the closeness of the estimated (i.e. interpolated) values $\hat{y}_i$ to the values y_i observed at sites $Ø_i$. Four easy-to-use measures are:

- The mean absolute error: $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$;
- The mean squared error: $MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$;
- The Euclidean distance: $D_1 = \sqrt{\sum_i (y_i - \hat{y}_i)^2}$;
- The correlation coefficient (r) between values y_i and $\hat{y}_i$ (eq. 4.7). In the case of a trend-surface model, the square of this correlation coefficient is the coefficient of determination of the model.

In the case of kriging, the above measures of fit cannot be used because the estimated and observed values are equal at all observed sites $Ø_i$. The technique of cross-validation can be used instead (Isaaks & Srivastava, 1989, their Chapter 15). One observation, say $Ø_1$, is removed from the data set and its value is estimated using the remaining points $Ø_2$ to $Ø_n$. The procedure is repeated for $Ø_2$, $Ø_3$, …, $Ø_n$. One of the measures of fit described above may be used to measure the closeness of the estimated to the observed values. If replicated observations are available at each sampling site (a situation that does not often occur), the F-test of goodness-of-fit described in Subsection 13.2.1 can be used with all interpolation methods.

13.3 Patches and boundaries

Multivariate data may be condensed into spatially-constrained clusters. These may be displayed on maps, using different colours or shades. The present section explains how clustering algorithms can be constrained to produce groups of spatially contiguous sites; study of the boundaries between homogeneous zones is also discussed. Prior to clustering, one must state unambiguously which sites are neighbours in space; the most common solutions to this problem are presented in Subsection 13.3.1.

1 — Connection networks

When sampling has been conducted on a regular rectangular grid, neighbouring points may be linked using simple connecting schemes whose names are derived from the game of chess (Cliff & Ord, 1981): rook's (rectangular: Fig. 13.21a), bishop's (diagonal: Fig. 13.21b), or king's connections (also called queen's: both rectangular and diagonal, Fig. 13.21c). Sampling in staggered rows leads to connecting each point

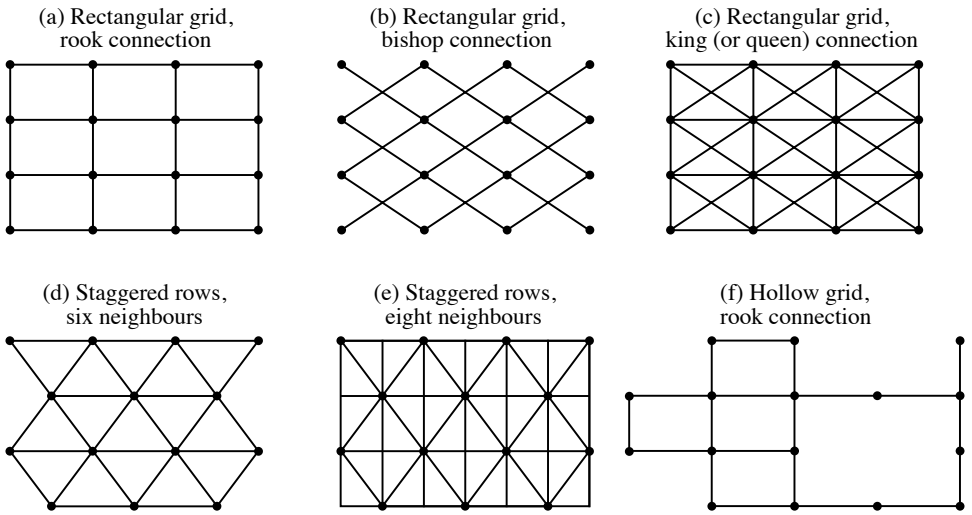


Figure 13.21 Connecting schemes for regular grids of points. See text.

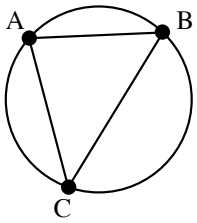
(except borders) to six (Fig. 13.21d) or eight neighbours (Fig. 13.21e). Algorithms may allow the construction of regular grids with missing points (Fig. 13.21f). When the objects represent irregularly-shaped land units covering a geographic area (e.g. types of ecosystems in a nature reserve), parcels sharing a common boundary are regarded as contiguous.

