ORIGINAL PAPER

# Clustering forest harvest stands on spatial networks for optimised harvest scheduling

**Thomas Smaltschinski · Ute Seeling · Gero Becker**

**Abstract**

· *Context* Clustering forest harvest stands reduces the movements of the harvesters, forwarders and staff. Moreover, it simplifies the subsequent log transport, when compared with dispersed stands.

· *Aims* Harvesting activities are generally based on silvicultural motivated planning data. The development of an analytical method to cluster harvest stands with respect to the spatial network of roads should improve the harvesting effort.

· *Material and method* The clustering of harvest stands was developed for Aracruz (Brazil) in 2004, where it is used there successfully since. The hierarchical method 'single linkage cluster analysis' is applied. As a distance function, the Euclidian distance was substituted by the shortest path on the spatial network.

· *Result* The clustering method is based on the minimum spanning tree, which is the spatial equivalent to the dendrogram of an ordinary cluster analysis. Applying the Delaunay triangulation to fill the distance matrix reduces the distance calculation time from $O(n^2)$ to $O(n)$. The method is illustrated by a planning district of the Aracruz enterprise.

· *Conclusion* Harvesting units are properly clustered spatially by the discussed method. Topographic obstacles are automatically avoided and the need to relocate machinery is reduced as well as the total driving distance.

**Keywords** Clustering harvest stands · Optimising harvest operations · Cluster analysis · Spatial network of roads

T. Smaltschinski (✉) · G. Becker
Institute for Forest Utilization and Work Science,
University of Freiburg,
Werthmannstr. 6,
79085 Freiburg, Germany
e-mail: thsm@gmx.de

G. Becker
e-mail: fobawi@fobawi.uni-freiburg.de

U. Seeling
German Centre for Forest Work and Technology (KWF),
Sprembergerstr. 1,
64823 Groß-Umstadt, Germany
e-mail: seeling@kwf-online.de

## 1 Introduction

The planning staff in a forest enterprise continuously attempts to devise methods for optimising the individual processes involved in timber production to ensure that these are as efficient as possible. This may relate to the operations themselves such as timber felling and log transport. The aim of this investigation is to develop an optimisation method for clustering the stands scheduled for harvesting during the planning phase.

Stands selected for harvest may be distributed irregularly across an entire forest area. However, the efficiency of timber harvesting, involving the organisation of machinery deployment and finally log transport, is significantly enhanced when harvest stands are clustered. By clustering the harvest stands the concentration of timber harvest activities saves time and costs.

A formal procedure to cluster harvest stands is often unnecessary in small forest enterprises. Here a forester successfully clusters stands according to distances estimated from forest maps or from local experience under simplified, comprehensible conditions. However, conventional methods become prone to error once the number of harvest stands increases.

INRA    ⌖ Springer

An analytical method to cluster harvest stands has not yet been described. The unfavourable topographic conditions and the great number of harvest stands at the Eucalyptus plantations of Aracruz, Brazil, stimulated the search for and the development of an approach for clustering harvest stands. The typical land use of this region is characterised by a complex spatial pattern of mutually nested plantations, gallery and alluvial forests. Rivers often separate the harvest stands. Despite Euclidian distances between the different stand locations, the actual road travel distances are much longer in order to circumnavigate the topographic obstacles (Fig. 1). The length of these routes affects the deployment of equipment and log transport considerably.

A cluster analysis method is presented in the following section, where the distances between harvest stands are the shortest paths on the spatial network of roads. Harvest stands are combined to a cluster if the distance between them is below a defined maximum distance. The size of this maximum distance depends on local conditions and the applied harvest technology. The method described was developed for the forest area at Aracruz, Brazil, in 2004.

The company provided the stand polygons, the line geometry for the forest roads and a list of stands scheduled for harvest. The data are represented as geodata sets of the geographic information system ArcGIS (ESRI). The implementation of the numerical methods as described later in this paper is a straightforward procedure using the programming interfaces of the extension technology of ArcGIS.

