Community detection in spatial networks: inferring land use from a planar graph of land cover objects

A.J. Comber[1]*, C.F. Brunsdon[2], C.J.Q. Farmer[3]

[1] Department of Geography, University of Leicester, Leicester, LE1 7RH, UK.
ajc36@le.ac.uk
[2]Department of Geography, University of Liverpool L69 3BX
Christopher.Brunsdon@Liverpool.ac.uk
[3] National Centre for Geocomputation, National University of Ireland, Maynooth, Ireland. carson.farmer@nuim.i.e.

* corresponding author

Abstract

This paper applies three algorithms for detecting communities within networks. It applies them to a network of land cover objects, identified in an OBIA, in order to identify areas of homogenous land use. Previous research on land cover to land use transformations has identified the need for rules and knowledge to merge land cover objects. This research shows that Walktrap, Spinglass and Fastgreedy algorithms are able to identify land use communities but with different spatial properties. Community detection algorithms, arising from graph theory and networks science, offer methods for merging sub-objects based on the the properties of the network. The use of an explicitly geographical network also identifies some limitations to network partitioning methods such as Spinglass that introduce a degree of randomness in their search for community structure. The results show such algorithms may not be suitable for analysing geographic networks whose structure reflects topological relationships between objects. The discussion identifies a number of areas for further work, including the evaluation of different null statistical models for determining the modularity of geographic networks. The findings of this research also have implications for the many activities that are considering social networks, which increasingly have a geographical component.

Keywords: OBIA; network; community detection; land cover to land use; modularity

1. Introduction

This paper applies graph partitioning methods to a network of land cover objects in order to identify area of homogenous land use. It introduces 'community detection' methods arsing from network sciences which use the only the internal structure of the graph for partitioning networks into sub-graph regions or 'communities'. Three algorithms employing different statistical operations were applied to a weighted network of land cover objects derived from an object based image analysis (OBIA). The land cover network was defined on land cover object topology (adjacency) and weights were generated from object attribute similarity.

OBIA is now a common approach in remote sensing. It uses object structures to represent areas on the ground that are homogenous to some degree. Objects may be generated through some segmentation process or they may be imported from ancillary data such as land ownership or cadastral boundaries. Typically segmentation parameters are specified heuristically through trial and error, although some automated methods are starting to emerge in the literature (eg Gao et al., 2011). Control over the segments is through adjustment of segment scale and image parameters (Van der Sande et al., 2003; Wang et al., 2010), with the aim replicating the areal units of interest on the ground. Some authors have commented that OBIA is better able to represent 'reality' as perceived by ecologists, field surveyors and air-photo interpreters than pixel-based remote sensing approaches (Lucas et al., 2007). OBIA produces structures with rich spatial and topological characteristics usually using object-level attribution (or metadata) in contrast to pixel based classifications (Blaschke, 2010). OBIA outputs are structures that can readily be recast into networks, specifically planar graphs, using their topological and thematic attributes (Benz et al, 2004; Zhou et al., 2008).

Networks are composed of vertices and edges that represent objects, agents or individuals and the interactions between them. Real networks are highly heterogeneous and have vertices with a wide distribution of degree values, for example. They are not regular lattices and are not random. Rather they are described as "objects where order coexists with disorder" (Fortunato, 2010, p2). Many geographic phenomena can be described using network structures. Individual objects, processes, areas and the relationships between can be represented as either nodes (*vertices*) or arcs (*edges*). Additionally, network edges can be weighted based on the strength of the relationship between objects or processes.

Over the last 10 years researchers from statistical physics and mathematics have developed a range of algorithms for analysing networks or graphs (the terms are used interchangeably here) in order to identify communities. Communities are sub-graph regions that are homogenous in some way. The algorithms use only the information encoded within the network (i.e. without any a priori knowledge of the system under consideration) and have been applied to co-authorship networks, protein-protein interactions, business organisational structures, cell phone networks and social networks. This paper introduces and compares three methods for identifying communities: Walktrap, Spinglass and Fastgreedy algorithms. These use different mathematical and statistical operations to explore graph structure but in each case, the strength of the different graph partitions they suggest are evaluated in terms of the *modularity* of the partition (Newman and Girvan, 2004). Modularity is described in full in Section 3. Essentially, it compares the possible partition of the network with a random version of the original graph with similar structural characteristics but no community structure. However, a number of researchers have expressed concerns about the reliability of the communities that are identified by such methods. For example, Porter et al (2009, p1098) note that "few methods have been developed to use or even validate the communities that we find" and Newman (2008) states that "methods for understanding what the communities mean after you find them are ... still quite primitive" (Newman, 2008, p38).

This research applies different community detection methods to a land cover network in order to select land cover objects to merge into areas of homogenous land use. By analysing an explicitly geographic case study, this research seeks to shed light on concerns over the reliability of the communities that are identified as expressed by Porter et al (2009) and Newman (2008) above. The paper proceeds as follows: Section 2 describes the land cover to land use background and case study. Section 3 introduces networks, the concept of modularity and three methods for identifying community structures. The results of applying the community detection algorithms to the case study are described in Section 4. Section 5 includes a discussion of the results and the methods before some conclusions are drawn in Section 6.

2. Land cover to land use case study

2.1 Land cover to land use

Accurate and reliable land use information is important. For example, recent climate change research has identified changes in land use to be one the major feedbacks into climate cycles and climate change (Strengers et al., 2010). However, the reliable identification of land use from remotely sensed data is a long-standing problems in remote sensing and geoinformatics (Comber, 2008) with a number of characteristics:

First, it is common in many remote sensing surveys for the concepts of 'Land Cover' and 'Land Use' to be confused and treated as if they are the same thing. Brown and Duh (2004) and Fisher et al (2005) document the nature and origins of this confusion.

Second, land cover can generally be distinguished directly from remotely sensed data as it relates to the physical properties of the earth's surface. By contrast, land use classes generally cannot as they describe socio-economic activities which may not be spectrally distinct. This is because any given land use class may be composed of many different land cover types, and any given land cover class may be a component of more than one land use class.

Third, as a consequence land use is commonly *inferred* from land cover data, where the creation of land cover is an intermediate step in land use mapping (Barnsley and Barr, 1996; Zhang and Wang, 2003).

Fourth, transforming land cover to land use requires rules to guide or constrain the transformation. For example, Lackner and Conway (2008) and Chilar and Jansen (2001) generated rules from expert knowledge and relating to the spatial configuration of land cover elements.