When the sites are positioned in an irregular manner, one can use geometric connecting schemes such as Delaunay triangulation, Gabriel graph, relative neighbourhood graph or minimum spanning tree, described below. There exists an inclusion relationship among these four connecting schemes: all edges that are members of a minimum spanning tree (MST) also obey the relative neighbourhood graph criterion; these are all members of a Gabriel graph, which in turn are all included in a Delaunay triangulation (Toussaint, 1980; Matula & Sokal, 1980; Gordon, 1996c):

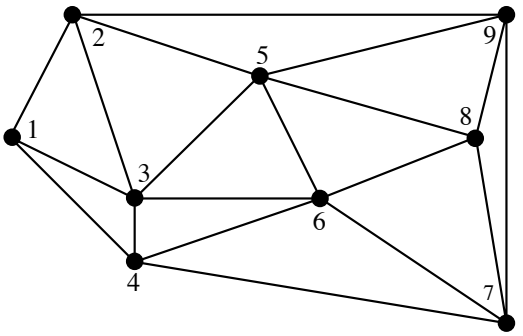
$$\text{MST} \subseteq \text{Relative neighbourhood graph} \subseteq \text{Gabriel graph} \subseteq \text{Delaunay triangulation}$$

Delaunay triangulation

• Delaunay triangulation. — The Delaunay triangulation criterion (Dirichlet, 1850; Upton & Fingleton, 1985) is illustrated in Fig. 13.22. For any triplet of points A, B and C, the three edges (i.e. lines) connecting these points are included in the triangulation if and only if the circumscribed circle (i.e. the circle passing through the three points; on the left in the figure) includes no other point. For example, the file of coordinates shown in the central part of the figure gives rise to the triangulation on the right. The triangulation is fully described by a list of pairs of points corresponding to its edges; this is how the information can be passed on to a computer program for space-constrained clustering (Subsection 13.3.2).



Point identifiers	Coordinates X	Y
1	0	3
2	1	5
3	2	2
4	2	1
5	4	4
6	5	2
7	8	0
8	7.5	3
9	8	5



19 edges form the Delaunay triangulation:

1–2	1–3	1–4	2–3	2–5	2–9	3–4
3–5	3–6	4–6	4–7	5–6	5–8	5–9
6–7	6–8	7–8	7–9	8–9		

Figure 13.22 Construction of a Delaunay triangulation for 10 points.

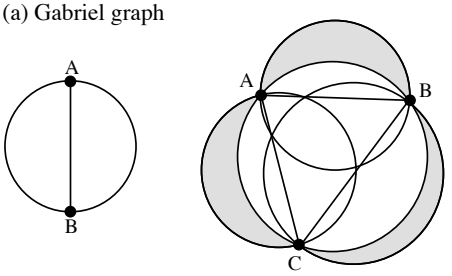
Long edges may be created at the outskirts of a set of points, simply because there is no other point located farther away in the sampling design; this is called a *border effect*. For example, edges 2–9 and 7–9 could have been removed from the triangulation in Fig. 13.22 by the presence of other points in the circumscribed circles of triangles (2, 5, 9) and (7, 8, 9) had the sampling extent been broader. Long peripheral edges can be removed by hand from the list, or by the computer algorithm.

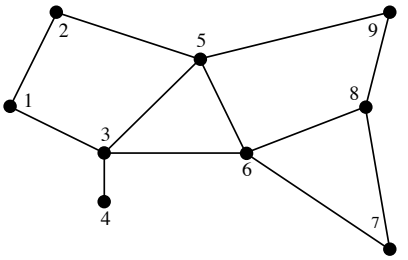
Gabriel graph

• Gabriel graph. — The Gabriel graph criterion (Gabriel & Sokal, 1969) differs from that of the Delaunay triangulation (Fig. 13.23a). Draw a line between two points A and B. This line is part of the Gabriel graph if and only if no other point C lies inside the circle *whose diameter is that line*. In other words, the edge between A and B is part of the Gabriel graph if $D^2(A, B) < D^2(A, C) + D^2(B, C)$ for all other points C in the study, where $D^2(A, B)$ is the square of the geographic distance between points A and B. Another way of expressing this criterion is the following: if CENTRE represents the middle point between A and B, the edge connecting A to B is part of the Gabriel graph if $D(A, B)/2 < D(\text{CENTRE}, C)$ for any other point C in the study.

