Water Resour Manage (2012) 26:1325–1346
DOI 10.1007/s11269-011-9961-4

Multicriteria Decision Support System for Regionalization of Integrated Water Resources Management

Ana Carolina Coelho · John W. Labadie · Darrell G. Fontane

Received: 20 July 2011 / Accepted: 13 December 2011 /
Published online: 11 January 2012

Abstract Successful implementation of integrated water resources planning and management (IWRM) requires delineation of regions that are relatively homogeneous with respect to multiple criteria, including hydrographic, physical-environmental, socioeconomic, and political-administrative aspects. The water resources planning and management (WARPLAM) DSS is presented as tool for regionalization in support of IWRM through: (1) GIS processing of spatial data related to multiple criteria for defining the homogeneity of clustered base units (e.g., catchments) with respect to multiple criteria; (2) application of fuzzy set theory to development of composite measures of homogeneity over all criteria for alternative clustering of adjacent base units; and (3) development of a modified dynamic programming clustering algorithm that guarantees consistent optimal solutions based on user preferences on the relative importance of the suite of criteria considered for regionalization. The viability of WARPLAM DSS as a tool for regional delineation in support of IWRM is demonstrated through a case study application to the Tocantins-Araguaia River Basin, the second largest in Brazil.

Keywords Decision support systems · Fuzzy sets · Multicriteria methods · Clustering methods · Dynamic programming · Integrated water resources management · Geographic information systems

1 Introduction

Integrated Water Resources Management (IWRM) is a new paradigm for water resources planning and management that has been defined by the Global Water Partnership (2000) as "…a process which promotes the coordinated development and management of water, land,

A. C. Coelho
Agência Nacional de Águas, Brasilia, DF, Brasil

J. W. Labadie (✉) · D. G. Fontane
Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO, USA
e-mail: labadie@engr.colostate.edu

and related resources in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems." Reflected in this statement is the realization that water is a multidimensional resource which cannot be confined to a hydrological context, but requires consideration of socioeconomic, political-administrative, and environmental impacts for sustainable water resources development and management.

With IWRM comes the recognition that the traditional fractured water management approaches must be replaced with holistic, coordinated decision making across multiple sectors and scales that promotes efficiency, equity, and environmental sustainability. IWRM is a multidisciplinary approach to water resources development and management that focuses on participatory methods for gaining a "shared vision" among competing interests that is flexible and adaptable to changing conditions, such as long-term global warming impacts.

A key element in the successful implementation of IWRM is the delineation of regions for cooperative water management that are relatively homogeneous with respect to multiple criteria. This is related to the concept of *problemsheds* proffered by Allan (2005), with boundaries established based on the predominant water management problems confronting the region that may transcend natural watershed boundaries. Ideally, these regional subdivisions should be sufficiently large to promote coordinated interregional water resources decision making, and yet small enough to foster integrated intraregional management among local public, private, and other interests. Unfortunately, attempts at such regionalization for the purposes of achieving IWRM are often carried out without sufficient scientific support or commonly agreed upon principles, and are overly biased by the political and historical context.

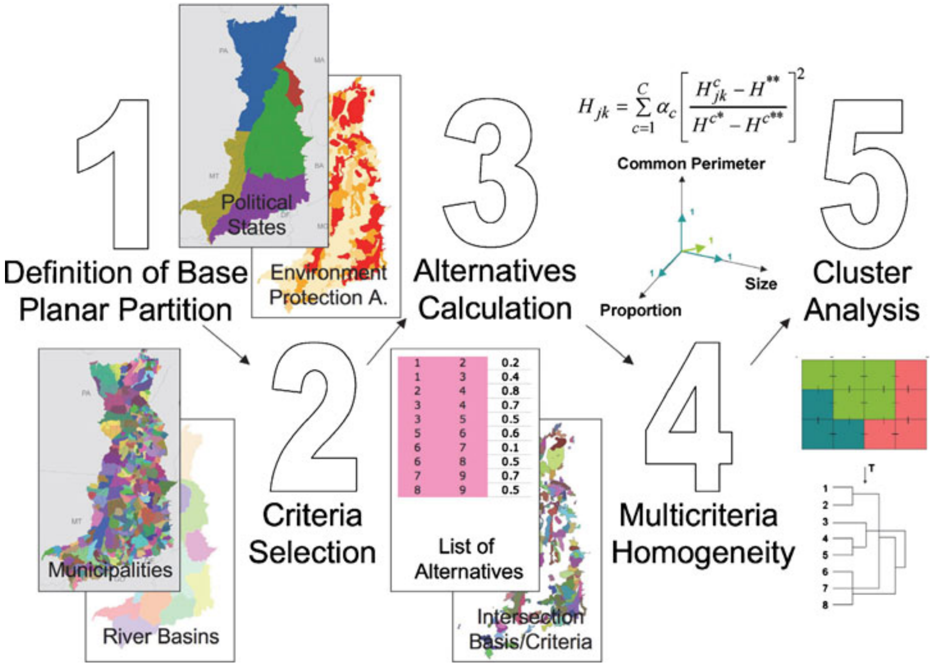
Mostert et al. (2008) affirm the central importance of boundaries in water resources management, recognizing they should be based on multiple criteria and not solely on watershed and river basin limits. Wiering et al. (2010) addressed the incentives and obstacles to regionalization as a means of fostering cross-border collaboration for IWRM, concluding that although regionalization planning is plausible, the actual implementation may be problematic. Allende et al. (2009) integrated a geographic information system (GIS) with cluster analysis for delineating regions in the Cuitzeo Lake Watershed of Central Mexico. Multicriteria decision analysis was applied to ranking the importance of the clustered subwatersheds based on specified ecological and geographical attributes of the clusters for the purpose of designing a hydrometeorological monitoring network. This study focused primarily on regionalization for the purpose of hydrologic analysis rather than integrated water management.

