Author(s): Äyrämö, Sami; Pölönen, Ilkka; Eskelinen, Matti

Title: Clustering Incomplete Spectral Data with Robust Methods

Year: 2017

Version:

Please cite the original version:

Äyrämö, S., Pölönen, I., & Eskelinen, M. (2017). Clustering Incomplete Spectral Data with Robust Methods. In E. Honkavaara, B. Hu, K. Karantzalos, X. Liang, R. Müller, E. Nocerino, I. Pölönen, & P. Rönnholm (Eds.), ISPRS SPEC3D 2017 : Frontiers in Spectral imaging and 3D Technologies for Geospatial Solutions (pp. 13-17). International Society for Photogrammetry and Remote Sensing. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-3/W3. https://doi.org/10.5194/isprs-archives-XLII-3-W3-13-2017

CLUSTERING INCOMPLETE SPECTRAL DATA WITH ROBUST METHODS

S. Äyrämö[a],*I. Pölönen[a], M.A. Eskelinen[a]

[a] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland - (sami.ayramo, ilkka.polonen, matti.a.eskelinen)@jyu.fi

Commission III, WG III/4

KEY WORDS: Robust, Clustering, Spectral data, Interpolation, K-means, nan-K-spatmed

ABSTRACT:

Missing value imputation is a common approach for preprocessing incomplete data sets. In case of data clustering, imputation methods may cause unexpected bias because they may change the underlying structure of the data. In order to avoid prior imputation of missing values the computational operations must be projected on the available data values. In this paper, we apply a robust nan-K-spatmed algorithm to the clustering problem on hyperspectral image data. Robust statistics, such as multivariate medians, are more insensitive to outliers than classical statistics relying on the Gaussian assumptions. They are, however, computationally more intractable due to the lack of closed-form solutions. We will compare robust clustering methods on the bands incomplete data cubes to standard K-means with full data cubes.

1. INTRODUCTION

Missing value imputation is a common approach for preprocessing incomplete data sets. In case of data clustering, imputation methods may cause unexpected bias because they modify the underlying structure of data. In order to avoid prior imputation of missing values computational operations must be projected on the available data values.

Hyperspectral imagers use different approaches for separating different wavebands from each other. Pushbroom cameras use variations of prism structures that divide the incoming radiation to the sensor cell. In filtering spectral imagers the data cube is formed by tuning or changing filters in front of the sensor cell and needed optics. There is variation in the kinds of filters and sensors used.

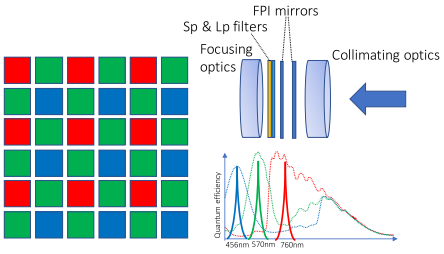

Figure 1. Working principle of Fabry-Perot interferometer. While changing the cap between mirrors, transmitted waveband and its orders passes system to the RGB cell. By selecting caps carefully it, is possible to capture three wavebands with one shot. Full resolution data cube actually contains missing data wavebands in different pixels due the Bayern matrix.

One possible filtering structure is a Fabry-Perot Interferometer with a colour CMOS cell (Saari et al., 2013). To gain full resolution images from the sensor, one must demosaic the Bayer pattern image using some interpolation method. This is due to the fact that the Bayer filter matrix acts to block certain wavebands from each pixel, with the precise configuration determined by the pattern of the matrix. Spectral imagers of this and derivative kinds have been developed by VTT, IMEC and Cubert among the others. On the other hand, some novel approaches for pushbroom cameras increase their frame rate by reading only every second or third band from the sensor cell. If the camera were to image every third band in an interleaved fashion, i.e. changing the set of imaged bands in each line, we would gain spectral images with different missing bands in each line.

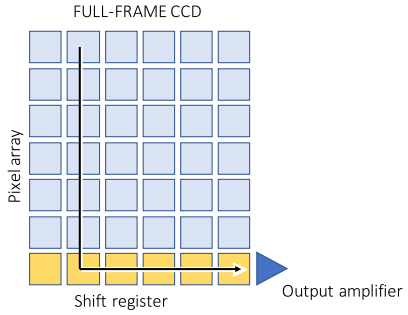

Figure 2. Readout circuit of the full-frame CCD array. Readout is connected to the frame rate. We can double the maximum frame rate, if we read only every second column from the censor cell.