The Gabriel graph in Fig. 13.23a is constructed for the same points as the Delaunay triangulation in Fig. 13.22. The 12 edges forming the Gabriel graph are a subset of the 19 edges of the Delaunay triangulation. Indeed, as shown by the sketch in the centre of the figure, the exclusion zone formed by the three circles corresponding to the Gabriel criterion (which have for diameters the edges A–B, B–C and A–C) may contain, in the shadowed areas outside the Delaunay circle (white inner circle), some points that the

(a) Gabriel graph

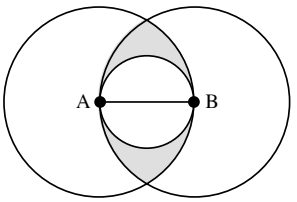


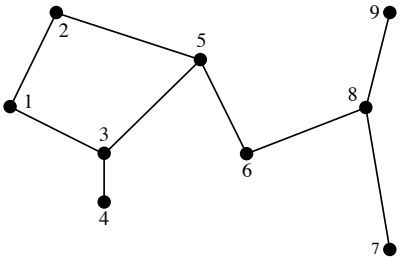


12 edges form the Gabriel graph:

1–2 1–3 2–5 3–4 3–5 3–6
5–6 5–9 6–7 6–8 7–8 8–9

(b) Relative neighbourhood graph





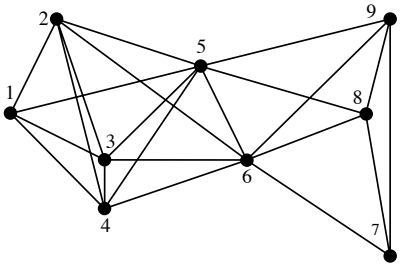
9 edges form the relative neighbourhood graph:

1–2 1–3 2–5 3–4 3–5 5–6
6–8 7–8 8–9

(c) Maximum distance graph

Criterion: $D \leq$ threshold

In this example, $D \leq 5$



22 edges form the maximum distance graph:

1–2 1–3 1–4 1–5 2–3 2–4 2–5 2–6
3–4 3–5 3–6 4–5 4–6 5–6 5–8 5–9
6–7 6–8 6–9 7–8 7–9 8–9

Figure 13.23 (a) Left: geometric criterion for the Gabriel graph. Centre: the zone of exclusion of the criterion, here for three points (grey zones + white inner circle), is larger than that of the Delaunay criterion (white inner circle). Right: graph for the example data, containing 12 edges. (b) Left: geometric criterion of the relative neighbourhood graph. The zone of exclusion of the criterion, here for two points (grey zones + white inner circle), is larger than that of the Gabriel criterion (white inner circle). Right: graph for the example data, containing 9 edges. (c) Left: criterion of the maximum distance graph. Right: graph for the example data with $D \leq 5$, with 22 edges.

Delaunay criterion circle does not exclude. This is why some edges that are authorized by the Delaunay criterion are excluded from the Gabriel graph.

Relative neighbourhood graph

• Relative neighbourhood graph. — The relative neighbourhood criterion is as follows (Toussaint, 1980; Fig. 13.23b). Draw a line between two points A and B. Draw a first circle centred over A and a second one centred over B, each one having the line from A to B as its radius. This line is part of the graph if no other point C in the study lies *inside the intersection of the two circles*. Points that fall on the circumference of one of the circles in the intersection zone do not count. In algebraic terms, the edge from A to B is part of the relative neighbourhood graph if and only if $D(A, B) \leq \max [D(A, C), D(B, C)]$ for all other points C in the study. For points forming an equilateral triangle, for instance, the three edges are included in the relative neighbourhood graph.