Fifth, the process of allocating land use is not always objective. As well as lacking an intrinsic relation to physical matter, membership of one land use class does not preclude membership of another (Bibby and Shepherd, 1999). Land cover may be allocated to specific land use classes for reasons such as institutional objectives, maximising profit or production factors (Monroe and Muller, 2007; Hoeschele, 2000). The way that this inference is conducted may not be transparent as the specific circumstances of any allocation may not be directly measurable (Anselin 2002).

A number of researchers have addressed the land cover to land use problem generating different rules and formalisms to infer land use from land cover. In a series of papers, Mike Barnsley and Stuart Barr explored a number of techniques for inferring land use including from land cover. These include applying a moving kernel to group clusters of pixels into discrete land use categories (Barnsley and Barr, 1996), an extended relational attribute graph model to infer land use from the spatial pattern of land cover objects (Barr and Barnsley 1997; Barnsley and Barr 1997) and analysis of the morphological properties of land cover derived from high-resolution satellite data (Barr and Barnsley 2000). Herold et al (2002) applied landscape metrics to identify urban land use structures. Jansen and Di Gregorio (2003) identified agricultural production systems based on the morphology of field patterns and building structures. Chilar and Jansen (2001) outlined conceptual and methodological issues related to interpreting land use categories based on their relation to mapped land cover categories. Brown and Duh (2004) noted the divergent semantic, geometric and spatial relations between land cover and land use and developed an approach for the semantic translation of land use to land cover using stochastic spatial simulation. Comber et al. (2008) analysed the conceptual overlaps between cover and use semantics associated with forest classifications. Lackner and Conway (2008) developed an object-based analysis of urban land use which re-cast land cover derived from high resolution imagery, using a series of derived layers (roads, *etc.*) and an extensive and iteratively applied rule base. In each of these and other similar analyses, specific rule sets and constraints were developed for each of the case studies.

The work of Barnsley and Barr is especially relevant to this study. First, their work pre-dated two developments in the information sciences: the increased use of object-oriented techniques in remote sensing, and the development of approaches for identifying communities in networks (described in the next section). The outputs of object-oriented remote sensing analyses can readily and intuitively be cast into networks, for instance defined on object topology. Second, despite concluding that land use can be identified through analysis of the spatial disposition of constituent land covers, and suggesting that a quantifiable mapping exists between form (land cover) and function (land use) (Barnsley and Barr, 2004), this work was not extended operationally – many of their analyses used simulated land cover data. Thus generic methods for translating land cover to land use are still lacking. Rather it is a process that requires consideration of:

- the different land covers that are associated with any given land use (thematic);
- the varying spatial characteristics of land use composition, for instance the impact of different ‘kernel’ or ‘window’ sizes on aggregations of land cover and the land uses they infer (spatial, granular); and
- knowledge of the local landscape and anthropogenic processes that result in specific cover / use combinations (temporal, knowledge-based).

2.2 Land cover data and pre-processing

A sample of Infoterra's LandBase© was provided for a study area in Leicester, UK. The dataset was chosen for the case study as it has high spatial resolution and contained neighbourhood and spatial context attributes. The sample has 5873 objects or segments, classified into one of 9 land cover classes (LandBase actually provides 10 classes, but there are no objects classified as 'Ocean' in or near the study area) and a minimum mapping unit of 50m^2. Landbase is constructed from an OBIA of a multiple layered image mosaic of Colour Infra-red Imagery, Natural Colour Imagery, a digital surface model and a digital terrain model. Each object carries extensive contextual attributes describing the proportions of each class in its neighbourhood, derived from the OBIA process. The neighbourhood is defined as a spectrally consistent area encompassing the segment and a 50m Euclidean distance from the extent of the object. A neighbourhood might typically encompass an agricultural field, a small woodland parcel or an urban block. If an object classified as 'Tree' has a *neighbourhood Tree* attribute value of 0.962 then 96% of the area surrounding the object is also trees, indicating perhaps an area of dense woodland. The neighbourhood attributes describe the proportions of the following land cover classes in the neighbourhood of each object: Inland Water, Artificial Surface, Buildings, Bare Ground, Herbaceous Vegetation, Sub Shrubs, Shrubs, Tall Shrubs and Trees. A sample of the case study data is shown in Table 1 (note that the there are zero values for some neighbourhood fields in this study area).

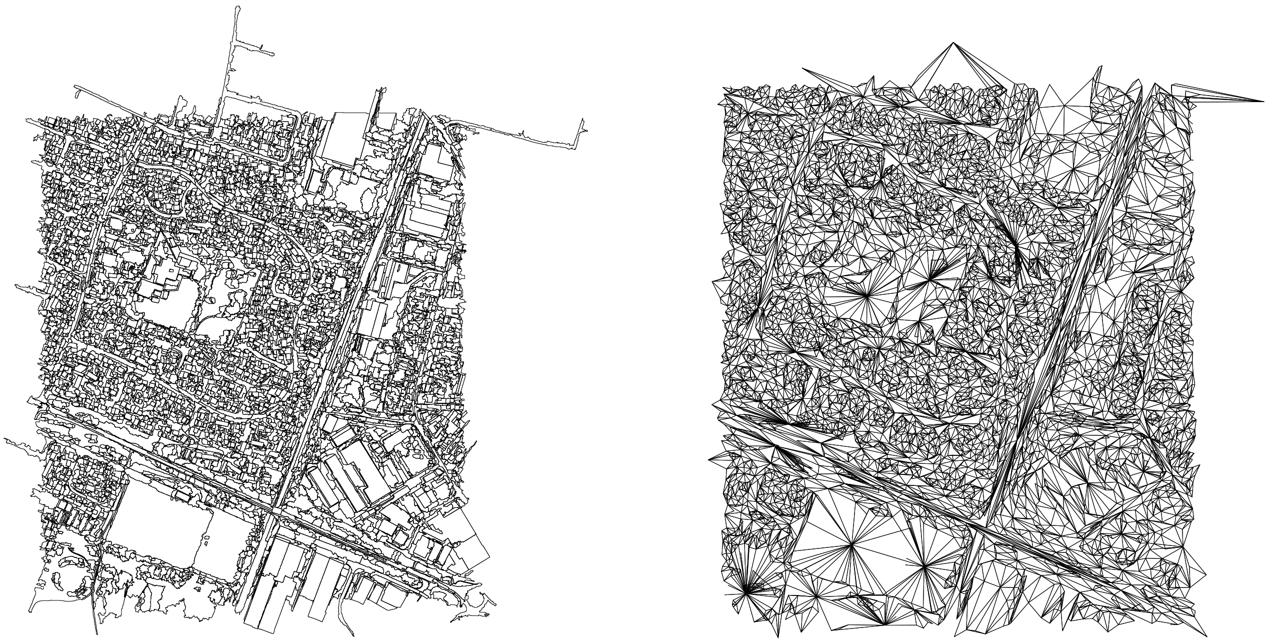
ID	**Water**	**Artificial Surface**	**Buildings**	**Bare Ground**	**Herbaceous**	**Sub-shrub**	**Shrub**	**Tall Shrub**	**Tree**
587	0	0.5227	0.2690	0	0.1933	0	0.0150	0	0