The Water Resources Planning and Management (WARPLAM) decision support system (DSS) is presented herein as a tool for federal and state governments, international commissions, and water councils in defining appropriate territorial limits for water resources planning and management that reflect multiple interests and criteria. Although river basin boundaries are generally considered to be the most suitable to achieve IWRM goals, WARPLAM provides the option for decision makers to include socioeconomic, political, and environmental aspects into the analysis. The result is improved dialog between multiple users and opportunities for integration of the water-related sectors as a means of overcoming the hurdles that interfere with strategic water uses. Application of the proposed DSS also improves understanding of the most critical water-related problems and priorities, as well as identification of the key stakeholders and interests groups in each region. WARPLAM facilitates analysis of the most suitable scale for water resources planning and management to encourage better integration among local, regional and national interests, particularly in federated countries and transboundary river basins. Public participation in the water resources planning and management process is also enhanced by the realization that the regionalization process may better reflect public interests.

The steps in the decision analysis process for IWRM regionalization as incorporated in WARPLAM DSS are presented, followed by a description of the GIS processing required for developing the measures of homogeneity of cluster alternatives. Multicriteria measures of homogeneity of clusters of the planar partitions (e.g., catchments) are derived as membership functions of fuzzy sets. Optimal clustering algorithms are then surveyed, with selection of a modified dynamic programming algorithm as the most suitable for this purpose. The clustering algorithm performs a multiobjective optimization based on subjective preference weights assigned by the decision makers and water planning experts to the various criteria, including physical-environmental, political-administrative, hydrographic, and socioeconomic categories. The viability of WARPLAM DSS for regional delineation in support of IWRM is demonstrated through a case study application to Tocantins-Araguaia River Basin, the second largest in Brazil in terms of drainage area and annual discharge exceeded only by the Amazon.

2 Decision Analysis Process for IWRM Regional Delineation

The decision analysis process for delineation of water resources planning and management regions in support of IWRM is conducted in WARPLAM through five general steps, as illustrated in Fig. 1. The first step is the application of GIS for definition of a base planar partition defined over the entire region designated for implementation of IWRM. For example, partitions may be defined as catchments and watersheds of a specified minimum threshold size represented as polygons in a geospatial database such that the union of the partitions represents the entire region with no gaps or overlaps (Haunert and Wolff 2010).



$$H_{jk} = \sum_{c=1}^{C} \alpha_c \left[\frac{H^c_{jk} - H^{**}}{H^{c*} - H^{c**}} \right]^2$$

1	2	0.2
1	3	0.4
2	4	0.8
3	4	0.7
3	5	0.5
5	6	0.6
6	7	0.1
6	8	0.5
7	9	0.7
8	9	0.5

Fig. 1 Five step decision analysis process for delineation of water management regions

Conversely, for highly developed regions, the adoption of elemental areas based on extents of municipal/industrial zones and political-administrative boundaries may provide the basis for analysis. The base planar partition serves as the starting point for performing regionalization through clustering and data segmentation procedures to delineate more generalized regions amenable to implementation of IWRM.

The second step is selection of a set of criteria reflecting the multiple interests and factors governing successful implementation of IWRM, such as hydrographic, political-administrative, socioeconomic, cultural, and environmental aspects. This is followed by development of a hierarchy of more detailed sub-criteria with associated weights reflecting user preferences. Selection of appropriate criteria and sub-criteria is aided by heuristic IF-THEN rules derived from international surveys and interaction with a broad spectrum of decision makers and experts in water resources planning and management from both the public and private sectors (Coelho 2010). This heuristic knowledge guides the user in selection and prioritization of various criteria, such as suggesting land use as an important socioeconomic factor influencing IWRM within a planar partition comprised of catchments.

As seen in Fig. 2, a land use map is represented as a classified GIS map layer, where the various land use categories are assumed to have the same priority within the general land use designation. For example, if agriculture is considered to be more important than the other land use categories in defining IWRM regions, then agriculture can be removed from the land use map and represented in a separate map layer with higher priority assigned. ArcGIS™ Desktop (ESRI, Inc.) is employed as the GIS platform for this study, with criteria maps maintained in the ESRI geodatabase framework for convenient geospatial data storage and management within in the decision analysis process for multicriteria delineation of IWRM regions.

The third step in the decision analysis process involves GIS operations of intersecting the criteria/sub-criteria map layers with the base planar partition and performing polygon-to-line operations, as illustrated in Fig. 3. These operations are required for defining *measures of closeness* or *homogeneity* with respect to each criterion for each adjacent pair of territorial units contained in the base planar partition. Each of these pairs constitutes a single clustering alternative, where each base unit may be included in more than one clustering alternative. Results of these GIS operations are stored in an ESRI personal geodatabase as a Microsoft Access database with sets of attribute tables designed for holding geodatabase metadata along with the feature geometry. In order to support the creation of a more functional and user-friendly interface, ESRI ModelBuilder is employed as a visual programming tool available in ArcGIS for automating the geospatial intersection and polygon-to-line processes and storing the results in a geodatabase (Fig. 4).

The fourth step in the decision analysis process is defining and applying *measures of homogeneity* for every alternative clustered pair with respect to each criterion based on results of the GIS operations stored in the geodatabase. This is followed by development and application of total weighted measures of homogeneity over all criteria for each clustering alternative using the Euclidean distance norm as scaled by the maximum and minimum measures of homogeneity for each criterion over all clustering alternatives. The fifth and final step is the application of an efficient clustering algorithm based on dynamic programming to define various clustering alternatives representing regionalization alternatives for implementation of IWRM. Included in this step is application of fuzzy logic to quantify the degree of uncertainty associated with the computed regionalization scenarios since the criteria are highly subjective. In addition, the polygon features of certain criteria, such as