The relative neighbourhood graph in Fig. 13.23b is constructed for the same set of points as in Figs. 13.22 and 13.23a. The 9 edges forming the relative neighbourhood graph are a subset of the 12 edges of the Gabriel graph. Indeed, as shown by the sketch on the left of the figure, the exclusion zone at the intersection of the two circles corresponding to the relative neighbourhood criterion (which have for radius the edge A–B) may contain, in the shadowed zone outside the Gabriel circle (white inner circle), some points that the Gabriel criterion circle does not exclude. This is why some edges authorized by the Gabriel criterion are excluded from the relative neighbourhood graph.

Maximum distance graph

• Maximum distance graph. — Another strategy is to select a distance threshold and connect all points that are within that distance of each other. The result is called a maximum distance graph or an influence circle graph (Fig. 13.23c). One possible criterion to choose the distance threshold is to make it equal to the range of a variogram model (Fig. 13.7) computed for univariate (Subsection 13.1.3) or multivariate response data (Subsection 13.1.4).

Minimum spanning tree (MST)

• Minimum spanning tree (MST). — This tree connects the n points in the study with $(n - 1)$ edges. The sum of the weights (i.e. distances) of the edges used in the tree is minimum, meaning that it is smaller than or equal to the sum of the edge weights of any other tree connecting these n objects. Its construction is described at the end of Section 8.2; one way of obtaining it is to list the edges forming the primary connections of a single-linkage dendrogram. For points forming an equilateral triangle, for example, only two of the edges are included in the minimum spanning tree, whereas the three edges are included in a relative neighbourhood graph; the choice of the edge to leave out is arbitrary. The edges of a minimum spanning tree are either the same as, or a subset of, the edges of a relative neighbourhood graph of the same points. The minimum spanning tree for the example data set is shown in Fig. 14.3.

The list of connecting edges (Figs. 13.22 and 13.23) may be written out to a file. The file may be modified to take into account other information that researchers may have about the study area. For example, one may wish to eliminate edges that do not make sense in terms of gene flow because they cross unsuitable areas (e.g. a sea or a

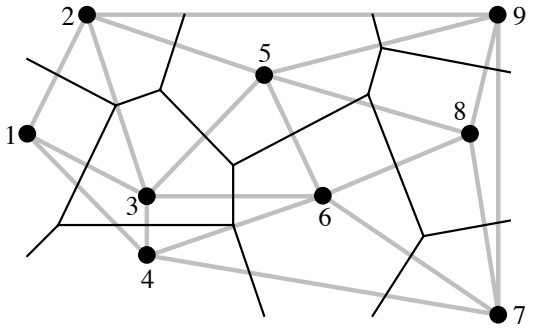


Figure 13.24 Delaunay triangulation (grey lines) and influence polygons (black lines) for the nine points of Fig. 13.22.

mountain range, in the case of terrestrial mammals). Or, one may wish to add connections that are potentially of interest although they do not imply first neighbours; for example, plants or animals may be able to cross water bodies (lake, sea) and settle in non-contiguous sites, which should be considered contiguous for the analysis because there is a direct path between them. Users of constrained clustering methods should not hesitate to modify lists of connections obtained from geometric criteria such as described above, to make the list of edges a better description of potential flow among sites, given the problem under study.

Influence polygons

It is sometimes interesting to determine the geometric zone of influence of each point on a map. The zone of influence of a point A includes all the other points of the surface that are closer to A than to any other point in the study. The zones of influence so defined have the shape of polygons, also called tiles, tessellae, or tesserae (singular: tessella or tessera). The resulting picture is called a mosaic or tessellation (adjective: tessellated); it is also referred to as a Dirichlet tessellation (1850), Voronoï polygons (1909), or Thiessen polygons (1911), from the names of the authors who described these mathematical structures.

Polygons are easily constructed from a Delaunay triangulation (Fig. 13.24). Draw the perpendicular bisector of each segment in the triangulation; the crossing points of the bisectors delimit the polygons (tiles). Computer algorithms may be used to calculate the surface area of each polygon, at least those that are closed; peripheral tiles may be open. Upton & Fingleton (1985) and Isaaks & Srivastava (1989) propose various applications of tessellations to spatial analysis.