We can treat both of these problems as problems of missing data. In this paper, we apply a robust nan-K-spatmed algorithm to the clustering problem on hyperspectral image data. Robust statistics are more insensitive to outliers than classical statistics relying on the Gaussian assumptions. They are, however, computationally more intractable due to the lack of closed-form solutions. We will compare robust clustering methods on the data cubes with missing bands to standard K-means with full data cubes.

*Corresponding author

Figure 3. A part of X-rite ColorChecker board composed from wavebands: 489, 534, and 618 nm

2. METHODOLOGY

2.1 Data cubes

Data sets are imaged with a framing spectral imager developed by VTT. In the VTT's camera the spectral separation is based on piezo-actualized Fabry-Perot interferometer (FPI). We used wavebands from 480 to 790 nm. Waveband calibration for FPI camera was done immediately after imaging (Saari et al., 2013). Reflectance images are calculated by dividing the radiance images with a white reference image. Below this data set will be referred to as *interpolated data set*.

An image was taken of a part of an X-rite ColorChecker board, which is shown in Figure 3.

After calibration the Bayer pattern is reconstructed from the data cube, so that values which result from the interpolation are replaced by NaN values. Below this data set is referred to as *missing data set*. To simulate pushbroom functionalities we used the same test set so that different wavebands were changed to NaN values on different lines of the image. We composed three data sets using this method. Every second, fourth and tenth lines and bands were used in the data sets. Thus the data sets contained $1/2$, $3/4$ and $9/10$ of missing values of the whole data set. Below these data sets are referred to *E2, E4 and E10 data set* . All the computation and data transformation were carried out on Matlab 2016b using custom-made functions as well as standard toolboxes.

As Figure 4 points out, dark margins of each color area create noise to the data. Thus, we also drew a sub-sample of size 100×100 pixels from each color area, which makes comparison between the results easier.

2.2 Robust clustering using spatial median

Real data are often incomplete, noisy and may contain even large outliers. There are several strategies to deal with incomplete data and outliers. In the presence of missing data one can either discard the incomplete data points, impute the missing values, or utilize all the available data values (Little and Rubin, 2014). Data contamination can be managed by data filtering, data cleaning, outlier detection, or using robust methods that are less sensitive to outlying values than classical methods. In this study we compare the performance of the K-spatialmedians clustering algorithm with missing data treatment on incomplete hyperspectral data to the well-known K-means method (MacQueen, 1967, Äyrämö, 2006). K-means is a classical partitional clustering method in which the clusters are represented by the sample means (MacQueen, 1967). Advantages of the K-means method are its algorithmic simplicity, computational efficiency, and interpretability of the results. K-spatialmedians is a variant of K-means in which each cluster prototype is represented by a robust multivariate estimate of location called the spatial median

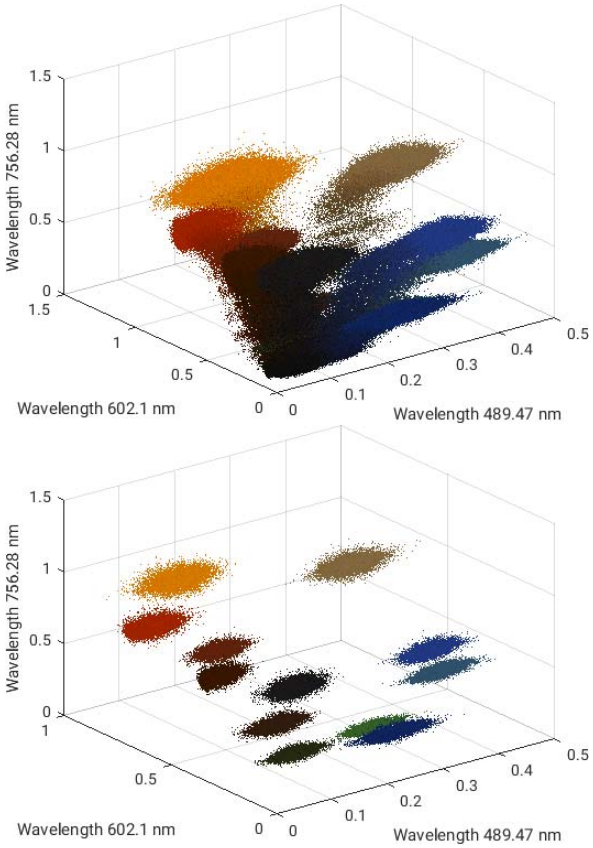

Figure 4. Visualisation of captured data set. The **upper** plot represents whole data, **lower** is sample taken from each color target. We can see that there is noise on data and border area of color checker card creates some disruption between clusters.