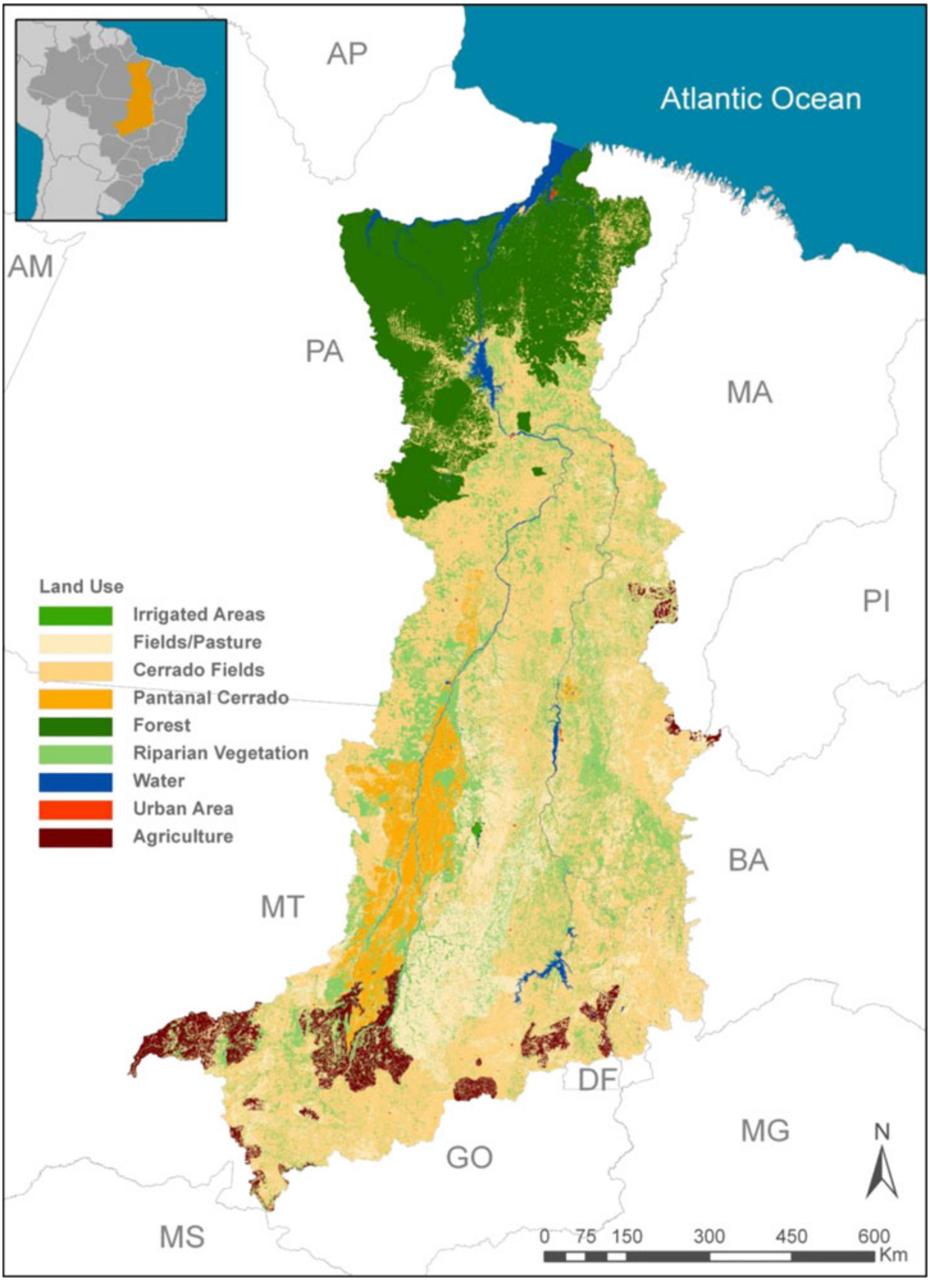


Fig. 2 Classified land use map for the Tocantins-Araguiaia River Basin, Brazil

soil maps, are represented as elements with precise boundaries, but are in fact continuously varying phenomena with indistinct borders.

These steps in the regionalization decision analysis process are embodied in WARPLAM, which is a spreadsheet-based DSS utilizing Excel™ (Microsoft, Inc.)