## 2 The development of cluster analysis on spatial networks

In the forest planning process, the selection of harvest stands is mainly motivated by customer demand or similar purposes and is usually carried out in a non-spatial way. Chen and von Gadow (2002) included the spatial distribution of harvest stands directly in the long-term planning process. Different stand treatment options result in different spatial patterns of thinning or clear-cut stands, with respect to regeneration areas, at a given point in time. Each spatial arrangement of a specific treatment option has an economic benefit. The combination of spatial arrangements and timber production is considered simultaneously, using the heuristic 'Simulated Annealing'.

With respect to a multi-objective long-term planning, Öhman and Lämas (2003) examined the spatial influence of clustered harvest activities. By clustering they identified several advantages such as habitat conservation and reduction of road maintenance and entry costs. In addition to this, dispersed stands are more expensive to harvest than clustered stands. The developed model is determined by two objectives: (1) maximizing the net present value from forest

management over an infinite time horizon and (2) maximising the clustering of harvest activities in space and time.

The introductory cluster procedures do not agree with an ordinary cluster analysis. Here clustering is only one weighted component among others of the whole objective function. Furthermore Euclidian distances are used as distance measure between the harvest stands. Obstacles, such as those shown in Fig. 1, can imply distorted clusters for operational planning. The movements of harvest equipment, piling and the subsequent wood transport are constrained by the spatial network of roads.

In a cluster analysis, events or data objects are arranged according to their relative similarity. A similarity cannot easily be expressed in numeric terms. It is usually substituted by a proximity measure, which is derived from metric parameters of the data objects. By clustering the data according to their closest proximity and by defining a cluster's proximity threshold, it is possible to obtain a defined maximum of mutual similarity.

For a given set of vectors $\{\mathbf{x}_1 \ldots \mathbf{x}_n\}$, the elements $\mathbf{x}_i$, $i=1\ldots n$, are coordinates of a multidimensional metric vector space of dimension $m$. The distance between two elements is denoted by $d(\mathbf{x}_i, \mathbf{x}_j)$ or more specifically $d_{ij}$. The distance measure between two coordinates has the general form (Steinhausen and Langer 1977: 61)

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)\right]^{1/r} = \left[\sum_{k=1}^{m} |x_{ik} - x_{jk}|^r\right]^{1/r} d_{ij}, r \in \mathbf{R} \tag{1}$$

For a plane of dimension $m=2$, the distance function (1) results in the Euclidian distance for $r=2$ and it corresponds to the 'taxi-driver'- or 'city-block'-distance for $r=1$. Distance functions similar to Eq. (1) with different values of $r$ usually do not represent a forest road network, where topographic obstacles require detours (Fig. 1). The shortest path between spatial locations $\mathbf{x}_i$ and $\mathbf{x}_j$ located on a spatial network of roads can be calculated by the Dijkstra algorithm (1959). For this purpose, the spatial network of roads must form a connected graph. The following distance function $dd$ is called the Dijkstra distance that corresponds to the shortest path between two points $(\mathbf{x}_i, \mathbf{x}_j)$ on the spatial network of roads

$$dd(\mathbf{x}_i, \mathbf{x}_j) = dd_{ij} \tag{2}$$

Generally the calculation procedure of a cluster analysis is independent of the chosen Eq. (1) or (2).

The analysis of objects or events on spatial networks was deduced from the analysis and description of point distributions in the 1950s. Clark and Evans (1954) investigated the Euclidian distances from points to nearest neighbours as a measure of spatial relationship in populations such as trees or seedlings in a forest. The distribution of points in a plane can

