Objective delineation of river bed surface patches from high-resolution spatial grain size data

Peter A. Nelson [pnelson@eps.berkeley.edu], Dino Bellugi, and William E. Dietrich
Department of Earth and Planetary Science, University of California, Berkeley, CA 94720

EP53A - 0606

1. Introduction

Gravel-bed rivers often display an organization of distinct textural patches of similar grain size and sorting. Patchiness is a primary control on bed mobility, hydrodynamic roughness, and the distribution of benthic habitat, but despite its importance, we lack answers to the fundamental questions like:

What is a patch?

and

How many patches (or patch types) should emerge on a river bed?

Objective delineation of bed surface patches from high-resolution spatial grain size data is an important step towards answering these questions and advancing our understanding of the morphodynamics of gravel-bed rivers.

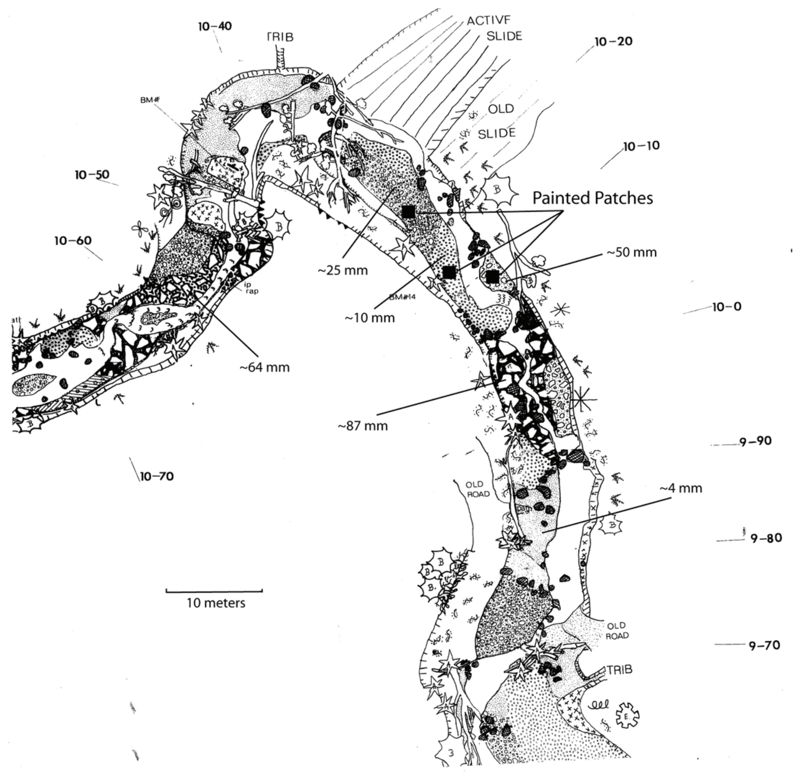


Figure 1.1. Hand-drawn patch map of Wildcat Creek, CA [drawn by Laurel Collins and presented in Dietrich et al., 2005].

2. Testing dataset

A. 1 x 1 mm bed surface DEM from a near-field scale flume experiment using a 2-45 mm gravel mixture [Nelson et al., 2010].