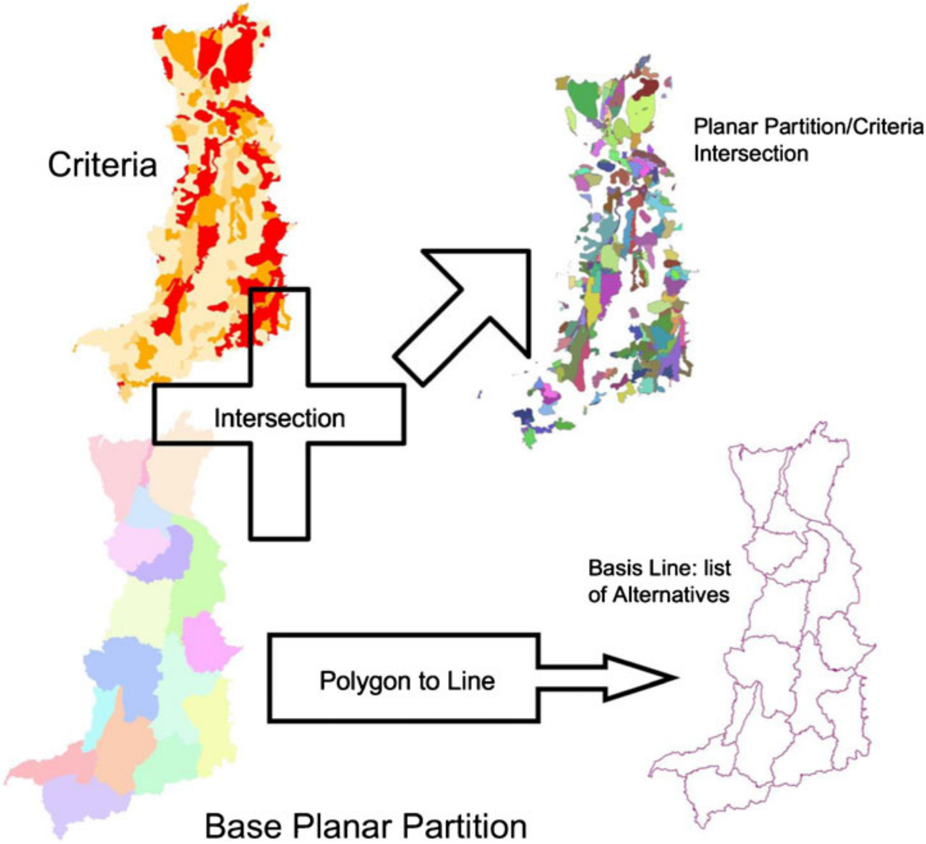
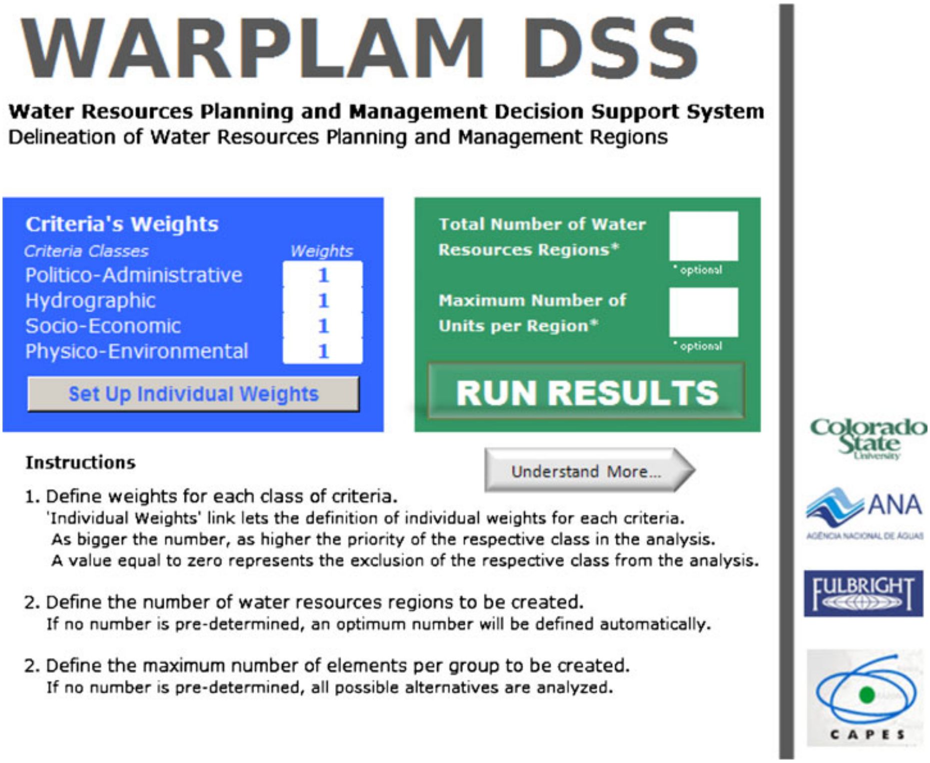

Criteria
Planar Partition/Criteria Intersection
Intersection
Basis Line: list of Alternatives
Polygon to Line
Base Planar Partition

Fig. 3 Illustration of GIS operations in step 3 of the decision analysis process

as the platform selected for integrating the user interface (dialog) subsystem, database management subsystem, and model base management subsystem. Figure 5 displays the user interface for WARPLAM in Excel, where the user sets preference weights for each main or sub-criterion, is alerted for any missing information, and directed to instructions on use of the DSS. The subsystems of WARPLAM DSS are illustrated in Fig. 5, where the database management system integrates geospatial data processed in ArcGIS™ Desktop GIS software (ESRI, Inc.) with nonspatial data maintained in MS Excel spreadsheets. All geospatial information is stored in ESRI geodatabase format and can be accessed and visualized through the ArcMap™ interface to ArcGIS. The ESRI geodatabase format provides an integrated structure for combining input data and output results from GIS spatial analysis and Excel data processing. The model-base subsystem is also implemented in Excel, with VBA scripts developed for data import, calculating measures of homogeneity for each clustered pair and each criterion, and execution of the external dynamic programming optimal clustering module. Results from the decision analysis processes are managed and stored in Excel, as well as exported to ArcMap for visualization and spatial analysis of regionalization scenarios.


WARPLAM DSS
Water Resources Planning and Management Decision Support System
Delineation of Water Resources Planning and Management Regions
Set Up Individual Weights
RUN RESULTS
Understand More...
Colorado State University
ANA
AGÊNCIA NACIONAL DE ÁGUAS
FULBRIGHT
CAPES

Criteria's Weights

Criteria Classes	Weights
Politico-Administrative	1
Hydrographic	1
Socio-Economic	1
Physico-Environmental	1

Total Number of Water Resources Regions* □ *optional

Maximum Number of Units per Region* □ *optional

Instructions

1. Define weights for each class of criteria.
'Individual Weights' link lets the definition of individual weights for each criteria.
As bigger the number, as higher the priority of the respective class in the analysis.
A value equal to zero represents the exclusion of the respective class from the analysis.

2. Define the number of water resources regions to be created.
If no number is pre-determined, an optimum number will be defined automatically.

2. Define the maximum number of elements per group to be created.
If no number is pre-determined, all possible alternatives are analyzed.