(Äyrämö, 2006, Kent et al., 2015). The benefit of substituting the spatial median for the sample mean is greater robustness against outlying points, whereas the cost is the increase in computational complexity (Kent et al., 2015). In the following we describe the nan-K-spatmed algorithm that is a generalized version of the K-spatialmedians method (Äyrämö, 2006).

Let us consider a set of $p-$dimensional data points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. In this study data point refers to the single spectra measured from the hyperspectral image. The goal of cluster analysis is to partition the set of data points $\mathbf{X}$ into set of K clusters $C = \{c_k, k = 1, \ldots, K\}$.

A general class of metric distance functions in the p-dimensional vector space is defined as:

$$l_q(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{p} |(\mathbf{x})_i - (\mathbf{y})_i|^q \right)^{1/q} = \|\mathbf{x} - \mathbf{y}\|_q, \quad q < \infty, \tag{1}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. The most common choices of q are 1, 2, and ∞, that gives us l1-norm, l2-/Euclidean-norm, and max norm, respectively.

The objective of the K-means method is to minimize the sum of the squared error over all K clusters. The squared error is

obtained from eq:lqnorm by choosing $q = 2$:

$$\mathcal{J}(\mathbf{u}) = \sum_{i=1}^{n} \|\mathbf{u} - \mathbf{x}_i\|^2. \tag{2}$$

The objective function of K-means is the squared Euclidean error over the K clusters (MacQueen, 1967):

$$\mathcal{J} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{x}_i - \mathbf{m}_k\|^2 \tag{3}$$

where $\mathbf{m}_k$ is the sample mean of the k^{th} cluster and r_{ik} is determined by:

$$r_{ik} = \begin{cases} 1, \text{if} \quad k = \underset{k}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{m}_k\|^2 \\ 0, \text{otherwise} \end{cases} \tag{4}$$

The batch-type of algorithm for minimizing the K-means objective function iterates between the following steps until the partition does not change:

1. Assign each data points to its closest cluster center c_k
2. Update the cluster centers $C = \{c_k, k = 1, \ldots, K\}$ by computing the sample mean of the assigned points

The initial cluster centers can be chosen randomly or by using some initilization strategy (Pena et al., 1999, Arthur and Vassilvitskii, 2007, Äyrämö et al., 2007).

Sensitivity of K-means toward outlying points is caused by the zero breakpoint point of the sample mean estimate that is used as the representative point for the cluster centers. A more robust error function for the problem of clustering partitioning is obtained by choosing $q = 1$ in (1). The point that minimizes the sum of Euclidean distances to n data points is known as the spatial median (Huber, 1981). The problem of the spatial median is defined as:

$$\min_{\mathbf{u} \in \mathbb{R}^p} \mathcal{J}(\mathbf{u}), \quad \text{for } \mathcal{J}(\mathbf{u}) = \sum_{i=1}^{n} \|\mathbf{u} - \mathbf{x}_i\|. \tag{5}$$

In statistics the spatial median is known as a robust multivariate estimate of location. Its breakdown point is 0.5, that is, more than 50 % of the data points must be contaminated to cause infinite influence on the estimate (Lopuhaä and Rousseeuw, 1991). If the data points are not collinear the spatial median is unique (Milasevic and Ducharme, 1987). If all points are concentrated on a line the spatial median reduces to the univariate median, which is generally not unique. The spatial median is also location and orthogonal equivariant, but not affine equivariant estimator of location.

Due to the lack of a closed-form solution to the problem (5) general optimization methods or problem-specific iterative solutions are needed (Kent et al., 2015). In this study we utilize an iterative over-relaxation variant of the Weiszfeld algorithm (SOR-Weiszfeld) that is extended by available case strategy for finding the minimum of (5) in the presence of missing values (Äyrämö, 2006).