**Fig. 1** Map of a section of the Aracruz forest: forest stands denoted in *grey*, rivers in *black*, road systems in *light grey* and gallery and protected alluvial forests in *white*. The shortest path on the spatial network of roads between two stands is shown as an example in *bold black*
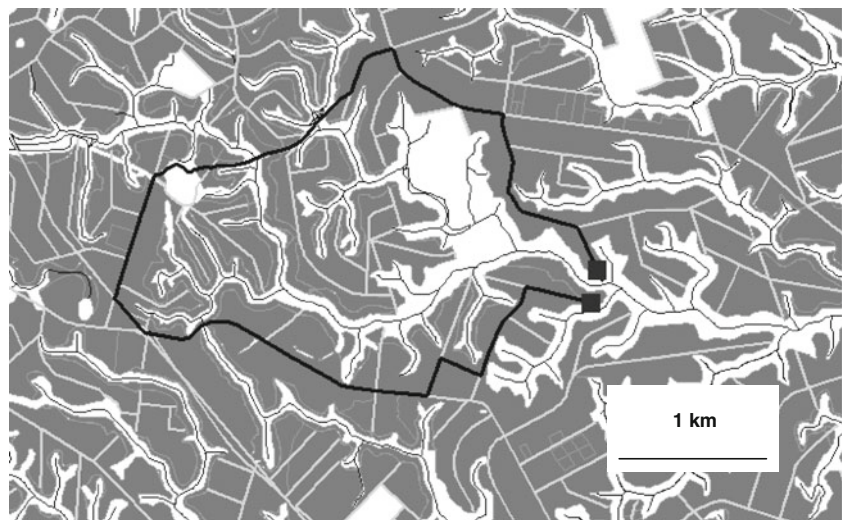
be regular, random, clustered or a combination these different distributions.

Applying the ideas from Clark and Evans (1954), Okabe et al. (1995) analysed points on spatial networks such as car accidents or the distribution of stores, using the shortest path (2) between points on the spatial network of roads. With respect to the distribution of the points (regular, random or clustered), Okabe and Kitamura (1996) developed a method for optimising the catchment areas of stores. Later Okabe and Yamada (2001) described the distribution of accidents on spatial networks using the 'K-function' derived by Ripley (1976) that more accurately describes spatial distributions of stationary point processes.

Urban and Keitt (2001) explored the habitat patches of the Mexican Spotted Owl (*Strix occidentalis lucida*) by 'single linkage cluster analysis' using Euclidian distances (1). The authors refereed to the application of the distance function (2) when considering transportation networks or routing applications. Yiu and Mamoulis (2004) presented a cluster analysis on a spatial network of roads using (2) as distance function and were able to identify clusters of objects, points or events. They analysed three clustering methods: partitioning-based clustering by the 'k-Medoid' method, the '∈-Link' algorithm by local selection of neighbouring points with a chosen search radius ∈ and the hierarchical 'single linkage cluster analysis' method. For the analysed spatial networks the 'k-Medoid' method was not effective, the '∈-Link' algorithm traverses the whole network only with an appropriate search radius, but the method 'single linkage cluster analysis' led to correct clusters. Okabe et al. (2006) integrated these methods in the software package SANET (Software for spatial Analysis on a Network).

Sughihara et al. (2008; 2011) transferred the approach of Yiu and Mamoulis (2004) to other hierarchical methods of cluster analysis, e.g. the closest-pair distance, the farthest-pair distance, the diameter distance, the average-pair distance, the median-pair distance and radius distance. The authors pointed

out three differences to a conventional cluster analysis with a distance Eq. (1):

– An object or event is tied to a geographic location,
– a measure of distance between events agrees to Eq. (2) and
– a cluster analysis of events on networks is oriented to geometric computations and is not determined by algebraic operations on attribute vectors.

They assumed that the spatial network of roads is planar, the number of network nodes largely exceeds the number of events and the distances between clusters X, Y and Z fulfil the following equation: $\min\{d(X,Z), d(Y,Z)\} \leq d(X \cup Y, Z)$.

There are many other studies for the analysis of spatial networks of roads. In these studies, rather the pattern of point distributions on the network has been studied as the formation of clusters (e.g. Yamada and Thill 2004; Okabe et al. 2006; Scheurer and Curtis 2008; Shiode and Shiode 2009; Dai et al. 2010).