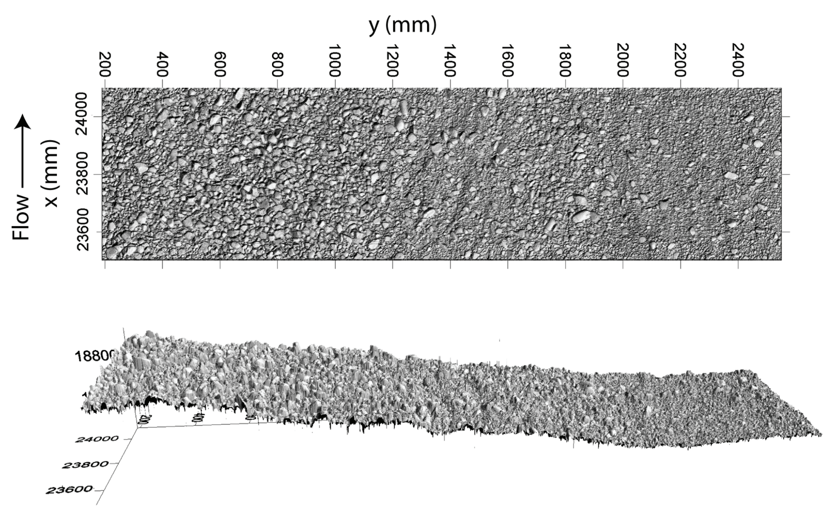


Figure 2.1. Example of the 1x1 mm DEM showing the strong sorting that developed over a bar.

B. Extract the boundaries of grains from hillshade image using automated procedure to generate a dataset of grain centroids (x,y) and diameters (D).

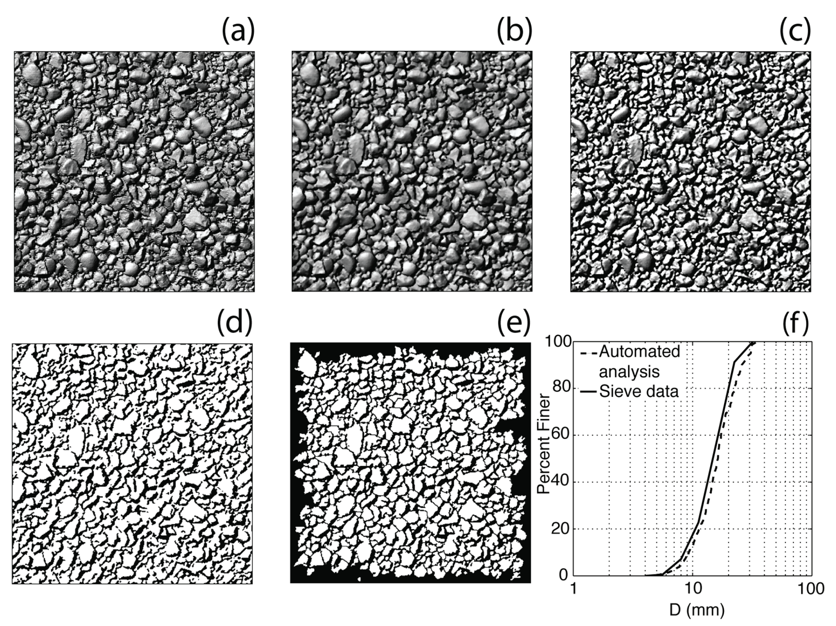


Figure 2.2. Example application of the methodology of Graham et al [2005] to a portion of the 1x1 mm DEM image [from Nelson et al., 2010].

C. The (x,y,D) data are used to generate a spatial grid of m grain size distributions:

$$\mathfrak{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$$

where the distributions are discretized into N size classes such that the ith grain size distribution is represented as:

$$\mathbf{f}_i = (f_{i,1}, \ldots, f_{i,N})$$

where $f_{i,k}$ is the cumulative percent of the kth grain size class in $\mathbf{f}_i$.

This grid of grain size distributions can now be fed into **clustering algorithms** to divide the bed into patches.