804	0	0.6633	0.0000	0	0.0990	0	0.0268	0	0.2109
847	0	0.5324	0.1657	0	0.2270	0	0.0602	0	0.0147
1615	0	0.5512	0.1436	0	0.2239	0	0.0721	0	0.0092
2619	0	0.4117	0.1465	0	0.2640	0	0.1581	0	0.0197
3531	0	0.6891	0.1373	0	0.1591	0	0.0145	0	0
3820	0	0.8254	0.0000	0	0.1746	0	0	0	0
4091	0	0.8680	0.0787	0	0.0224	0	0.0309	0	0
4567	0	0.2754	0.2881	0	0.3135	0	0.1229	0	0
4892	0	0.5458	0.1905	0	0.2554	0	0.0083	0	0

Table 1. An example of the neighbourhood attribution, describing the proportions of the different land cover classes in the neighbourhood of each segment.

2.3 Weighted land cover network

A number of pre-processing steps were required to convert the data into a network. For each land cover segment the adjacent segments were identified, using a Queen's rule (i.e. where a single shared boundary point indicates contiguity), and the result converted into an undirected graph. The vertices in the graph represented each segment and the edges between them indicated an adjacency relation, as shown in Figure 1. In this case there were 5873 vertices (one for each land cover object) and 17026 edges between them.



a) Land cover segments

b) Adjacency network

Figure 1 a) land cover segments b) centroids of adjacent segments form network vertices joined by lines or graph edges.

The neighbourhood attributes of the land cover objects were used to generate edge weights. The weights for each edge were created from the neighbourhood attributes as follows:

- The Euclidian distance of each segment in attribute space to each other segment was calculated;
- The distances were rescaled to a range [0,1];
- The rescaled distance were subtracted from 1.01 to ensure that the minimum value for any distance was 0.01 (and not to remove any edges spuriously);
- The matrix of the distances was multiplied with the binary matrix indicating presence of edges between vertices (1 for the presence of an edge, 0 for an absence).

The resulting matrix was then converted to an undirected graph, with weights for each of the edges derived from the distance measures.

3. Networks

3.1 Introduction

Network approaches in remote sensing and geo-information have typically been associated with networks representing flows along linear features, such as roads or rivers. Research using network or graph-based approaches for analysing networks describing landscape processes such as land cover or land use is limited. Some work has used them to explore landscape connectivity (Urban and Keitt, 2001) and the impacts of land cover change on species dispersal corridors (Pinto and Keitt, 2008). De Cola (2010) applied graph structures to GIS data to model and visualise the arrangement of land cover patches. Rae et al. (2007) applied a graph-based landscape model to optimize landscape connectivity networks. McRae and Beier (2007) used the concept of ‘resistance distance’ to create an analogy between connected land areas and electrical circuits to predict gene flow in animal and plant populations.

Many geographic phenomena can be described and represented using network structures. Spatial databases of geographic objects can readily be recast into networks based on the object topology. As yet no work has considered the application of recent community detection methods arising from network sciences in a remote sensing / OBIA context, nor applied them to an explicitly geographical network – ie one that describes object topology.

The identification of communities from networks has become a prominent area of research in network science. Three methods for identify communities within networks are presented below but the interested reader is directed to Porter et al. (2009), Newman (2006a) and Leicht and Newman (2008) for overviews of recent research in this area and Fortunato (2010) for a comprehensive review. The different methods analyse graph structure in different ways but in each case, the strength of the different graph partitions they suggest are evaluated by comparing the distribution edges with those expected in random or null model with the same structural characteristics (Modularity). The partitioning algorithms analyse the characteristics of the network and identify possible communities using a number of metrics: degree, connectivity, graph cohesion and adhesion. The ‘degree’ of each vertex is the number of edges connected to it. The connectivity of a graph describes the number of edges or vertices that can be removed to disconnect the remaining nodes from each other. Vertex connectivity describes the number of vertices that need to be removed to remove all paths between any two vertices and defines the *cohesion* of a graph or sub-graph region. Edge connectivity describes the number of edges that need to be removed to remove all paths between any two vertices and defines the *adhesion* of a graph or sub-graph region.

3.2 Modularity

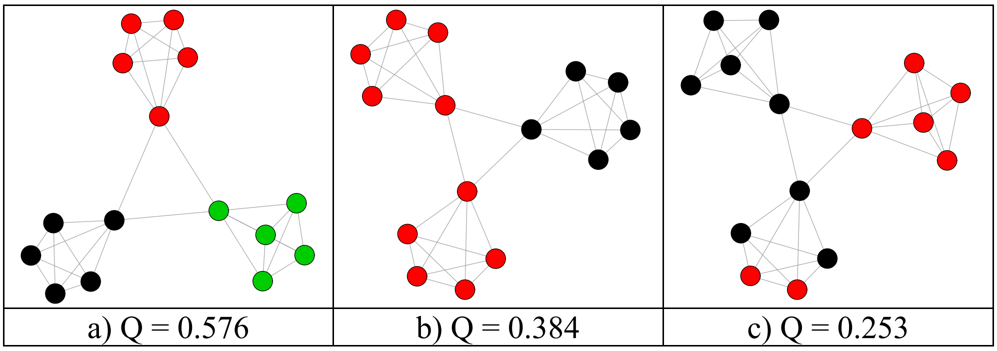
It is possible to partition any given network in a number of different ways. The key issue relates to the quality of any given partition, given the many possible partitions for only a moderately complex network. Newman and Girvan (2004) proposed modularity as a quality measure for a partitioned network. Modularity, *Q*, compares the actual density of edges in a possible partition to the density one would expect given a null model of randomness – a version of the original graph with similar structural characteristics but no community structure – and as is defined as follows:

$$Q = \frac{1}{2m}\sum(A_{ij} - P_{ij})\delta(C_i, C_j) \text{ (eqn 1 from Newman and Girvan, 2004)}$$

where the sum runs over all pairs of vertices, *A* is the adjacency matrix, *m* the total number of edges of the graph, and P_{ij} is the expected number of edges between vertices *i* and *j* in the null model. The *δ*-function returns 1 if vertices *i* and *j* are in the same community (i.e. $C_i = C_j$), and zero otherwise. Note that for weighted graphs, *m* is replaced by $W = ½ \sum_{ij} A_{ij}$, a factor describing the total edge strength in the network.