Fig. 4 WARPLAM DSS user interface

3 Multicriteria Measures of Homogeneity for Clustering Alternatives

As mentioned previously, each pair of adjacent base units constitutes a single alternative for the cluster analysis. The clustered pairs are evaluated in order to determine a measure of homogeneity with respect to each criterion, which is assumed to be related to the areal extent of each criterion within each of the paired base units. In addition, it is assumed that the measure of homogeneity should increase with the arc length of the shared boundary between the adjacent units, indicating a stronger geographic connection between the adjacent pairs. Guided by these principles, the following equation was proposed by Coelho et al. (2005) as the measure of homogeneity for cluster analysis:

$$H_{jk}^{c} = \frac{2 \cdot L_{jk}}{L_j + L_k} \cdot \frac{A_j^c}{A_j} \cdot \frac{A_k^c}{A_k} \text{for } j,k = 1,\ldots,n; j \neq k; c = 1,\ldots,C \tag{1}$$

where H_{jk}^{c} is the measure of homogeneity for the adjacent pair j, k of base units with respect to criterion c; $A_j^c$is the area of the intersection of criterion c with base unit j; A_j is the area of base unit j; L_{jk} is the arc length of the shared boundary between adjacent base units j and k; L_j is the length of the perimeter of base unit j; n is the total number of units in the planar partition; and C is the total number of criteria/subcritera.

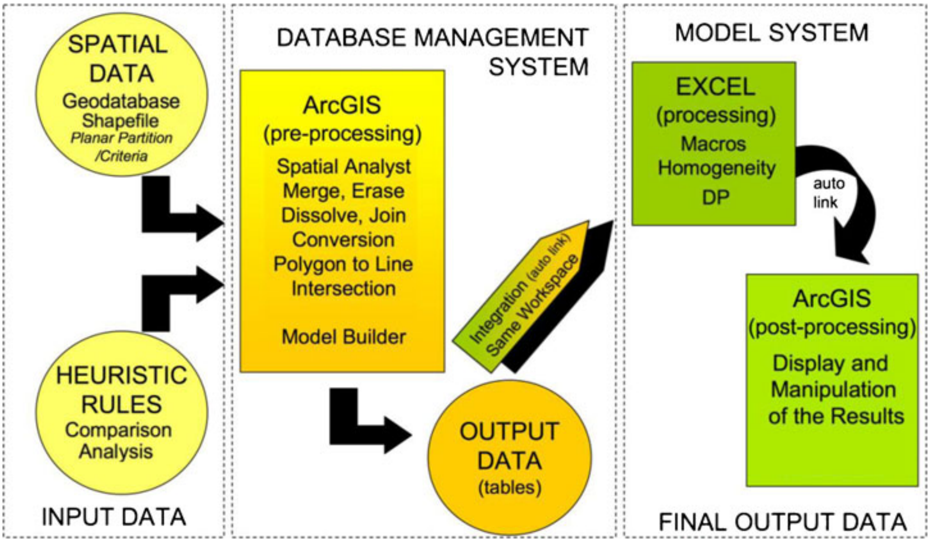


Fig. 5 Database Management and Model Subsystems in WARPLAM DSS

Since the measures of homogeneity H^c_{jk} vary between 0 and 1, they can be considered as membership functions of the fuzzy set of homogeneous adjacent pairs, where a value of 1 represents 100% truth to the assertion that the adjacent units are homogeneous with respect to criterion *c*, and 0 represents no truth to that assertion. The data used for calculating the measures of homogeneity are the result of the GIS intersection and polygon-to-line operations in ArcMap as stored in the ESRI geodatabase.

The next task is to combine the fuzzy membership functions for each adjacent pair (*j*, *k*) in order to create a composite measure of homogeneity that considers all the criteria/subcriteria *c*=1,…,*C*. The weighted Euclidean distance norm is utilized for calculating the composite measure of homogeneity H_{jk} for all order-dependent distinct pairs (*j*, *k*), excluding (*j*=*k*), as normalized by the absolute difference of the maximum and minimum measures of homogeneity over all adjacent pairs (*j*, *k*) for each criterion *c*:

$$H_{jk} = \sum_{c=1}^{C} \alpha_c \left[\frac{H^c_{jk} - H^{**}}{H^{c*} - H^{c**}} \right]^2 \text{for} j, k = 1, ..., n;\ j \neq k \qquad (2)$$

where

$$H^{c*} = \max_{\substack{(j,k) \\ j,k=1,...,n}} H^c_{jk} (c = 1, ..., C)$$

$$H^{c**} = \min_{\substack{(j,k) \\ j,k=1,...,n}} H^c_{jk} (c = 1, ..., C)$$

and weights α_c represent the subjective relative importance of the criteria/subcriteria as specified by the users.

Since the homogeneity calculations are based on adjacent pairs of units in the planar partition, a means of calculating the intra-cluster measure of homogeneity is required based

