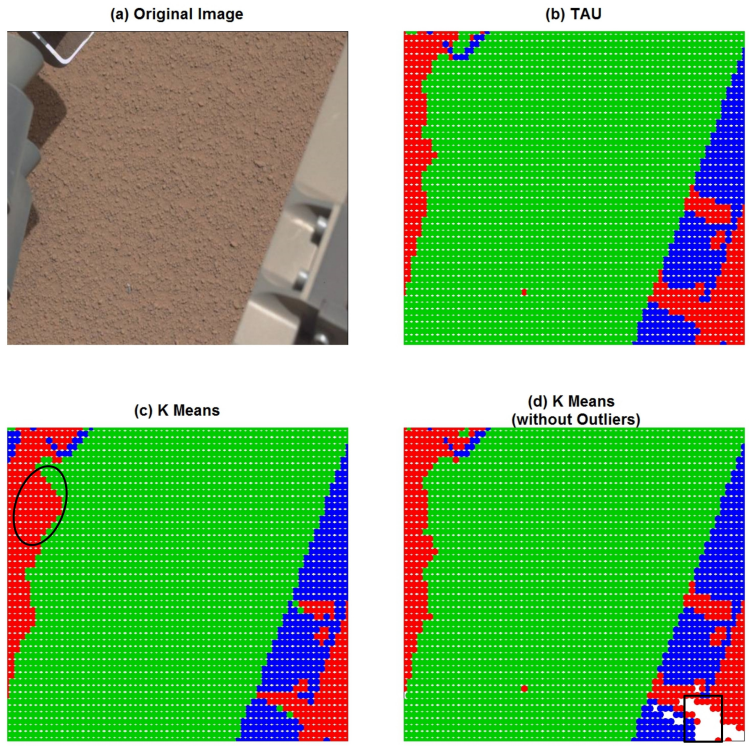


Cluster Centers



Effect of Ouliers

- The shaded sand region is assigned to the OPM–cluster by K-means.
- Recomputing the K-means clusters after removing these outliers validates this reasoning.
- Now the K-means results are consistent with those of the robust clustering procedures.



Automatic detection of missing objects

- The robust clustering sorted the cells as follows:
  - $n_1 = 2500$ SND-cluster cells
  - $n_2 = 405$ OPM-cluster cells
  - $n_3 = 329$ SHM-cluster cells

- The screw is made of a material – opaque metal – that makes up 13% of the image

"Geographic" Step

- Restrict attention to the $n_2 \times 2$ *Geographic Data Matrix* with the position (latitude and longitude) of the OPM–cluster cells
- Perform a second robust cluster analysis on these Geographic data
- Isolated outliers in this second robust clustering are likely to locate the missing piece.

The geographic data