2 — Space-constrained clustering

The delineation of clusters of contiguous objects has been discussed in Section 12.6 for time series and spatial transects. The method of chronological clustering, in

particular, was described in Subsection 12.6.4; it proceeds by imposing to a clustering algorithm a constraint of contiguity along the time series. Constraints of contiguity have been applied to spatial clustering by several authors, including Lefkovitch (1978, 1980), Monestiez (1978), Lebart (1978), Roche (1978), Perruchet (1981) and Legendre & Legendre (1984c). In the present subsection, it is generalized to two- or three-dimensional spatial data and to spatio-temporal data.

Constrained clustering differs as follows from its unconstrained counterpart:

- Unconstrained clustering methods (Chapter 8) only use the information in the similarity or distance matrix computed among the objects. In hierarchical methods, a local criterion is optimized at each step; in all methods included in the Lance and Williams general model, for instance, the objects or groups clustered at each step are those with the smallest fusion distance or the largest fusion similarity. In partitioning methods, a global criterion is optimized; in *K*-means, for instance, the algorithm looks for *K* groups that feature the smallest sum of within-group sums-of-squares E_K^2;

- Constrained clustering methods take into account more information than the unconstrained approaches. In the case of spatial or temporal contiguity, the only admissible clusters are those that obey the contiguity relationship. Spatial contiguity may be described by one of the connecting schemes of Subsection 13.3.1. The criterion to be optimized during clustering is relaxed to give priority to the constraint of spatial contiguity. It is no surprise, then, that a constrained solution may be less optimal than its unconstrained counterpart in terms of the clustering criterion, e.g. E_K^2. This is balanced by the fact that the resulting clusters are likely to more readily interpretable.

It is fairly easy to modify clustering algorithms to incorporate a constraint of spatial contiguity (Fig. 13.25). As an example, consider the clustering methods included in the Lance and Williams general agglomerative model (Subsection 8.5.9). At the beginning of the clustering process, the vector of group membership has each object in a different group (Fig. 13.25, right). Proceed as follows:

1. Compute a distance matrix (**D**) among objects using the non-geographic information. Turn it into a similarity matrix **S** using one of the equations of Subsection 7.2.1. This transformation will make step 3 of the procedure possible.

2. Choose a connecting scheme (Subsection 13.3.1) and produce a list of connection edges as in Figs. 13.22 and 13.23. Read in the file of edges and transform it into a *contiguity matrix* containing 1's for connected sites and 0's elsewhere.

3. Compute the Hadamard product of these two matrices, i.e. their product element by element (Section 2.5). The resulting matrix contains similarity values in the cells where the contiguity matrix contained 1's, and 0's elsewhere.

4. The largest similarity value in the matrix resulting from step 3 determines the next pair of objects or groups (*h* and *i*) to be clustered. Modify the vector of group