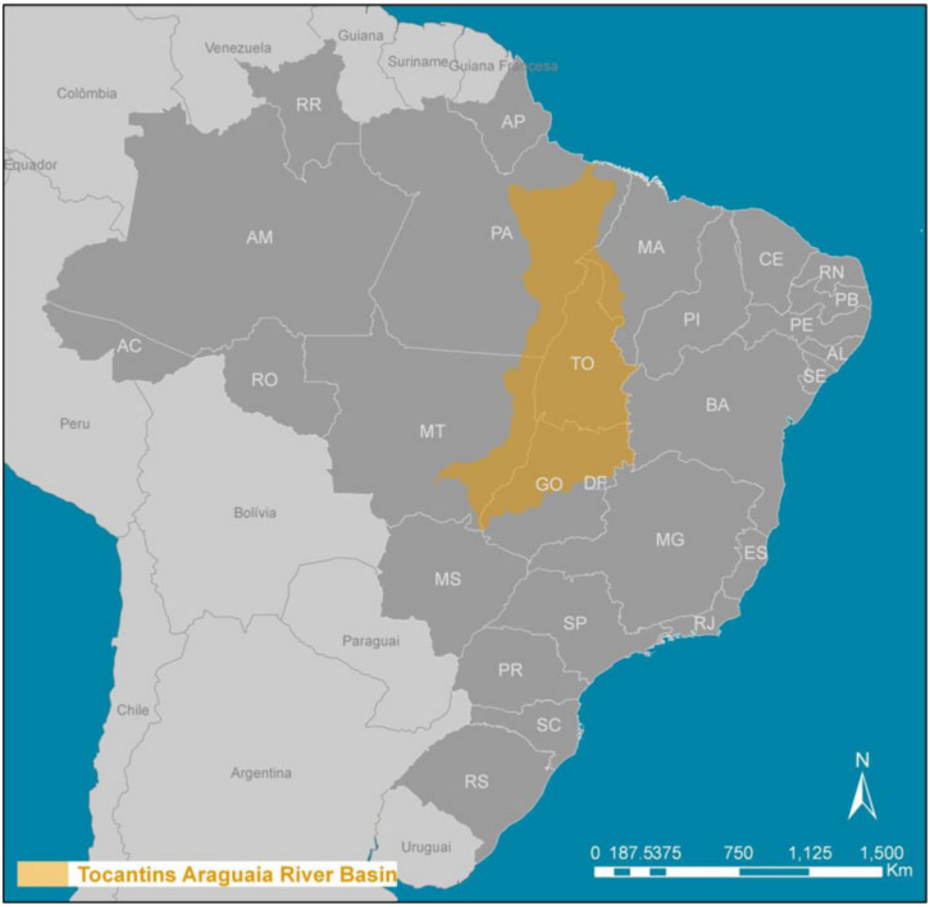


Fig. 6 Tocantins-Araguaia River Basin. Brazil

on the composite homogeneity matrix H_{jk}. Only the upper triangle of the matrix is used for these calculations since pair (j, k) has the same composite measure of homogeneity as pair (k, j), so the focus is on unique combinations of pairs rather than permutations. Let the positions with value 1 in binary string $\mathbf{s}_i$ represent the set of base planar units comprising cluster i. The intra-cluster homogeneity *benefit* B_i is defined as the average of the measures of homogeneity of each order-dependent unique pair of base elements in cluster i represented in the upper triangle of the H_{jk} :

$$B_i = \frac{\sum\limits_{\substack{j \in \mathbf{S}_i; \\ j \neq n}} \sum\limits_{\substack{k \in \mathbf{S}_i; \\ k > j}} H_{jk}}{car(\mathbf{S}_i)} \tag{3}$$

where $\mathbf{S}_i$ is the set of positions in cluster string $\mathbf{s}(i)$ with values of 1, car($\mathbf{S}_i$) is the cardinality of set $\mathbf{S}_i$ (i.e., the number of elements in the set).

If a cluster is comprised of a single element, then no pairs can be defined in that cluster and the intra-cluster measure of homogeneity in that case is set to zero.

Table 1 Example composite homogeneity matrix H_{jk} for an eight element base partition

	Base unit k							
Base unit j	1	2	3	4	5	6	7	8
1		0.2	0.4					
2	0.2			0.8				
3	0.4			0.7	0.5			
4		0.8						
5			0.5					
6					0.6		0.1	0.5
7						0.1		
8						0.5		

Although these singleton cluster alternatives could be removed from the feasible set, they are retained to guarantee the feasibility of clustering alternatives, but with the assurance that there is little chance they would be included in the final clustering solution. The inter-cluster homogeneity for combining the intra-cluster measures of homogeneity for distinct clusters is a simple operation of averaging the (averaged) intra-cluster homogeneity values to provide an overall measure of homogeneity to a clustering solution.

As an example of these calculations, Table 1 gives the composite homogeneity matrix H_{jk} for an eight element base partition. Notice that the matrix is symmetrical and values along the diagonal relate to non-applicable (j, j) pairs. For this example, the intra-cluster homogeneity of each of three clusters is evaluated according to

Table 2 Calculation of intra-cluster measure of homogeneity B_i

(j, k) pairs	H_{jk}	Cluster 1 {1, 2, 3, 4} string $\mathbf{s}_i$							
		1	1	1	1	0	0	0	0
(1,2)	0.2								
(1,3)	0.4								
(2,4)	0.8								
(3,4)	0.7								
sum	2.1								
B_1 (ave.)	**0.525**								
(j, k) pairs	H_{jk}	cluster 2 {5, 6, 7} string $\mathbf{s}_i$							
		0	0	0	0	1	1	1	0
(5,6)	0.6								
(6,7)	0.1								
sum	0.7								
B_2 (ave.)	**0.35**								
(j, k) pairs	H_{jk}	cluster 3 {8} string $\mathbf{s}_i$							
		0	0	0	0	0	0	0	1
(8,8)	0								
B_3	**0**								
$\sum B_i$	0.875								
ave.	**0.292**	Inter-cluster measure of homogeneity							

Table 2, based on the composite homogeneity matrix H_{jk} in Table 1. The final inter-cluster homogeneity is calculated as the average of the B_i values, but with the intra-cluster homogeneity for single element cluster 3 assigned a value of 0.

4 Optimal Clustering Algorithm

4.1 Alternative Clustering Methods

The composite measures of homogeneity over all criteria for each adjacent pair provide the basis for grouping the clustering alternatives into regional delineations that enhance the viability of IWRM. The hierarchical agglomerative approach, often referred to as a *bottom up* clustering method, begins with the pairing of similar elements as initial clusters, which are then successively merged or agglomerated with similar clusters until a single cluster of all elements is produced (Everitt et al. 2001). A disadvantage of this approach is the requirement of specifying a threshold parameter T representing the maximum distance or degree of dissimilarity between elements that should be clustered, where varying the T threshold values can produce nonunique clustering strategies. In addition, use of differing metrics for measuring distances between clusters can also generate different results. Hierarchical dissociative or divisive clustering is a *top–down* approach to clustering which, according to Xu and Wunsch (2009), is less popular than the hierarchical agglomerative approach due to the increased computational burden.