In order to utilize all the available data values we need to first define a diagonal matrix $\mathbf{P}_i$ for each data point $\mathbf{x}_i$. In order to project operations on the available values we define $(\mathbf{P}_i)_{j=k} = 0$ if j^{th} element of data vector $\mathbf{x}_i$ is missing and otherwise $(\mathbf{P}_i)_{j=k} = 1$.

The iterative SOR-Weiszfeld method is based on a smooth "ε-perturbed" formulation in which first-order necessary conditions for the stationary point are given by (Äyrämö, 2006)

$$\sum_{i=1}^{n} \left(\frac{\mathbf{P}_i(\mathbf{u} - \mathbf{x}_i)}{\sqrt{\|\mathbf{P}_i(\mathbf{u} - \mathbf{x}_i)\|^2 + \varepsilon}} = 0 \right. \tag{6}$$

.

$\mathbf{v}$ can be then solved by a "linearized" equation

$$\sum_{i=1}^{n} \alpha_i^t \mathbf{P}_i(\mathbf{v} - \mathbf{x}_i) = 0, \tag{7}$$

where α_i^t defines the explicit weights for the denominator of (6):

$$\alpha_i^t = \frac{1}{\sqrt{\|\mathbf{P}_i(\mathbf{u}^t - \mathbf{x}_i)\|^2 + \varepsilon}}.$$

The solution at t^{th} iteration is obtained by the over-relaxation step:

$$\mathbf{u}^{t+1} = \mathbf{u}^t + \omega(\mathbf{v} - \mathbf{u}^t), \quad \omega \in [0, 2], \tag{8}$$

where ω is the over-relaxation parameter, $(\mathbf{v} - \mathbf{u}^t)$ is the search direction, and $\mathbf{v}$ is obtained from (7). The steps are iterated until the stopping criterion is satisfied.

2.3 The K-spatialmedians for incomplete data

The objective function of the basic K-spatialmedians clustering problem is obtained from (3) by simply replacing the sample mean with the solution of (5)(Äyrämö, 2006). Imputation or discarding of incomplete data points is avoided in nan-K-spatmed by projecting the Euclidean norm to the existing values. The objective function of nan-K-spatmed is then defined as:

$$\mathcal{J} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mathbf{P}_i(\mathbf{x}_i - \mathbf{m}_k)\| \tag{9}$$

where $\mathbf{m}_k$ is the spatial median point of the k^{th} cluster and r_{ik} is determined by:

$$r_{ik} = \begin{cases} 1, \text{if} \quad k = \underset{k}{\operatorname{argmin}} \|\mathbf{P}_i(\mathbf{x}_i - \mathbf{m}_k)\| \\ 0, \text{otherwise} \end{cases} \tag{10}$$

The algorithm used to minimize the nan-K-spatmed objective function follows the same basic steps as the K-means method. The sample mean, that is computed using the available case strategy, was input as the initial guess to the SOR-Weiszfeld algorithm (8). The K-means type of algorithms end up occasionally with one or more empty clusters. In our implementation of nan-K-spatmed, an empty cluster is always discarded and K-1 clusters are returned. In order to find the best possible partition from the target data set the nan-K-spatmed algorithm was initialized using the furthest first principle in which the mutually K most distant data points are being selected as the initial cluster centers. Since incomplete data points do not necessarily lie in the same space (indices of missing elements may not match in pairs of data vectors), an artificial set of complete data points is created by computing the spatial medians points of 10000 random samples (each

of size 10000). In order to minimize computational effort of estimating large numbers of spatial medians the maximum number of over-relaxation iterations (8) was set to five. The furthest first algorithm was then applied to this approximated set of complete spatial median points yielding a set of initial points for the nan-K-spatmed algorithm on the full data.