## 3 Applied method

The harvest stands should be grouped into clusters in a way that the distances between stands within a cluster do not exceed a defined *maximum distance*. This maximum distance depends on the costs for moving the harvesting machines between the stands and on the road trafficability. Following the suggestions of Yiu and Mamoulis (2004) the hierarchical 'single linkage cluster analysis' method has been applied. Using this method the proximity measure is determined by the distances between the harvest stands with respect to the spatial network.

The 'distance between stands' is an ambiguous measure. The spatial network of roads corresponds to an undirected weighted graph consisting of edges and nodes. The Dijkstra distance (2) can be calculated only if its start and end points

lie on the edges of the graph. The harvest stands are polygons without connection to the graph. Therefore, *stand identifiers* are calculated as points, which represent the harvest stands and lie on the edges of the graph according to Okabe et al. (2006). For this purpose, first all edges adjacent to a given harvest stand are selected. After that, the point of the closest edge with shortest distance to the stand's centroid is taken as stand identifier and is allocated the same identification as the stand. The stand identifiers are simply referred to as $x_i$, $i=1…n$, where the selected harvest stand is denoted by the index i.

Generally, the distances between all stand identifiers are to be stored in the distance matrix D=$(dd_{ij})$ with $n(n-1)/2$ independent values. The computing time to fill the matrix D increases by $O(n^2)$ (Yiu and Mamoulis 2004). Because of this quadratic relation, the distance matrix grows rapidly with increasing number of harvest stands. Therefore, the method is not very feasible for any kind of application with a substantial number of harvest stands.

By using the 'single linkage cluster analysis' method it is possible to reduce the just quoted amount of distance calculation. The dendrogram as the result of this method is a subset of the geometries from the Delaunay triangulation using Euclidian distances (1) (Steinhausen and Langer 1977) and is equivalent to the *minimum spanning tree* (Gower and Ross 1969; Urban and Keitt 2001; Kruskal 1956). This binary tree connects all the harvest stands with a minimum total length. A weak-filled matrix D by triangulation is sufficient to perform the 'single linkage cluster analysis' method or to calculate the minimum spanning tree respectively.

Concerning these considerations the described procedure of the 'single linkage cluster analysis' method using the Euclidian distances (1) can be transformed to a method with the distance measure shortest path on a spatial network of roads (2). The filling process of the matrix D with distances (2) comprises the following three steps:

Step 1. Calculation of stand identifiers representing the harvest stands
First, all stand identifiers of the harvest stands are to be calculated as described above.

Step 2. Delaunay triangulation of these stand identifiers
The stand identifiers are meshed by a Delaunay triangulation using Euclidian distances (1) (Delaunay 1934). The neighboured stand identifiers of the triangles are called *natural neighbours* (Gower and Ross 1969).

Step 3. Adding distances to D for adjacent stand identifiers
The natural neighbours to each stand identifier are used only for the selection of presumed natural neighbours on the spatial network and are examined with consideration of the distance function (2). To this, all shortest paths are computed on the spatial

network from a stand identifier to its natural neighbours. Subsequently it is checked, if a shortest path intersects other stand identifiers. If other stand identifiers were intersected, the shortest path $dd_{ij}$ is decomposed into partial paths $dd_{i1}$, $dd_{12}$ ... $dd_{kj}$ that are entered in the matrix D. Otherwise the shortest path $dd_{ij}$ is stored in the matrix D. The partial paths themselves are part of the shortest connections between the stand identifiers.

The advantage of this approach is a remarkable reduction of distance calculations (2). The number of natural neighbours calculated for 5,000 analysed harvest stands in Aracruz varied between 3 and 12. In average only six shortest paths from each stand identifier to other stand identifiers were required. This resulted in an almost linear processing time during calculation. The time required increased roughly by a factor of $O(n)$ instead of $O(n^2)$.