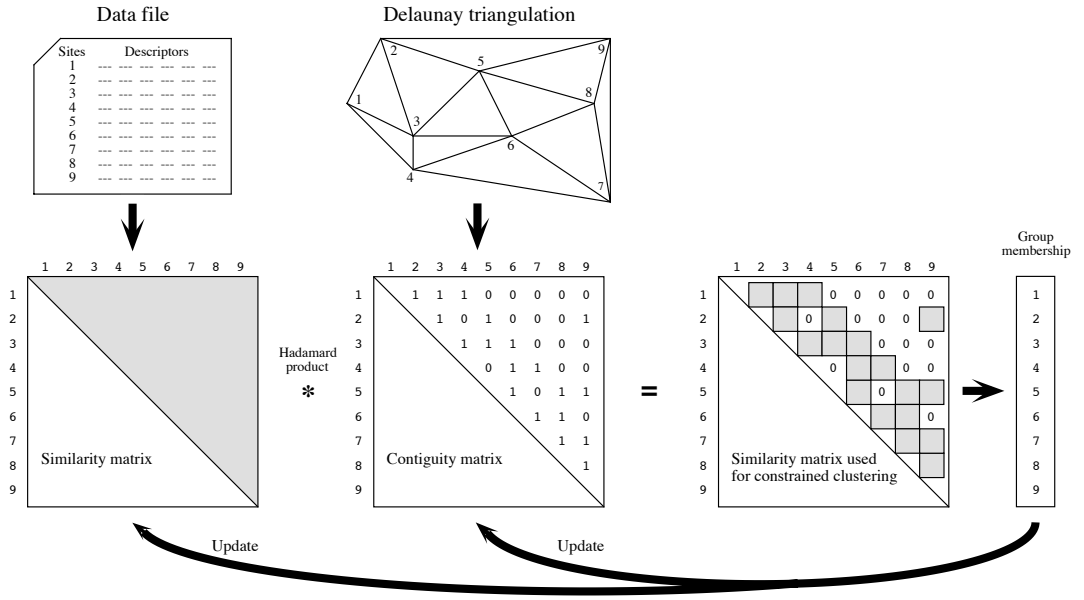


Figure 13.25 Summary of the spatially-constrained clustering procedure for methods included in the Lance and Williams general clustering model. The vector of group membership is represented on the right; at the start of the clustering process, each object is in a different group (numbers 1 to 9 in the example). Locations of the points are the same as in Fig. 13.22.

membership (right of the figure), giving the same group label to all members of former groups *h* and *i*.

5. Update the similarity matrix using eq. 8.12.

6. Update also the contiguity matrix. All objects that were neighbours to *h* are now also neighbours to *i* and vice versa.

7. Go back to step 3. Iterate until all objects are members of a single group.

8. Determine the most informative number of clusters, either by visual inspection of space-constrained clustering maps, or after calculating one of the indices mentioned at the end of Section 8.8 (available in R function ***clustIndex()*** of package CCLUST). Among those indices, the Calinski-Harabasz criterion was recommended by Gordon (1999) for constrained clustering. Pawitan & Huang (2003) proposed a permutation procedure to test the significance of successive partition levels in constrained clustering. Cross-validation seems another promising way of identifying the most informative partition in constrained clustering.

Ferligoj & Batagelj (1982) showed that the introduction of relational constraints (e.g. spatial contiguity) may occasionally produce reversals with any of the hierarchical clustering methods included in the Lance & Williams algorithm (Subsection 8.5.9], except complete linkage. Additional constraints may be added to the algorithm, for example to limit the size or composition of any group (Gordon, 1996c). *K*-means partitioning algorithms (Section 8.8) can also be constrained by the contiguity matrix shown in Fig. 13.25.

Space-constrained clustering is useful in a variety of situations. Here are some examples.

- In many studies, there are compelling reasons to force the clusters to be composed of contiguous sites; for instance, when delineating ecological regions, administrative units, or resource distribution networks.

- One may wish to relate the results of clustering to geographically-located potential causal factors that are known to be spatially autocorrelated, e.g. geological data.

- One may wish to cluster sites based upon environmental variables, using a constraint of spatial contiguity, in order to design a stratified biological sampling program to study community composition.

- To test the hypothesis that neighbouring sites are ecologically similar, one may compare unconstrained and constrained clustering solutions using the modified Rand index (Subsection 8.12.2). De Soete *et al*. (1987) give other examples where such comparisons may help test hypotheses in the fields of molecular evolution, psycholinguistics, cognitive psychology and evolution of languages.

- Constrained solutions are less variable than unconstrained clustering results, which may differ in major ways among clustering methods. Indeed, the constraint of spatial contiguity reduces the number of possible solutions and forces different clustering algorithms to converge onto largely similar clusters (Legendre *et al*., 1985).

Constrained clustering can also be used for three-dimensional or spatio-temporal sampling designs (e.g. Planes *et al*., 1993). As long as the three-dimensional or spatio-temporal contiguity of the observations can be accurately described as a file of edges as in Figs. 13.22 and 13.23, constrained clustering programs have no difficulty in computing the solution; the only difficulty is the representation of the results as three-dimensional or spatio-temporal maps. Higher-dimensional extensions of the geometric connecting schemes presented in Subsection 13.3.1 are available in the literature. In addition, space-constrained clustering can be used to detect discontinuities in spatial transects or time series, a topic that has been discussed in Section 12.6.