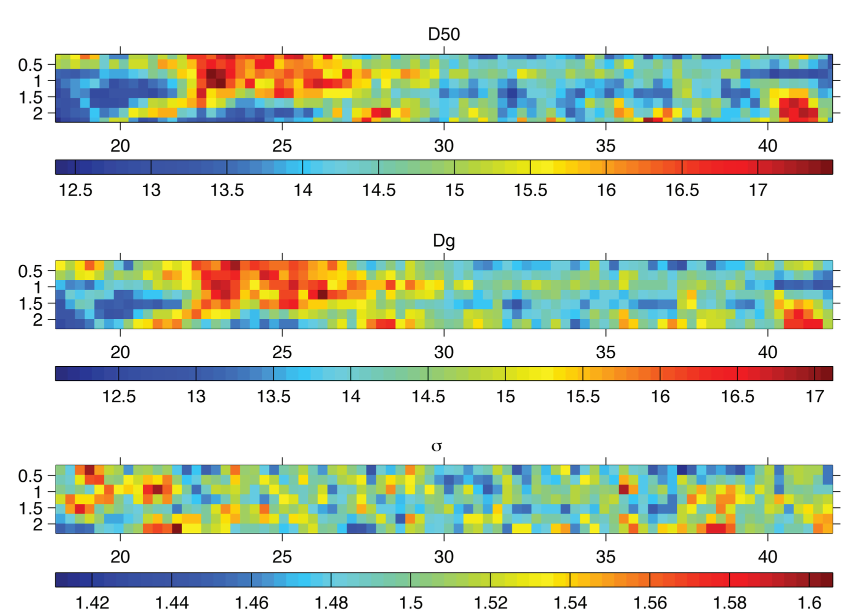
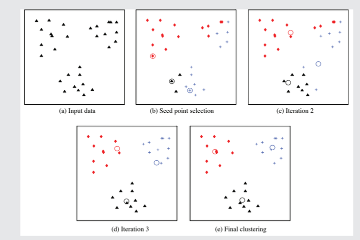


Figure 2.3. Maps of median grain size D_{50}, geometric mean grain size D_g, and geometric standard deviation σ, resulting from the automated image analysis. Scale bars for D_{50} and D_g are in mm.

3. Naive partitional clustering with *k*-means

Basic idea: Given a set of grain size distributions (GSDs), assign each one to a patch type based solely upon its distribution characteristics, then map the result.

How it works:
Iteratively assign distributions $\mathbf{f}_i$ to clusters c_α, with cluster mean μ_α, until the sum of squared errors over all k clusters is minimized: $\sum_{\alpha=1}^{k} \sum_{\mathbf{f}_i \in c_\alpha} \|\mathbf{f}_i - \mu_\alpha\|_2^2$



[from Jain, 2009]

Although this is the simplest possible method of delineating patches based on GSDs, the method tends to produce a quasi-random scattering of different patch types, suggesting spatial information must be included during classification.

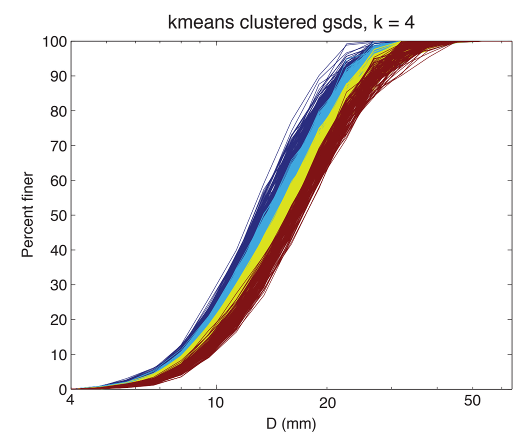


Figure 3.1. Cumulative distribution functions (CDFs) classified by *k*-means, with *k*=4. Colors correspond to the map in Figure 3.2.