The identification of the *minimum spanning tree* for a given distance matrix D is the result of the 'single linkage cluster analysis' method and was derived quantitatively by the following procedure according to Kruskal (1956). The n stand identifiers result initially in the set of sets $C^0 = \{\{1\}, \{2\}… \{n\}\}$ with n clusters. The elements of the distance matrix D are sorted by increasing distance and saved in a distance list. Furthermore, the indices of the rows and columns are included in the sorted list. In any case, the first element of this list with the stand identifiers $(p, q)$, $(p<q) \wedge p, q \in \{1 … n\}$ is an element of the minimum spanning tree. The union of the two individual sets or clusters $\{p\}$ and $\{q\}$ results in the new cluster $\{p, q\}$. $C^0$ then converts to $C^1 = \{… \{p, q\} …\}$. Three cases can arise for the next list element:

(a) $\{r\}$ and $\{s\}$ remain individual clusters in $C^1$ and are united in the cluster

$$\{r, s\} : C^2 = \{… \{p, q\} …. \{r, s\} …\}.$$

(b) $\{r\}$ or $\{s\}$ already belong to a cluster with more than one element. Should, for instance, s $\in \{p, q\}$, then the union $\{p, q\} \cup \{r\}$ is part of the minimum spanning tree. This then results in the quantity: $C^2 = \{… \{p, q, r\} …\}$.

(c) r and s already belong to a set (cluster).

In the cases (a) and (b), the elements of the sorted distance list are marked and case (c) prevents from creating cycles. The sorted distance list is searched until the set $\{\{1, 2, … n\}\}$ is reached. The marked elements in the distance list form the minimum spanning tree from which the clusters are generated. Deleting all the distances greater than the
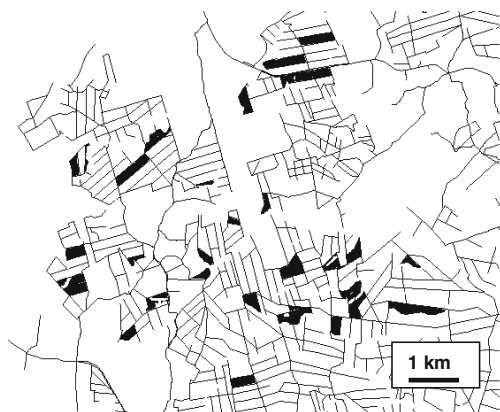
Fig. 2 Spatial network of roads (*grey lines*) and a selected set of 47 harvest stands (*black polygons*)



Fig. 4 The reduced distance matrix D of the distances by Eq. (2) (*black lines*) as a line-layer and the minimum spanning tree (*bold black lines*)
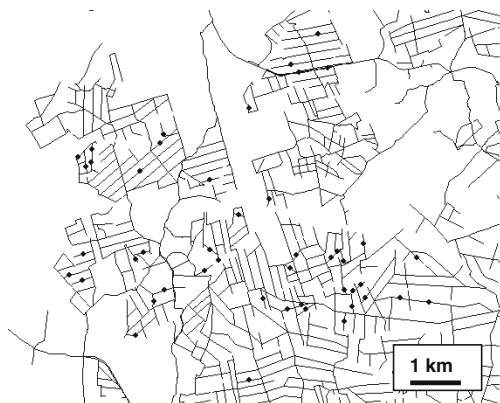
chosen maximum distance from the minimum spanning tree leaves only those groups, in which the distance to each other is smaller than the chosen maximum distance.

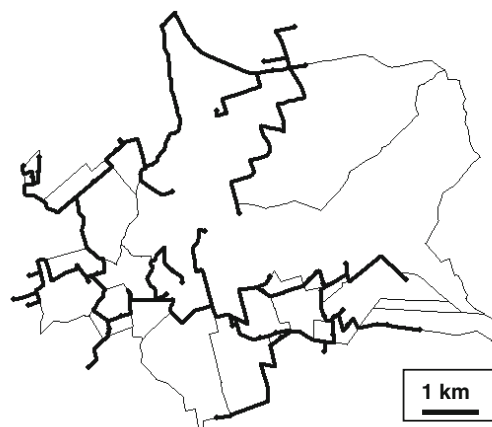In the following, the presented method is illustrated by an example.