K-means clustering is a popular partitioning method with the advantage of *a priori* selection of the number of clusters K, rather than the less intuitive specification of a threshold parameter T as with the agglomerative hierarchical approach (Kaufman and Rousseeuw 1990). An additional advantage is the computational efficiency of the method, particularly for the analysis of large datasets. However, although *a priori* specification of K is an advantage of the method, it can also be considered a disadvantage since it may be difficult to predict what the correct value of K should be. In addition, different dataset partitions can be generated with each run of the algorithm due the dependence on an initial random specification of cluster means, with no assurance that globally optimum clustering has been achieved.

Optimization-based clustering methods include applications of evolutionary methods such as genetic algorithms (Maulik and Bandyopadhyay 1999) and dynamic programming (Esogbue 1986). Since evolutionary algorithms rely on heuristic operators and random processes, it is not possible to guarantee attainment of global optima or even convergence to consistent solutions. In addition, the number of clusters K must generally be specified *a priori*. On the other hand, dynamic programming (DP) solves the clustering problem as a sequential decision process which, if properly formulated, guarantees convergence to the global optimum while automatically determining the optimal number of clusters, as well families of optimal solutions for a wide range of K. For these reasons, the DP method is applied in this study using the generalized dynamic programming software package CSUDP developed by Labadie (2003).

An attempt was made to solve the optimal clustering problem using the dynamic programming (DP) formulation suggested by Bellman (1973) and Esogbue (1986). In the forward-looking DP recursion over stages $i=1,\ldots,N$ (assuming $K<N$), stage i represents the number of clusters generated at that stage of the sequential decision process, state variable x_{i+1} is the total number of elements clustered through stage i, and decision variable u_i is the number of elements in the i-th cluster. The optimal number of clusters K is determined as that stage where the

maximum *average* of the intra-cluster measures of homogeneity over all K clusters occurs with all elements included in a cluster. This implies that forcing more than K clusters actually reduces the clustering benefit.

As described in Coelho (2010), uniqueness problems emerge with this formulation whereby solutions in intermediate stages of the DP algorithm result in ties that required an arbitrary tie-breaking procedure. Unfortunately, all possible nonunique solutions cannot be carried forward in the sequential decision process due to the explosive increase in combinations of nonunique solutions over several stages. Although the optimal clustering decision at an intermediate stage is among the set of nonunique solutions, the likelihood of its selection in the arbitrary tie-breaking procedure is low, ultimately resulting in suboptimal solutions to the clustering problem A modified DP algorithm is proposed that guarantees unique solutions at each stage, and therefore assures attainment of the global optimal solution.

4.2 Modified Dynamic Programming Clustering Algorithm

The modified DP formulation is initiated by creating a table of all possible clustering alternatives at any stage i as a list of binary strings $\mathbf{s}_j$ (j=1,…,N) of length n, where n is the number of elements to be clustered and N is the total number of unique binary strings with a maximum of m ($< n$) elements with bit values=1, representing the elements included in the cluster. The total number of combinations of binary strings with a maximum of m nonzero elements is:

$$M = \sum_{k=1}^{m} \begin{bmatrix} k \\ n \end{bmatrix} \tag{4}$$

The pre-calculated intra-cluster measures of homogeneity B_j associated with string $\mathbf{s}_j$. is included as a column in the table. The rows are ordered in relation to strings with a single nonzero bit value to strings with m nonzero bits, where the ordering within each cluster of strings with the same number of nonzero bits is arbitrary. All infeasible strings with B_j=0 (i.e., nonadjacent elements included) are then removed from the table, except for those strings with a single element in the cluster. The latter are allowed primarily as means of insuring feasible solutions during initial stages of the DP algorithm. Further processing of the table involves removal of any strings $\mathbf{s}_j$ such that the number of elements in the cluster exceeds a user specified limit on the number of elements allowed in any cluster, followed by sequential renumbering of the integer codes for the remaining strings. Table 3 is a sample of feasible clustering alternatives represented as nine element bit strings $\mathbf{s}(u)$ with associated unique integer code u and pre-calculated intra-cluster measures homogeneity $B(u)$, assuming a maximum of four nonzero elements in each cluster.

The objective function for the optimal clustering problem is:

$$\max_{\substack{K, u_i, \\ i=1,\ldots,K}} \frac{1}{K} \sum_{i=1}^{K} B_i(u_i) \tag{5}$$

subject to:

$$\sum_{i=1}^{K} s_\ell(u_i) = 1; \text{for } \ \ell = 1, \ldots, n \tag{6}$$

Table 3 Sample table of feasible cluster alternatives as binary strings assigned a unique integer code

Integer code u	Intra-cluster homogeneity $B(u)$	No. of Elements	Binary string $\mathbf{s}(u)$
1	0.000	1	0 1 0 0 0 0 0 0 0
2	0.000	1	0 0 1 0 0 0 0 0 0
3	0.000	1	0 0 0 1 0 0 0 0 0
4	0.000	1	0 0 0 0 1 0 0 0 0
5	0.000	1	0 0 0 0 0 1 0 0 0
6	0.000	1	0 0 0 0 0 0 1 0 0
7	0.000	1	0 0 0 0 0 0 0 1 0
8	0.000	1	0 0 0 0 0 0 0 0 1
9	0.800	1	1 1 0 0 0 0 0 0 0
10	0.700	2	0 1 0 1 0 0 0 0 0
11	0.700	2	0 0 1 1 0 0 0 0 0
12	0.600	2	0 0 0 0 0 0 1 0 1
13	0.500	2	0 0 0 0 1 1 0 0 0
14	0.500	2	0 0 1 0 1 0 0 0 0
.			
.			
.			
36	0.533	4	1 0 1 1 1 0 0 0 0
37	0.533	4	0 0 1 0 1 1 0 1 0
38	0.533	4	0 0 0 0 1 1 1 0 1
39	0.500	4	1 0 1 0 1 1 0 0 0
40	0.467	4	0 0 1 0 1 1 1 0 0
41	0.467	4	0 0 0 0 1 1 0 1 1
42	0.467	4	0 0 0 0 1 1 1 1 0
43	0.450	4	0 0 0 0 0 1 1 1 1