Legendre (1987b) suggested a way of introducing spatial proximity into clustering algorithms which is less stringent than the methods described above. The method consists in weighting the values in the ecological similarity or distance matrix by some function of the geographic distances among points, before clustering. The idea was

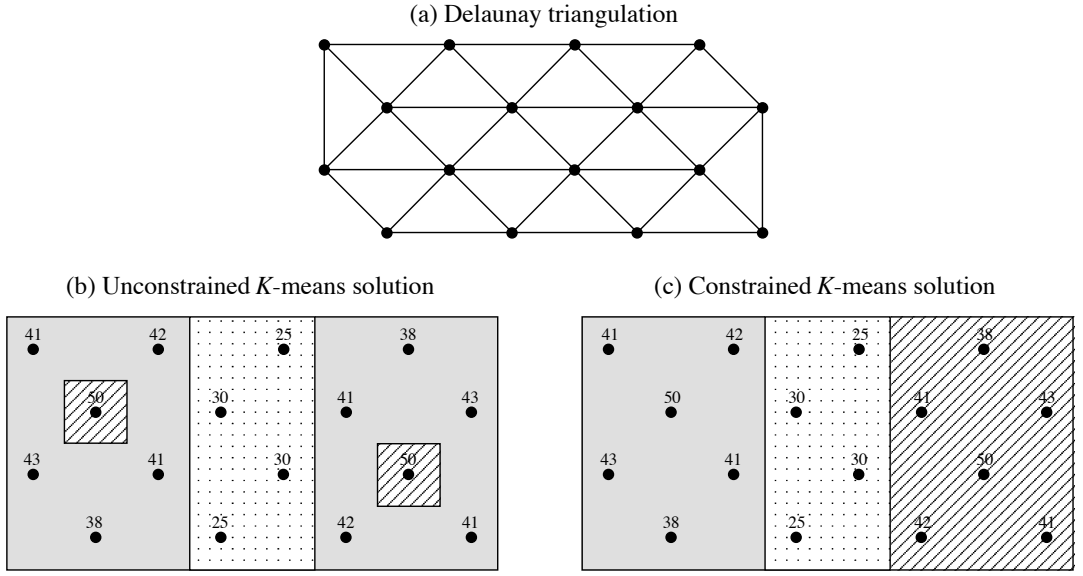


Figure 13.26 Numerical example showing the difference between (b) unconstrained and (c) constrained clustering solutions. (a) Delaunay triangulation with 35 edges, which were used as constraint in (c). The values of the artificial variable are given in panels (b) and (c); the three groups obtained by unconstrained and constrained *K*-means are identified by shadings.

implemented by Bourgault *et al*. (1992) who proposed to use a multivariate variogram or covariogram as spatial weighting function prior to clustering. Large ecological distances between sites that are close in space are downweighted to some extent by this procedure. It is then easier for clustering algorithms to incorporate somewhat diverging sites into neighbourhood clusters. Oliver & Webster (1989) suggested to use a univariate variogram for the same purpose. Constrained classification methods were reviewed by Gordon (1996c, 1999) and algorithms were surveyed by Murtagh (1985). Formal aspects were discussed by Ferligoj & Batagelj (1982, 1983). Generalized forms of constrained clustering were described by De Soete *et al*. (1987).

Numerical example. An artificial set of 16 sites was constructed to represent staggered-row sampling of a distribution with two peaks. From the geographic positions of the sites, a Delaunay triangulation (35 edges) was computed (Fig. 13.26a). A single variable was attributed to the sites. For three groups, the unconstrained *K*-means solution has a sum of within-group sums-of-squares $E^2_K = 53$ (Fig. 13.26b). The constrained *K*-means solution, for three groups, has a value $E^2_K = 188$ (Fig. 13.26c) which is higher than that of the unconstrained solution, for reasons explained above. The two partitions are interesting in different ways. The unconstrained solution identifies sites with similar values, whereas the constrained solution brings out the two regions with high values plus a region with lower values forming a valley between the peaks.