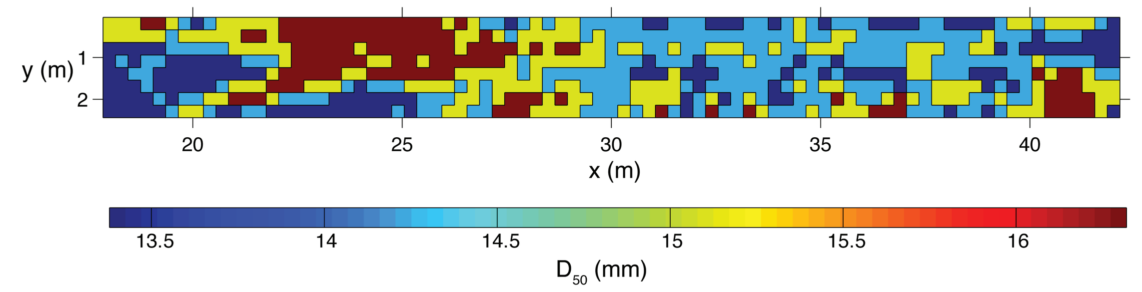


Figure 3.2. Map of *k*-means clustered GSDs corresponding to CDFs in Figure 3.1, for *k*=4. Color corresponds to patch-averaged D_{50}, as shown in the color bar.

4. Spatially-constrained agglomerative clustering

Basic idea: Initially assign each grain size distribution to its own cluster. Then merge the two *adjacent* clusters that are the most similar, and repeat until all distributions comprise the same cluster. This creates a hierarchy of clusters, which can be thresholded to produce a partition of the bed.

Here, the similarity d_{ij} between clusters i and j (with grain size distributions $\mathbf{f}_i$ and $\mathbf{f}_j$) is computed with the Minkowski distance of order p:

$$d_{ij} = \left\{ \sum_{k=1}^{N} |f_{i,k} - f_{j,k}|^p \right\}^{1/p} = \|\mathbf{f}_i - \mathbf{f}_j\|_p$$

Here, if $p = 1$ the distance is the "Manhattan" distance while if $p = 2$ it is the Euclidean distance. In this exercise we use $p = 1$.

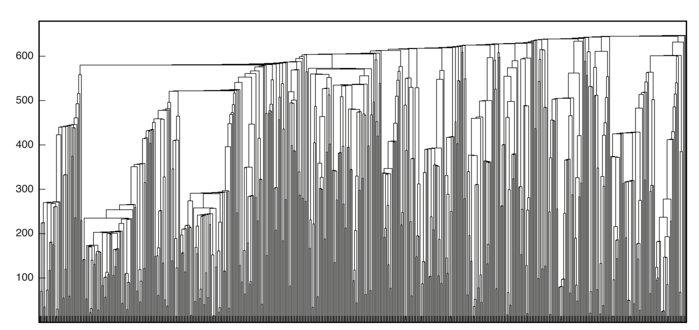
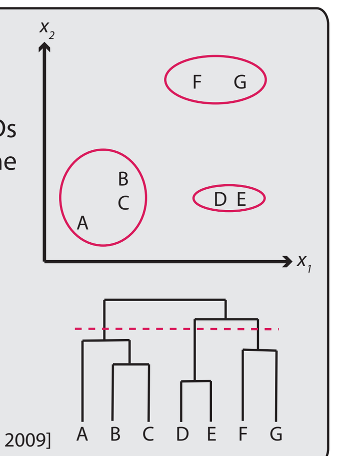


Figure 4.1. Dendrogram from spatially constrained agglomerative clustering. Individual GSDs lie on the bottom axis, the heights of branches signify the order of clustering.

How it works:
1. Assign each GSD to its own cluster.
2. Calculate the similarity between clusters d_{ij}
3. Merge the two adjacent (8-neighborhood) GSDs that are the most similar to form a new cluster. The GSD of the new cluster is the area-weighted average of its components.
4. Repeat steps 2-3 until all GSDs are in the same cluster.

This produces a dendrogram showing the hierarchy of merges, which is thresholded to produce a final partitioning of the data.