Modularity defined in this way is based on the notion that a random graph is not expected to have a community structure, so the possible existence of communities is determined by comparing the actual density of edges in a sub-graph and the density one would expect to have in the sub-graph if the vertices of the graph were attached, regardless of community structure. It measures the fraction of the edges in the network that connect vertices of the same type (ie within-community edges) minus the expected value of the same quantity in a network with the same community divisions but with random connections between the vertices.

Modularity provides a precise measure of the total strength of connections within communities versus those between communities. Figure 2 shows the modularity for different communities as indicated by vertex colour. Figure 2a shows three identical communities, each containing a ‘gate-keeping’ vertex that is connected to the other two. Figures 2b and 2c show two communities with different vertex memberships. The modularity scores in Figures 2 a) to c) reflect the degree of unexpectedness associated with graph of the same structure in each case (i.e. the same number of vertices, edges and their distributions). The varying modularity scores reflect the extent to which allocation of individual vertices to different communities reflect that structure.



a) Q = 0.576 | b) Q = 0.384 | c) Q = 0.253

Figure 2. Examples of modularity values for various different network communities, a) with 3 communities, b) and c) with 2 communities.

Three community detection methods are outlined below. They are illustrative rather than exhaustive. Each is available in the igraph package implemented in R and developed by Gábor Csárdi and Tamás Nepusz a description of which is at the R website (http://cran.r-project.org/web/packages/igraph/igraph.pdf). The different methods have different underlying assumptions and approaches for partitioning, edge removal and merging the network into structures. In each cases, the quality of the partitions is evaluated using modularity.

3.3 Greedy approaches ('Fastgreedy')

The Fastgreedy algorithm (Clauset et al., 2004) is an agglomeration algorithm that uses a 'greedy' optimization of modularity. That is, one that makes the locally optimal choice at each stage in the hope of finding the global optimum. The algorithm finds the changes in modularity that would result from the amalgamation of each pair of communities after identifying the largest of them, and performs the corresponding merge. This method takes advantage of the fact the matrices are sparse, resulting in computational efficiency. The algorithm maintains 3 data structures:

- A sparse matrix containing changes in modularity (ΔQ_{ij}) for each pair of communities (i, j) with at least one edge between them;
- An array, H, containing the largest element of each row of the changes in modularity (ΔQ_{ij}) – i.e. a max-heap – along with the identifiers for the corresponding pair of communities (i, j);
- An ordinary vector array with elements of a_i, the fraction of ends of edges that are attached to vertices in each community, i.

The algorithm proceeds as follows:

i) Calculate the initial values of ΔQ_{ij} and a_i and populate the max-heap with the largest element of each row of the matrix ΔQ;

ii) Select the largest ΔQ_{ij} from the largest element in row, join the corresponding communities, update the matrix ΔQ, the heap H and a_i and increment Q by ΔQ_{ij};

iii) Repeat step ii) until only one community remains.

At each iteration of steps i) and ii) resulting in a merge, modularity for the network is calculated.

3.4 Random Walks ('Walktrap')

Pons and Latapy (2005) proposed the Random Walk algorithm. It assumes that if a strong community exists within a network, then a random walker would spend a longer time inside the community due to the density of within-community edges and the high number of possible paths in that community. That is, the random walker gets 'trapped' in densely connected parts of the network corresponding to communities. The algorithm measures the structural similarity between vertices and between communities, defining a distance between them that is calculated from the probabilities that the random walker moves from one vertex to another in a fixed number of steps. The number of steps has to be large enough to allow a significant portion of the network to be explored. The method proceeds as follows. The network is partitioned into communities, each reduced to a single vertex. This partition evolves by repeating the following operations for the (n – 1) steps (where n is the number of vertices):

i) Choose two communities in the partition according to a criterion based on the distance between the communities;

ii) Merge these two communities into a new community and create a new partition;

iii) Update the distances between communities;

iv) After (n − 1) steps, the algorithm finishes and a partition of all vertices is obtained.

Each step defines a partition of the network into communities and each vertex is associated with a particular merging of communities. Pons and Latapy (2005) note that the key characteristic of this algorithm is the way that the communities to merge are chosen and the efficient updating of distances: only adjacent communities (having at least an edge between them) are merged and the two communities that are merged are those that minimize the mean of the squared distance between each vertex and its community.

3.5 Spinglass

Reichardt and Bornholdt (2006) reformulated the problem of community detection in networks as one of finding the ground state of a spinglass model. In physics, particles that possess a magnetic moment are called ‘spins’. They interact with other spins either ferromagnetically (ie ordered because they seek to align) or antiferromagnetically (disordered because they seek to have different orientations). Reichardt and Bornholdt (2006) noted that optimizing modularity in a network is mathematically equivalent to minimizing the energy (known as finding the ground state of the Hamiltonian) of spin system. They combined this with simulated annealing – a probabilistic optimization approach for finding ‘good enough’ solutions to very complex problem – that seeks to improve the current solution with one that is randomly chosen from a sample set of probabilistically similar solutions. The new solution may be accepted depending on how well the new improves on the current one – a probability that depends on a global parameter that is gradually decreased during the process. The Spinglass consists of a simulated annealing algorithm that tries to minimize the following Hamiltonian:

$$H\{\sigma\} = \sum_{i<j} J(A_{ij} - \gamma p_{ij})\delta(\sigma_i, \sigma_j)$$ (eqn 2 from Reichardt and Bornholdt, 2006)

where J is a constant expressing the coupling strength, A_{ij} are vertices in the network, $\gamma > 0$ describes the relative contribution to the energy (or weight) from existing and missing edges, and p_{ij} is the expected number of links connecting i and j for a null model. That is, the Hamiltonian compares the actual distribution of edges in network with the expected distribution given by a particular null model which defines p_{ij}. Under this method the definition of a community is slightly different but with the same effect as the other methods presented here: a community is defined as a group of vertices with the same spin state.

4. Results

The three community detection algorithms were applied to the network described in Section 2 of adjacent land cover objects weighted by their attribute similarity. The land cover objects were allocated to thematically coarse land use classes using simple rules that were applied to the attributes of the graph partition, composed of merged objects (Table 2). The rules do not relate to a specific classification but were established so that land cover to land use translation process could be illustrated. Partition attributes were created from summaries of the attributes of their constituent land cover objects.