## 4 Result of a clustering in practice

The forest area of Aracruz comprises nearly 200,000 ha Eucalyptus plantations with nearly 24,000 stands. Every year, about 3,000 stands are harvested.

From these stands, an example is given here with $n=47$ harvest stands (Fig. 2) to apply the method 'single linkage cluster analysis' using distance Eq. (2). In the Aracruz forests, the quality and state of roads allow a maximum distance of 3,000 m for the movements of the crawler harvester between the stands. Therefore, the quantity of the chosen maximum distance depends on the local conditions and the applied harvest technology. The calculations are based on the locations of harvest stands and on the spatial network of roads.

At first the point geometries of the stand identifiers, representing the harvest stands on the spatial network, are calculated (Fig. 3).

The natural neighbours of stand identifiers can be determined by triangulation to calculate the reduced number of the shortest paths or the shortest partial paths respectively between stand identifiers. The shortest paths are stored in the attribute table of a line-layer instead of a distance matrix D. The attribute table of this layer contains the *dd* values and the identifications of the affected harvest stands. For the 47 harvest stands, only 99 distances, instead of 1,081 (= $n\,(n-1)/2$), needed to be computed to derive the minimum spanning tree. The lines representing the minimum spanning tree are shown in bold (Fig. 4).

Harvest stands can only be part of a cluster if the distances between them are below a defined maximum distance. For this example, a maximum distance of 3,000 m has been chosen. The clusters of harvest stands are generated by removing all the edges greater than 3,000 m of the minimum spanning tree (Fig. 5).



Fig. 3 Spatial network of roads (*grey lines*) and the calculated stand identifiers (*black points*) representing the harvest stands as point-layer



Fig. 5 The deletion of distances greater than 3,000 m from the minimum spanning tree results in clusters (*bold black lines*) of stand identifiers or harvest stands respectively. Three of them are individual clusters (*black points*)

INRA    🕊 Springer

In Fig. 5, the bold lines indicate how five harvest clusters might be connected. Within these harvest clusters the distances between the stand identifiers are less than the chosen 3,000 m using distance equation (2).

In addition to the harvest clusters, single stand identifiers can be seen (Fig. 5). They represent stands that do not fulfil the distance requirements. These stands must be harvested separately, according to the distance requirements.

## 5 Discussion

The advantage of the presented method is the modification of the standard distance (1) in favour of a more realistic measure (2). This is achieved by using the shortest path between two points on the existing network of roads instead of the Euclidian distance. Consequently, the access to an accurately digitised road network as digital dataset is required. The method programmatically developed in the years 2003–2004 has been tested and evaluated at Aracruz (Brazil), where it has been used successfully since then.

The presented method is appropriate for larger enterprises that need to schedule a large number of stands for harvesting in a given planning period and want to concentrate harvesting in specific areas due to the technically complex nature of operations. A plan intended harvest stands can be checked with respect to their spatial distribution and the mutual distances. Individual stands (Fig. 5) can be removed from the plan and replaced by stands with better spatial location. The clustering of harvest stands and the concentration of harvesting operations include significant logistical advantages. Yet smaller enterprises may also benefit from harvesting in groups wherever topographic obstacles prevail in the forest area. The method can be applied to improve multi-objective long-term planning models with spatial objectives (Chen and von Gadow 2002; Öhman and Lämas 2003).

Generating clusters of harvest stands leads to a reduction of timber harvesting costs mainly because the time for site transfers and total driving distances can be reduced. The site transfer times and the related costs vary and depend on the type of transfer: self-driving of the machine or transporting the machine by a flatbed trailer.

In the meantime it has become obvious that the clustering of harvest stands is necessary also in other countries. In Germany for example 2009 the contractors worked in an area with a mean radius of 143 km. The average transport cost by flatbed trailer is 8.50 €/km (Narjes 2009). Therefore, the economic break-even-point for using a flat-bed trailer as opposed to driving the machine itself is 15 km. Thus, a contractor who has an action radius of more than 140 km can use the method for clustering stands when planning his harvest activities and may decide to take 15 km as the maximum distance for clustering the stands.