[Figure adapted from Jain, 2009]

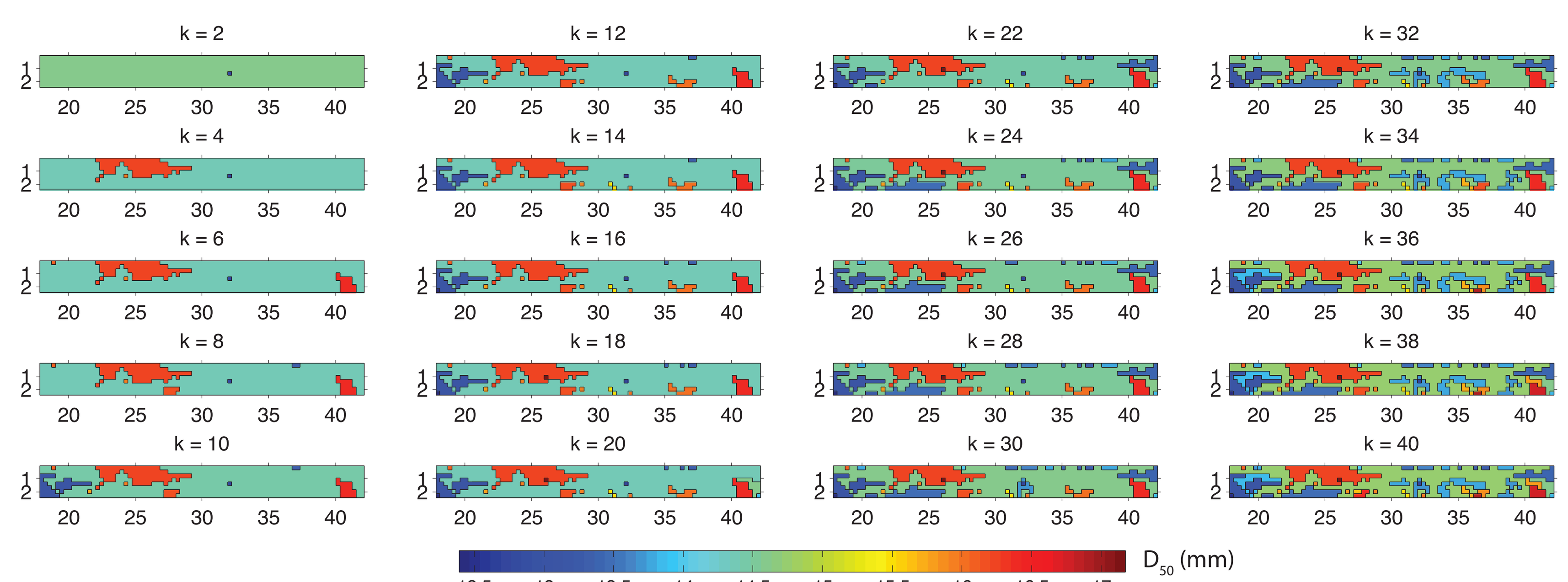


Figure 4.2. Patch maps from agglomerative clustering for *k*=2 to 40 total clusters. Patch color corresponds to the patch-averaged D_{50} shown on the color scale.

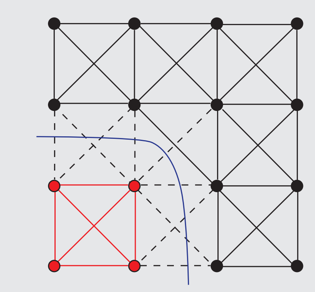
5. Spectral clustering using Normalized Cuts

Basic idea: Each GSD is a node on a graph, connected to other GSDs with edges with weights that correspond to the similarity between the GSDs. Edges with low weights are selectively cut so that within-cluster similarity is maximized while between-cluster similarity is minimized.

How it works:
Normalized Cuts [Shi and Malik, 2000] minimizes the cost function:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

In the weighted graph on the right, the blue line segments the graph into *A* (red nodes) and *B* (black nodes). *cut*(*A*,*B*) is the sum of the dashed (removed) edges, *assoc*(*A*,*V*) is the sum of the red edges and dashed edges, and *assoc*(*B*,*V*) is the sum of the black edges and dashed edges.