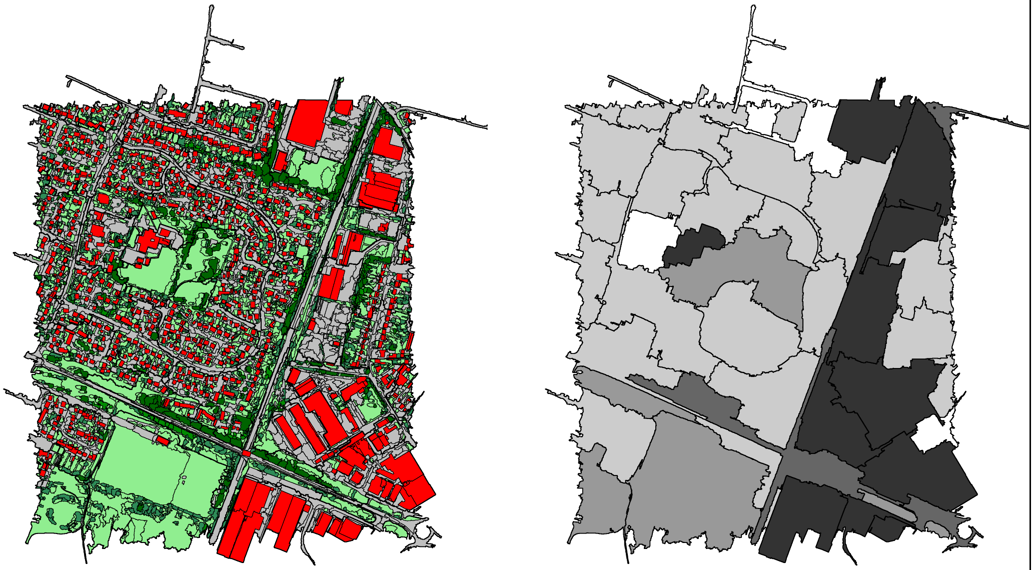
Land Use Class	Land cover	Areal	Operator	Spatial
Infrastructure(Transport)	Artificial Surface	High proportion	AND	High Shape Index
Residential	Buildings	Low mean area	OR	Low Shape index
Industrial	Buildings	High mean area	-	-
Recreation (Leisure)	Herbaceous Vegetation	High proportion	-	-

Table 2. Rules for allocating communities to generic land use classes.

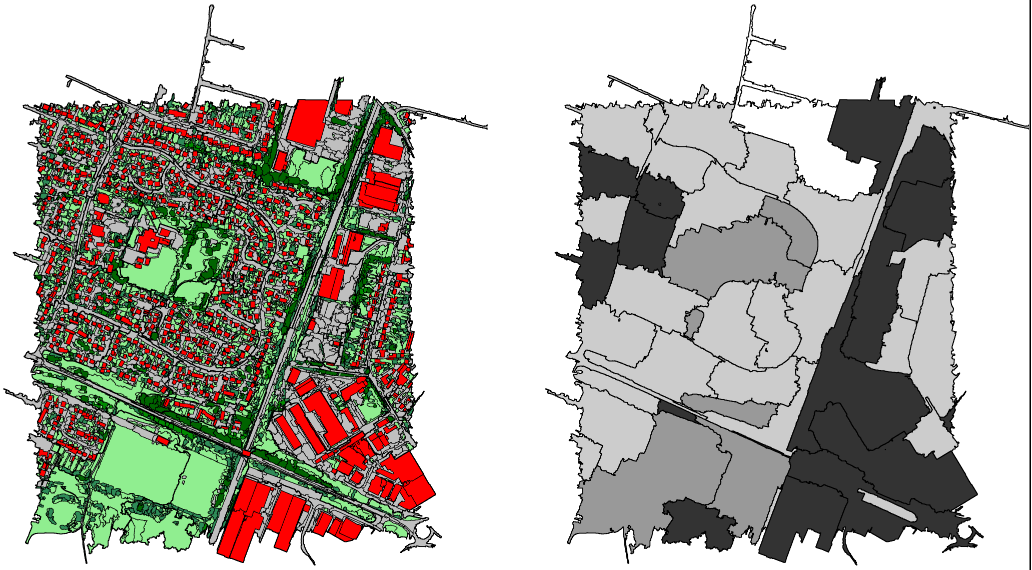
The results of applying the rules to the communities identified by each algorithm are shown in Figure 3. For each of the algorithms some similar patterns are evident. First, there are distinct boundaries between the partitions relating to the road network running from the north east of the study area to the south. The boundaries are those areas of the weighted network where discontinuities between groups of vertices exist. The weighted network was defined on in attribute space (adjacency weighted by similarity in neighbourhood attribute space). Thus, in these areas only weakly weighted edges exist between vertices across such boundaries. Second, similar patterns of land use are evident and through visual inspection the results be seen to reflect the actual land use of the study area:

- Industrial land use to the East and Southeast, with some Infrastructure and some Residential;
- Recreation to the Southwest and centre;
- Infrastructure running Northeast to South and West to Southeast;
- Residential mainly in the centre, North and Northwest but with smaller areas to the East and Southwest.

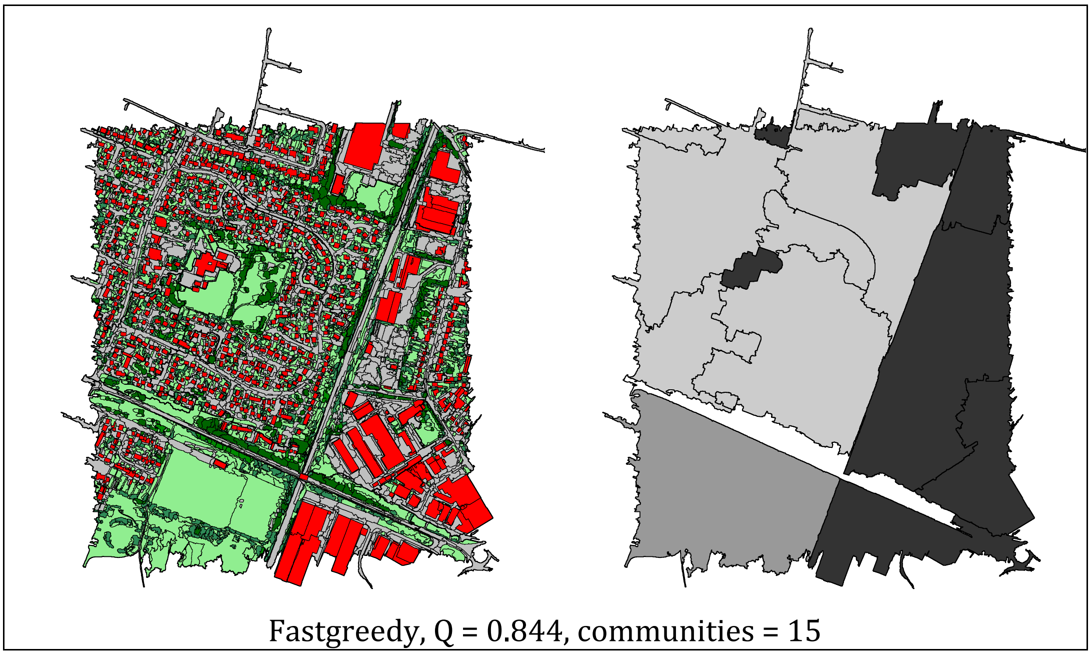
The differences amongst the community detection algorithms are in the number of communities they identify and their associated spatial characteristics. The Walktrap algorithm identified 48 communities, the Spinglass 27 and the Fastgreedy 15 communities. Each algorithm produced markedly different results in terms of the number of land use communities, the optimisation of modularity, the nature of the merges that were performed and the homogeneity of the land use communities that were identified.