The following parameters were chosen for the nan-K-spatmed algorithm:

- Number of clusters $K = 12/13/14$ (depending on data sets)
- Maximum number of clustering iterations $= 100$
- SOR-stepsize $= 1.5$
- Stopping tolerance SOR $1e - 5$
- Maximum number of SOR iterations = 100

3. RESULTS AND VALIDATION

We studied the performance of the nan-K-spatmed algorithm on hyperspectral data with missing wavebands. Figure 5 summarizes classification results between different subsets of data taken from each of the 12 color targets. The first row corresponds to the ground truth of each target color. On the second and third row we can see that K-means and nan-K-spatmed with missing data set perform equally. Approximately $1/3$ of data points are missing from the incomplete data sets. Both K-means on the interpolated data and nan-K-spatmed on the incomplete data find 11 clusters correctly, but both methods divide one cluster into two parts. K-means makes mistake with the cluster 12 that contains a lot of noise due its location on the corner of original image (see fgure 3), whereas nan-K-spatmed make worst mistake with the cluster number 8 which is contains darkest color.

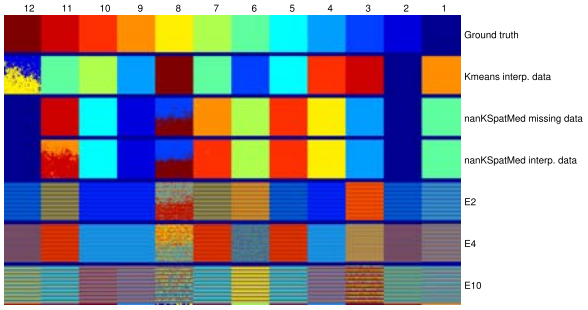

Figure 5. Clustering results for different data sets and methods using 12 subsets from each colors.

Surprisingly, nan-K-spatmed performed even worse on the whole interpolated data set and divided also color 11 into two different clusters. With data sets E2, E4 and E10 nan-K-spatmed seems to fail in general and only partly succeed. It is capable of finding clusters, but not all. Closer examination showed that the method is incapable of connecting consecutive lines to the same cluster. This observation suggests that the nan-K-spatmed algorithm is not able to handle data sets with more than 50% of values missing. Figure 9 shows the clustering results for the whole sets E2, E4 and E10 when $K = 13$. In case of E2 data set nan-K-spatmed is capable of detecting continuous clusters, which is somehow in conflict with the results obtained on the subsets. This could be related to the presence of noise and the dark borders surrounding the color areas.

Figures 6, 7 and 8 show the clustering results for the whole image using K values 13 and 14. Here, in general, it seems that nan-K-spatmed outperforms K-means. In the closer examination we can see that the borders of the color areas are difficult to cluster for both methods.

4. CONCLUSION

We have shown that if spectral data include missing wavebands or values K-spatialmedians method with available case strategy can be applied in clustering. The results are meaningful at least when there is not too many missing values in the data set. We tested also novel initialization for finding the initial cluster centers to start the iterative clustering algorithm. Our approach showed reasonable performance and it was comparable with K-means on the data sets without missing values.

nan-K-spatmed is an appropriate method for clustering in such cases where the sensor itself produces missing values to the data set, but it can also be applied to data sets which for some reason have some missing values on wavebands. For example, specular reflection can cause disturbances on some wavebands only. We can easily replace these values by empty values (NaN) and use nan-K-spatmed for the clustering data set.

The present study points out that initialization and noise level of data affect the clustering results. When noise-to-signal-ratio (SNR) is high K-means and nan-K-spatmed approaches perform equally, but in the presence of low SNR nan-K-spatmed seems to outperform K-means. If data is biased or includes outliers all the clusters may not found, at least, without proper initialization due to the lack general of robustness of the K-means type of partitioning methods (Garcia-Escudero and Gordaliza, 1999).

REFERENCES

Arthur, D. and Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, pp. 1027–1035.

Äyrämö, S., 2006. Knowledge Mining using Robust Clustering. PhD thesis, University of Jyväskylä.

Äyrämö, S., Kärkkäinen, T. and Majava, K., 2007. Robust refinement of initial prototypes for partitioning-based clustering algorithms. *Recent Advances in Stochastic Modeling and Data Analysis: Chania, Greece, 29 May-1 June 2007* p. 473.

Garcia-Escudero, L. A. and Gordaliza, A., 1999. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association* 94(447), pp. 956–969.

Huber, P., 1981. *Robust statistics*. John Wiley & Sons.

Kent, J. T., Er, F. and Constable, P. D. L., 2015. *Algorithms for the Spatial Median*. Springer International Publishing, Cham, pp. 205–224.