Method:
1. Compute the edge similarity weights w_{ij}. We use a Gaussian:

$$w_{ij} = \begin{cases} \exp\left(\frac{-d_{ijnorm}}{\sigma_1}\right) \cdot \exp\left(\frac{-\|\mathbf{X}_i - \mathbf{X}_j\|_2}{\sigma_2}\right) & \text{if } \|\mathbf{X}_i - \mathbf{X}_j\|_2 < r \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{X}_i$ are the (x,y) coordinates of the ith node, r is a threshold Euclidean distance beyond which similarity is forced to be zero, d_{ijnorm} is the Minkowski distance between grain size distributions normalized to scale between 0 and 1, and σ_1 and σ_2 are parameters that scale the relative importance of the differences between grain size distributions and the physical distance.
2. Summarize similarity in the matrices **W** and **D**, where **W**(i,j) = w_{ij} and **D** is a diagonal matrix with **d** on the diagonal where $\mathbf{d}(i) = \sum_j w(i,j)$
3. Solve (**D** - **W**)**x** = λ**Dx** for eigenvectors **x** with the smallest eigenvalues λ.
4. The eigenvector with the second-smallest eigenvalue is used to bipartition the graph by finding a splitting point such that *Ncut* is minimized.
5. The segmented parts of the dataset are recursively partitioned until a prespecified number of segments is achieved.

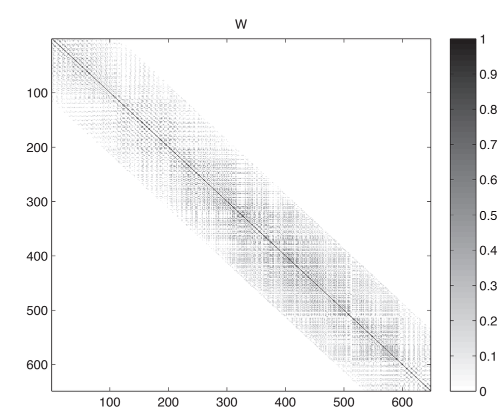


Figure 5.1. Similarity matrix **W** computed with r = 10 nodes, σ_1 = 0.008, σ_2 = 10, and p = 1 (Manhattan distance between GSDs).