Walktrap, Q = 0.904, 48 communities



Spinglass, Q = 0.889, 27 communities



⬜	Artificial Surface
🟥	Buildings
🟩	Herbaceous Vegetation
🟩	Shrubs
🟩	Tall Shrubs
🟩	Trees

⬜	Residential
▨	Recreation
▨	Infrastructure
⬛	Industrial
□	Unclassified

Figure 3. The communities identified by the different algorithms with the underlying land cover structures (left hand side) and the inferred land uses (right hand side).

Walktrap identified 48 communities, separating most of the distinct land use areas. There were some mixed land use communities (in the central area near the greenspace), and through visual inspection only few of the actual Infrastructural land use areas were identified. The 4 Unclassified areas have mixed patterns of land uses. Most of the Residential, Industrial and Recreational land use areas were correctly identified.

Spinglass identified 27 communities and most of the land use regions were correctly identified. However, it is apparent that more areas are delineated (Figure 3) than the stated 27 communities. Inspection of the results revealed that some of the communities were geographically split. Further investigation showed that this was due to the operation of the Spinglass approach. It uses simulated annealing to minimise the Hamiltonian by randomly replacing the current solution with a probabilistically nearby solution, which may not be nearby geographically. Some of this randomness can be controlled but not enough and this point will be returned to in the discussion. The splitting of communities resulted in a number of misclassifications (Residential areas in the Northwest, a smaller area of Recreation misclassified as Industrial, some Residential allocated to Recreation, Infrastructural and Industrial). For example Figure 4 shows two split communities, both with separate and different underlying land uses. However the general pattern of the modelled land uses is correct: with Industrial, Recreational and Residential land uses in the East, South and Northwest respectively.

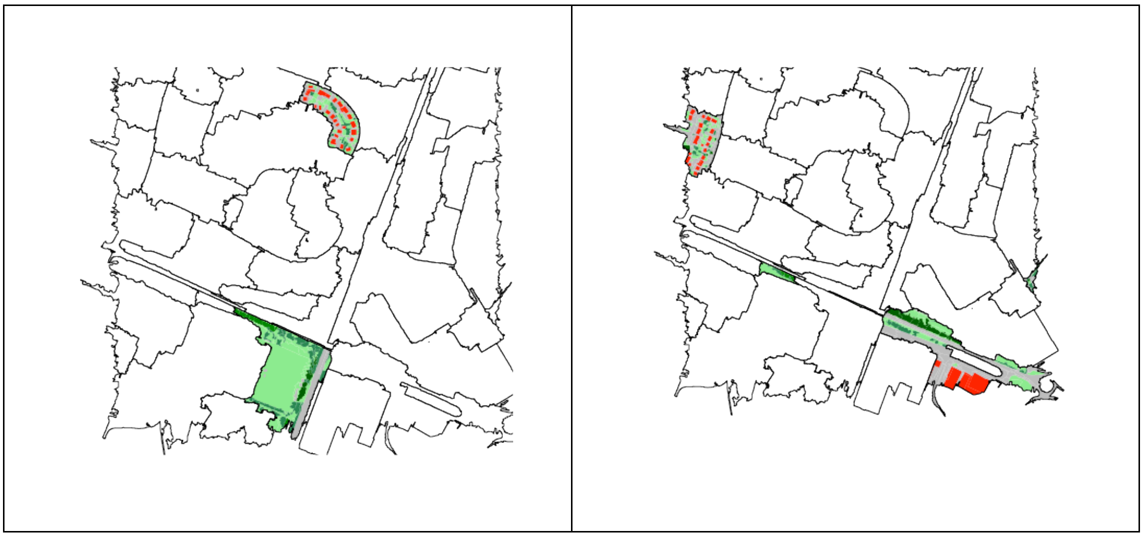


Figure 4. Two examples of single communities identified by the Spinglass method that are split geographically as a result of the randomness introduced by the algorithm.

Fastgreedy identified just 15 communities. The spatial pattern is coarser than the others with large areas of land use identified. However, within this spatial pattern, the algorithm identified homogenous areas of land use. Some of the detail apparent in the other approaches was inevitably lost with fewer communities. For example the areas of Recreational land use in the centre of the study area and the small areas of Residential land use to the East. Additionally because of the coarseness, none of the communities represented long thin areas of Artificial Surface land cover indicating transport infrastructure land uses. The one Unclassified community relates to a road and its grassy verge.

5. Discussion

The application of different community detection algorithms for partitioning graphs to a geographic case study – inferring land use from land cover – results in different merges, given the same input and the same objective function of maximizing modularity. The allocation of those communities, identified from the inherent properties of the network, to a land use class based on the summary statistics of the constituent objects, allows some insight into the operation of the different algorithms, in the context of geographical networks which are in this case planar graphs:

- Fastgreedy. The choice of a locally optimal partition over one that is globally optimal, results in large areal merges of objects, with only relatively large differences in network edge weight providing high differences in modularity.
- Spinglass. The random replacement of the solution with a probabilistically nearby one produces spatially inconstant merges. This 'jumping' to other, less strongly connected portions of the network is problematic when analysing geographic networks.
- Walktrap. The Walktrap algorithm only merges adjacent vertices or communities. Merging choices are made to maximise the movement of the random walker in a fixed number of steps, in this case specified to maximise modularity. The number

of steps determines the number of merged objects that are identified as communities and has an explicitly spatial property: the optimal number of steps (and communities) is derived from an analysis of their topological network weighted by attribute similarity that results in the highest modularity. It relates to the granularity of the objects.