Little, R. J. and Rubin, D. B., 2014. *Statistical analysis with missing data*. John Wiley & Sons.

Lopuhaä, H. P. and Rousseeuw, P. J., 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* 19(1), pp. 229–248.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Milasevic, P. and Ducharme, G. R., 1987. Uniqueness of the spatial median. *The Annals of Statistics* 15(3), pp. 1332–1333.

Pena, J. M., Lozano, J. A. and Larranaga, P., 1999. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters* 20(10), pp. 1027–1040.

Saari, H., Pölönen, I., Salo, H., Honkavaara, E., Hakala, T., Holmlund, C., Mäkynen, J., Mannila, R., Antila, T. and Akujärvi, A., 2013. Miniaturized hyperspectral imager calibration and uav flight campaigns. Vol. 8889, pp. 88891O–88891O–12.

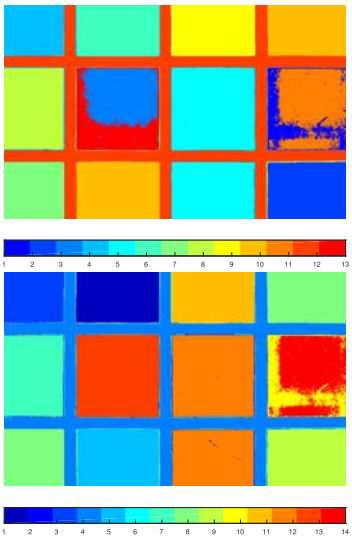

Figure 6. Kmeans clustering result when K is 13 and 14. Border areas of color targets are hard. Three colors are clustered as same.

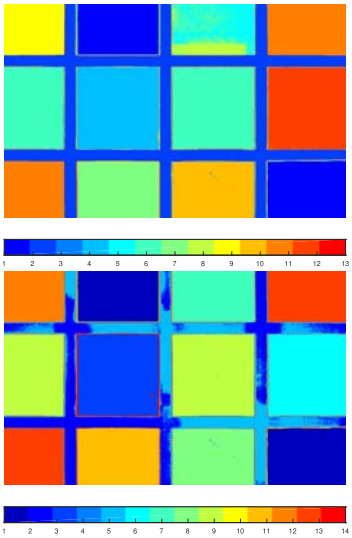

Figure 7. K-spatialmedian clustering results when K is 13 and 14 for incomplete FPI data

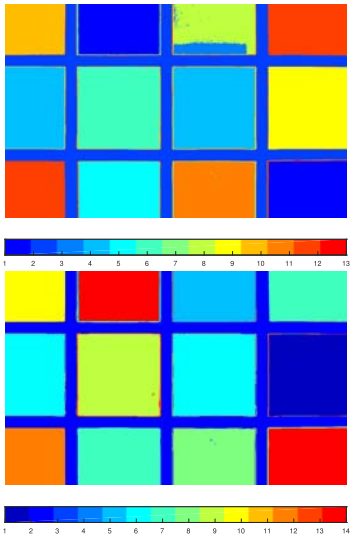

Figure 8. K-spatialmedian clustering results when K is 13 and 14 for the original interpolated data.

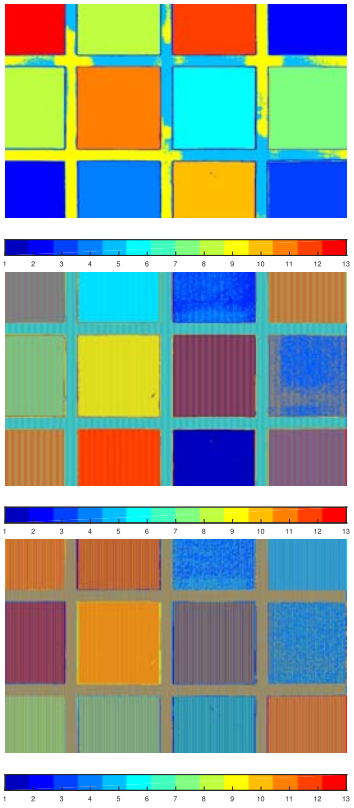

Figure 9. K-spatialmedian clustering results when $K = 13$ for simulated pushbroom data, where every second band is read from second line (E2 data set), every fourth band from fourth line (E4 data set) and every tenth band from tenth line (E10 data set).