After the harvest stands are successfully divided in clusters, it is possible to further optimise machinery deployment within a harvest cluster. This corresponds to the classical 'Travelling Salesman Problem' or the 'Chinese Postman Problem'. Heuristic procedures like 'simulated annealing', 'tabu search' or 'genetic algorithms' are available to identify the optimal circuit within and between the clusters (Pirlot 1996; Michalewicz 1999).

Clustering harvest stands is a sub-process of the whole wood supply chain. The optimisation of a sub-process, however, does not necessarily lead to the optimisation of the whole process. As timber harvesting procedures involve felling, forwarding, storing and transporting logs to the mill, these processes need to be coordinated in order to optimise the whole wood supply chain.

## References

Chen BW, von Gadow K (2002) Timber harvest planning with spatial objectives, using the method of simulated annealing. Forstw Cbl 121:25–34

Clark PJ, Evans FC (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453

Dai D, Taquechel E, Steward J, Strasser S (2010) The impact of built environment on pedestrian crashes and the identification of crash clusters on an urban university campus. West J Emerg Med 11:294–301

Delaunay B (1934) Sur la sphere vide. Bulletin of Academy of Sciences of the USSR, 793–800

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271

Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. J Royal Stat Soc Ser C, Appl Stat 18:54–64

Kruskal J (1956) On the shortest spanning subtree and the travelling salesman problem. Proc Am Math Soc 7:48–50

Michalewicz Z (1999) Genetic algorithms+data structures=evolution programs. 3rd rev. and extended ed. Springer, Berlin, p 387

Narjes HJ (2009) Presentation at LIGNA Fair, Hannover, unpublished

Öhman K, Lämas T (2003) Clustering of harvest activities in multi-objective long-term forest planning. For Ecol Manag 176:161–171

Okabe A, Kitamura M (1996) A computational method for market area analysis on a network. Geogr Anal 28:330–349

Okabe A, Yamada I (2001) The K-function method on a network and its computational implementation. Geogr Anal 33:271–290

Okabe A, Yomono H, Kitamura M (1995) Statistical analysis of the distribution of points on a network. Geogr Anal 27:152–175

Okabe A, Okunuki K, Shiode S (2006) SANET: a toolbox for spatial analysis on a network. Geogr Anal 38:57–66

Pirlot M (1996) General local search methods. Eur J Oper Res 92:493–511

Ripley BD (1976) The second-order analysis of stationary point processes. J Appl Probab 13:255–266

Scheurer J, Curtis C (2008) Spatial network analysis of multimodal transport systems: developing a strategic planning tool to assess the congruence of movement and urban structure. Research Monograph, Perth, p 52

Shiode S, Shiode N (2009) Detection of multi-scale clusters in network space. Int J Geogr Inf Sci 23:75–92

Steinhausen D, Langer K (1977) Clusteranalyse: Einführung in Methoden und Verfahren der automatisierten Klassifikation. Walter de Gruyter, Berlin-New York, p 205

Sugihara K, Okabe A, Satoh T (2008) Computational method for the point cluster analysis on networks. Centre for Spatial Information Science, University of Tokyo, Discussion paper no. 90, pp 25

Sugihara K, Okabe A, Satoh T (2011) Computational method for the point cluster analysis on networks. GeoInformatica 15:167–189

Urban D, Keitt T (2001) Landscape connectivity: a graph-theoretic perspective. Ecology 82:1205–1218

Yamada I, Thill J-C (2004) Comparison of planar and network k-functions in traffic accident analysis. J Transp Geogr 12:149–158

Yiu ML, Mamoulis N (2004) Clustering objects on a spatial network. In: Proceedings of the ACM conference on management of data (SIGMOD). Paris, June 2004, 443–454