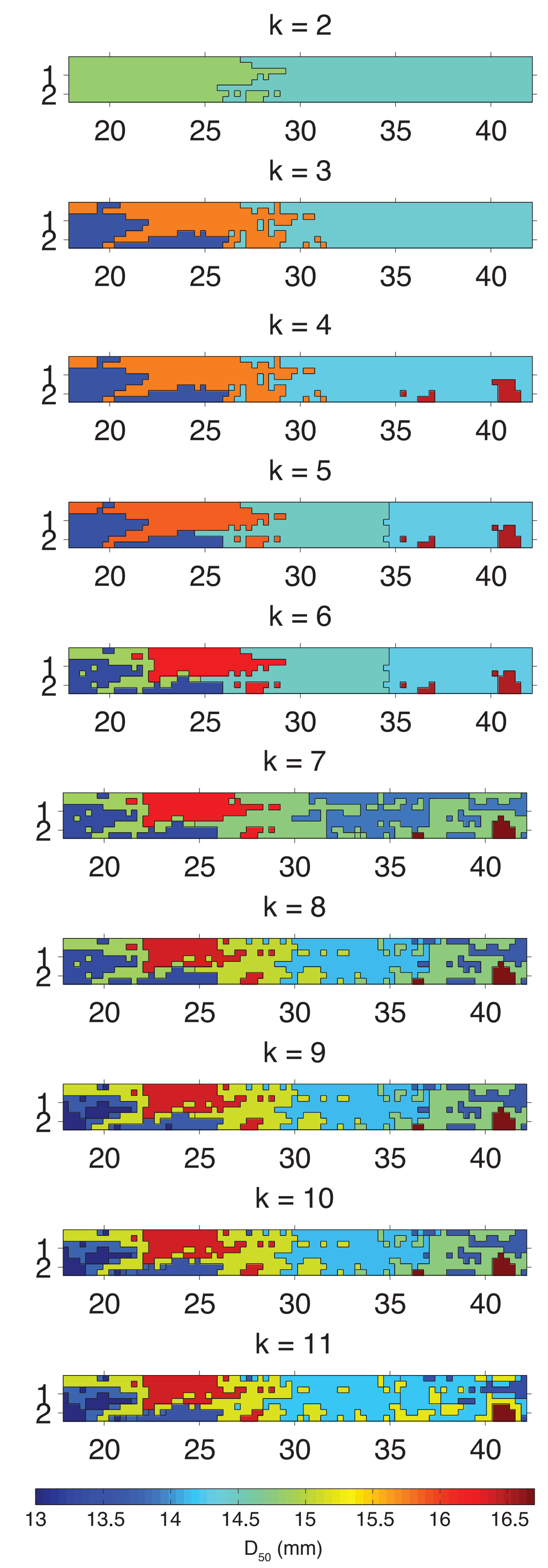


Figure 5.2. Bed discretization from *Ncut* using the similarity matrix shown in Fig. 5.1, for *k*=2 to 11 patches. Patch color corresponds to patch-averaged D_{50} as shown in the color scale.

6. Discussion and Conclusions

The Kolmogorov-Smirnov test can be used to compare the patch-averaged GSDs produced by the clustering algorithms to test whether they come from the same underlying distribution.

The statistical test suggests that there is a tendency for the channel to form a **finite number of patch types** that become distributed throughout the reach.

Clustering techniques can use information about spatial location and GSD characteristics to produce maps of bed surface patches objectively, which can then be used to inform assessments of habitat, sediment transport, and channel morphology.

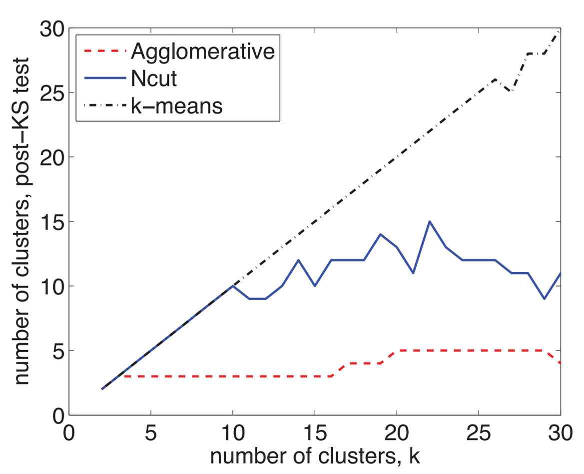


Figure 6.1. Number of statistically-different patch types resulting from the application of the two-sample Kolmogorov-Smirnov test to all pairs of clusters resulting from each clustering technique for a given value of *k*. Tests were performed at an $\alpha = 0.01$ significance level.

References:
Dietrich, W.E., P.A. Nelson, E. Yager, J.G. Venditti, M.P. Lamb, and L. Collins (2005), Sediment patches, sediment supply, and channel morphology, in *RCEM 2005*, edited by G. Parker and M.H. Garcia, pp. 79-100, Balkema, Rotterdam.
Graham, D.J., I. Reid, and S.P. Rice (2005), Automated sizing of coarse-grained sediments: image processing procedures, *Mathematical Geology*, *37*(1), 1-28.
Jain, A.K. (2009), Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, *31*, 651-666, doi:10.1016/j.patrec.2009.09.011.
Nelson, P.A., W.E. Dietrich, and J.G. Venditti (2010), Bed topography and the development of forced bed surface patches, *Journal of Geophysical Research, 115*, F04024, doi:10.1029/2010JF001747.
Shi, J., and J. Malik (2000), Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888-905.

Acknowledgements:
Funding was provided by the National Center for Earth-surface Dynamics (NCED) and a National Science Foundation Graduate Research Fellowship.