The results also demonstrate varying spatial characteristics: Walktrap identified more detailed communities, Spinglass fewer but potentially for non-spatially contiguous communities, and Fastgreedy identifies fewer and spatially coarser regions. Further, algorithms with heuristic searches, such as Spinglass, introduce some randomness, which need to be constrained over geographic space. Investigations of the algorithm parameters controlling the degree of randomness and the extent to which within group links are rewarded and between group links are penalised, could not eliminate the geographic discontinuities.

The modularity function was used as a stopping criteria for merges in each of the algorithms applied here. It evaluates the quality of the partition by comparing the distribution of the *within* and *between* community connections (edges) against their expected distribution in a random network. Modularity 'embeds in its compact form all essential ingredients and questions, from the definition of community, to the choice of a null model, to the expression of the "strength" of communities and partitions' Fortunato (2010, p100). However, modularity is not without criticism in the literature. Good et al. (2010) showed that maximum modularity increases if the size of the network increases or if the number of good communities increases. Others have similarly argued that high values of modularity may not indicate good partitions as partitions of random graphs can still result in high modularity values (Reichardt and Bornholdt, 2006). There are also questions about whether modularity can detect good partitions on the basis of a single criterion, especially as community structure and size vary so much in the real world. Brandes et al. (2006) note that although several techniques use modularity as a criterion for detecting communities, they do not necessarily provide a globally optimal partition. Other research has found that resolution limits to modularity may exist (Porter et al., 2009; Arenas et al., 2008; Fortunato and Barthélemy, 2007; Ruan and Zhang, 2008). Additionally, other community detecting techniques exist. For example, Edge Betweenness (Newman and Girvan, 2004) uses the number of shortest paths between vertices or communities running through an edge to identify and remove edges. The Leading Eigenvector method developed by Newman (2006b) uses the largest positive eigenvector of the so called 'modularity matrix' to iteratively partition a network into communities.

Future work will i) explore these and other algorithms; ii) compare optimal modularity with optimal partitions of networks of land cover objects defined in other ways; iii) explore the use of *grouping* genetic algorithms (as in Comber et al., 2011) as a method for optimising aggregation into communities; iv) consider alternative 'null' statistical models which may be more appropriate for geographic networks. For example, if weighted-edge based null models are used the segments will always be aggregated on the basis of adjacency. Whereas a more reasonable null model might be a non-random graph with unweighted adjacency-defined edges, such that the 'baseline' for comparison is a set of segments with the same topological structure as input network, but where there is no information distinguishing the characteristics of the segments.

One interesting characteristics of the weighted network approach is that merges are based on the relative difference of the attributes of adjacent segments, rather than on the absolute values of the segments themselves. This is in contrast to traditional cluster analysis, such as *k*-means where membership is generally based on absolute differences compared with the entire dataset, which does account for the spatial structure of the data. This suggests that graph-based divisions may rely more heavily on dissimilarity of one attribute in one region than in another, depending on the local dissimilarities of the other attributes, and that partitioning using graph-based approaches may be more sensitive to local differences.

This work indicates that community detection methods arising from network sciences may offer a set of tools for merging OBIA objects. As the methods use the internal structure of the network to identify communities, the need for a formal rule base is reduced, although the structure and pattern of the merged objects will depend on the nature and granularity of the original objects. The OBIA implications of this work suggests alternative methods for generating a range of merged objects using the properties of the original objects and modularity as an evaluation function, with little need for a rule base. The wider implications of this work indicate the need for careful consideration and analysis of networks with explicit geography and spatial components (for example, much social network data has a location tag). The result of this research shows that geographic space may not be appropriately treated by methods that introduce some randomness or that this needs to be geographically constrained.

6. Conclusions

This research applied a selection of community detection algorithms to and land cover network in order to infer areas of homogenous and contiguous land use, The networks were partitioned into sub-graph regions based on the internal properties of the graph – edge and vertex structure with weights. The results showed that community detection algorithms result in different land cover object aggregations, with variations in granularity of the land use areas. The Fastgreedy algorithm produced the spatially coarsest results and Walktrap the most detailed. The results also showed that community detection / graph partitioning algorithms cannot be universally applied to geographic networks. This is because many geography networks are planar – with an explicit 2 dimensional structure – and algorithms that introduce random replacement of partitions and merges with one that is probabilistically close to the original, such as Spinglass, produce spatially inconstant merges. Such randomness violates the topological properties of the network, where sub-graph partitions have to be geographically contiguous.

Acknowledgements

The authors would to thank Infoterra / Andrew Tewkesbury for access to the LandBase© data and the anonymous reviewers whose comments greatly improved the original manuscript.

References

Anselin, L., (2002). Under the hood - Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3): 247-267.

Arenas, A, A Fernández, and S Gómez. (2008). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics* 10, no. 5 (5): 053039. doi:10.1088/1367-2630.

Barnsley, M.J. and Barr S.L, (1997). Distinguishing Urban Land-Use Categories In Fine Spatial Resolution Land-Cover Data Using A Graph-Based, Structural Pattern Recognition System. *Computers Environment and Urban Systems*, 21(3/4): 209-225.

Barnsley, M.J. and Barr S.L., (1996). Inferring Urban Land Use from Satellite Sensor Images Using Kernel-Based Spatial Reclassification. *Photogrammetric Engineering & Remote Sensing*, 62(8): 949-958.

Barnsley, M.J. and Barr, S.L., (2000). Monitoring Urban Land Use By Earth Observation. *Surveys in Geophysics* 21: 269-289.

Barr, S.L. and Barnsley, M.J., (1997). A region-based, graph-theoretic data model for the inference of second-order thematic information from remotely-sensed images. *International Journal of Geographical Information Science*, 11(6): 555-576.

Barr, S.L. and Barnsley, M.J., (2000). Reducing structural clutter in land cover classifications of high spatial resolution remotely-sensed images for urban land use mapping. *Computers & Geosciences* 26: 433-449.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing* 58 (3–4), 239–258.

Bibby, P. and Shepherd, J., (1999). GIS, land use, and representation. *Environment and Planning B: Planning and Design*, 27: 583-598

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1): 2-16.

Brandes, U., Delling, D., Gaertler, M., Gorke, R. Hoefer, M., Nikoloski, Z. and Wagner, D., (2006).*Maximizing Modularity is hard*. http://arxiv.org/abs/physics/0608255v2.

Brown, D.G. and Duh, J.D. (2004). Spatial simulation for translating from land use to land cover. *International Journal of Geographical Information Science*, 18(1): 35-60.

Chilar, J., and Jansen, L. J. M., (2001). From land cover to land use: a methodology for efficient land use mapping over large areas. *Professional Geographer*, 53(2), 275–289.

Clauset, A., Moore, C., and Newmanm, M.E.J., (2004). Finding community structure in very large networks. http://arxiv.org/abs/cond-mat/0408187v2.

Comber, A., (2008). The separation of land cover from land use with data primitives. *Journal of Land use Science*, 3(4): 215–229.

Comber, A., Medcalf, K., Lucas, R., Bunting, P., Brown, A., Clewley, D., Breyer, J. and Keyworth, S., (2010). Managing uncertainty when aggregating from pixels to objects: context sensitive mapping and possibility theory. *International Journal of Remote Sensing*, 31(4): 1061-1068.

Comber, A.J., Fisher, P.F. and Wadsworth, R.A., (2008). Using semantics to clarify the conceptual confusion between land cover and land use: the example of ‘forest’. *Journal of Land use Science*, 3(2-3): 185-198.

Comber, A.J., Sasaki, S., Suzuki, H. and Brunsdon, C., (2011). A modified grouping genetic algorithm to select ambulance site locations. *International Journal of Geographical Information Science*, 25(5): 807–823.

De Cola, L., (2010). A Network Representation of Raster Land-Cover Patches. *Photogrammetric Engineering And Remote Sensing,* 76(1): 61-72.

Fisher, P.F., Comber, A.J., Wadsworth, R.A., (2005). Land use and Land cover: Contradiction or Complement. Pp. 85-98 in *Re-Presenting GIS*, (eds. Peter Fisher, David Unwin), Wiley, Chichester.

Fortunato, S., (2010). Community detection in graphs. *Physics Reports*, 486(3-5): 75-174.

Fortunato, S., and Barthélemy. M., 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 104(1): 36-41. doi:10.1073/pnas.0605965104.

Gao, Y., Mas, J.F, Kerle, N. and Pacheco, J.A.N., (2011). Optimal region growing segmentation and its effect on classification accuracy, *International Journal of Remote Sensing*, 32(13): 3747-3763.

Girvan, M. and Newman, M.E.J., (2002). Community structure in social and biological networks, *Proceedings of the National Academy of Sciences*, 99: 7821-7826.

Good, B. H., Y. de Montjoye, and Clauset, A., (2010), The performance of modularity maximization in practical contexts. http://arxiv.org/pdf/0910.0165.

Herold, M., Scepan, J. and Clarke, K.C. (2002). The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environment and Planning A*: 34 (8): 1443-1458.

Hoeschele, W., 2000. Geographic Information Engineering and Social Ground Truth in Attappadi, Kerala State, India. *Annals of the Association of American Geographers*, 90(2): 293-321.

Jansen, l.m. and Di Gregorio, A., (2003). Land-use data collection using the ‘land cover classification system’ results from a case study in Kenya. *Land Use Policy*, 20: 131–148.

Lackner, M. and Conway, T.M., (2008). Determining land-use information from land cover through an object-oriented classification of IKONOS imagery. *Canadian Journal of Remote Sensing*, 34(2): 77-92.

Leicht, E.A. and Newman, M.E.J., (2008), Community structure in directed networks, *Physical Review Letters*, 100: 118703.

Lucas, R., Rowlands, A., Brown, A., Keyworth, S. and Bunting, P. (2007). Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping, *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(3): 165-185.

McRae, B.H. and P. Beier (2007). Circuit theory predicts gene flow in plant and animal populations. Proceedings of the National Academy of Sciences of the USA 104,19885-19890.

Monroe, D.K. and Muller D., (2007). Issues in spatially explicit statistical land-use/cover change (LUCC) models: Examples from western Honduras and the Central Highlands of Vietnam. *Land Use Policy*, 24: 521–530.

Newman, M.E.J and Girvan, M., (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69: 026113.

Newman, M.E.J., (2006a). Modularity and community structure in networks, *Proceedings of the National Academy of Sciences*, 103: 8577-8582.

Newman, M.E.J., (2006b). Finding community structure in networks using the eigenvectors of matrices. http://arxiv.org/abs/physics/0605087v3.

Newman, M.E.J., (2008). The physics of networks, *Physics Today,* 61(11): 33-38.

Pinto, N. and Keitt, T.H., (2008). Beyond the least-cost path: evaluating corridor redundancy using a graph-theoretic approach. *Landscape Ecology*, 24(7): 253-266.

Pons, P. and Latapy, M., (2005). Computing communities in large networks using random walks. http://arxiv.org/abs/physics/0512106v1.

Porter, MA, Onnela, J-P and Mucha, PJ, (2009). Communities in Networks. *Notices of the AMS*, 56(9): 1082-1166.

Rae, C., Rothley, K. and Dragicevic, S., (2007). Implications of error and uncertainty for an environmental planning scenario: A sensitivity analysis of GIS-based variables in a reserve design exercise. *Landscape And Urban Planning*, 79(3-4): 210-217.

Reichardt, J. and Bornholdt, S., (2006). Statistical Mechanics of Community Detection. http://arxiv.org/pdf/cond-mat/0603718.

Ruan, J. and Zhang, W., (2008). Identifying network communities with a high resolution. *Physical Review E 77*, 1: 016104. doi:10.1103/PhysRevE.77.016104.

Strengers, B. J., Müller, C., Schaeffer, M., Haarsma, R. J., Severijns, C., Gerten, D., Schaphoff, S., van den Houdt, R. and Oostenrijk, R. (2010). Assessing 20th century climate–vegetation feedbacks of landuse change and natural vegetation dynamics in a fully coupled vegetation–climate model. *International Journal of Climatology*, 30(13): 2055–2065.

Urban, D. and Keitt, T., (2001). Landscape connectivity: a graph-theoretic perspective. *Ecology*, 82(5): 1205-1218.

Van der Sande, C.J., de Jong, S.M. and Roo, A.P.J., (2003). A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment. *International Journal of Applied Earth Observation and Geoinformation*, 4: 217–229.

Wang, J., Chen, Y., He, T., Lv, C. and Liu, A., (2010). Application of geographic image cognition approach in land type classification using Hyperion image: A case study in China. *International Journal of Applied Earth Observation and Geoinformation*, 12S: S212–S222.

Zhang, Q. and Wang, J., (2003). A rule-based urban land use inferring method for fine-resolution multispectral imagery. *Canadian Journal of Remote Sensing*, 29:(1) 1-13.

Zhou, W., Troy, A., Grove, J.M., (2008). Object-based land cover classification and change analysis in the Baltimore metropolitan area using multi-temporal high resolution remote sensing data. *Sensors,* 8: 1613–1636.

List of Tables and Figures