$$\sum_{i=1}^{K} u_i \leq x_{\max} \tag{7}$$

where the integer code u_i selected in stage i (i=1,…,K) is associated with bit string $\mathbf{s}(u_i) = (s_1(u_{i1}), ..., \mathrm{s}_n(u_{in}))$ representing a unique cluster alternative; $B_i(u_i)$ is the intra-cluster measure of homogeneity for the bit string represented by integer code u_i ; n is the total number of elements to be clustered; K is the total number of clusters; and $x_{\max}$ is an arbitrary upper bound on the total accumulated integer codes for binary strings selected over each stage. Solution of Eq. 5 is subject to the constraint that each of the n elements is a member of exactly one cluster. This formulation is complicated by K being considered as a decision variable, as well as the objective function defined as the *average* of the intra-cluster measures of homogeneity over all clusters. A reasonable initial estimate is $x_{\max}$ is $\frac{N}{2} \cdot K$, but this can be increased if results indicate that the selected value for $x_{\max}$ is over-constraining the solution. Equation 6 specifies that all elements must be included in exactly one cluster and Eq. 7 simply requires that the accumulation of the integer codes selected for all stages cannot exceed $x_{\max}$.

Defining state variable x_i as the accumulation of integer codes of binary strings selected in stages $1,\ldots,i-1$, Equation 7 is equivalently represented as a state equation for solution by dynamic programming:

$$x_{i+1} = x_i + u_i \text{ for } i = 1, \ldots, N \tag{8}$$

where and $x_1=0$, $x_{K+1} \leq x_{\max}$, and the total number of stages N for the dynamic programming recursive calculations is assumed to be $>K$. A reasonable estimate is to set $N=\text{INT}(n/2)$, where N is an upper bound on the number of clusters such that each unique pair of base units defines a minimal cluster. If n is an odd number, then the lone unclustered base element is not included in the N count.

The dynamic programming solution to this problem uses a forward recursion for calculation of the dynamic programming optimal return function $F_i(x_{i+1})$ for stages $i=1,\ldots, N$ and for all integer $x_{i+1} \leq x_{\max}$ using the *inverted form* of the state equation:

$$F_i(x_{i+1}) = \max_{0 \leq x_i \leq x_{\max}} [B_i(u_i) + F_{i-1}(x_i)] \tag{9}$$

subject to:

$$u_i = x_{i+1} - x_i \geq 0 \tag{10}$$

$$\left.\begin{array}{l} s_\ell(u_i) \neq s_\ell\left(u^*_{k-1}(x_k)\right) \ \forall \ell \ s_\ell \ (u_i) = 1 \\ x_{k-1} = x_k - u^*_{k-1}(x_k) \end{array}\right\} \text{for } k = i, \ldots, 1 \tag{11}$$

$$\textbf{store} \text{ optimal } u^*_i(x_{i+1}) \text{ from stage } i$$

The recursive calculations begin with the assumed boundary conditions $F_0(x_1)=0$. Since the optimal clustering policies $u^*_{k-1}(x_k)$are stored for the previous stages $k=i-1,\ldots,1$, Equation 11 represents a traceback calculation process over the stored optimal clustering policies from the stages previous to stage i so as to insure that elements selected for clustering in stage i are currently unclustered.

The dynamic programming optimal value function defined in the recursion relation of Eq. 9 is inconsistent with the optimal clustering problem objective function of Eq. 5 since Eq. 9 maximizes the total accumulated intra-cluster homogeneity, rather than maximizing the average of the homogeneity measures over all clusters. Lee and Labadie (2007) prove that modification of the recursion relation as given in Eq. 12 succeeds in maximizing the average intra-cluster homogeneity over all clusters:

$$F_i(x_{i+1}) = \max_{0 \leq x_i \leq x_{\max}} \left\{ \left(\frac{1}{i}\right) B(u_i) + \left(\frac{i-1}{i}\right) F_{i-1}(x_i) \right\} \tag{12}$$

In these forward computations through stages $i=1,\ldots,N$, termination may occur prior to reaching the final stage N if feasible solutions cannot be found at that stage. Infeasible solutions encountered at stage i can occur if forcing a solution comprised of exactly i clusters is unattainable since a user-defined maximum number of elements m is allowed in any cluster.

The optimal number of clusters K is found from the stored optimal return function values $F_i(x_{i+1})$ after solution over N stages, or up to the stage where termination occurs due to infeasibility:

$$F_K(x^*_{K+1}) = \max_{i,x_{i+1}} F_i(x_{i+1}) \tag{13}$$

subject to:

$$\sum_{i=1}^{K} s_\ell(u^*_i) = 1;\ \text{for } \ell = 1,\ldots,n \tag{14}$$

where Eq. 14 requires that each element must be included in exactly one cluster. Traceback solutions through the optimal stored integer codes gives the optimal integer codes u^*_i and associated cluster strings $\mathbf{s}(u^*_i)$ by sequentially retrieving the stored optimal clustering policies for each stage i , as illustrated in the following pseudo code:

```
For $i = K,...,1$
    retrieve $u_i^* \ (x_{i+1}^*)$
    table lookup $\mathbf{s} \ (u_i^*)$
    IF $i > 1$ , calculate $x_i^* = x_{i\ +1}^* - u_i^* \, (x_{i+\,1}^*)$
Loop end
```

4.3 Implementation of DP Clustering Algorithm in WARPLAM DSS

The optimal clustering problem is automatically setup and executed in the MS Excel-based WARPLAM DSS, with VBA commands transferring the necessary data files to the CSUDP program and executing it as an external program. The CSUDP results are then automatically imported into WARPLAM for further analysis and display. This procedure allows users to access the powerful capabilities of CSUDP without requiring any knowledge or understanding of the dynamic programming algorithm. In addition, although CSUDP produces the optimal number of clusters K, users are also provided the optimal clustering structure for any user-specified preference for the desired number of clusters, as well as the maximum number of elements in any cluster. This is consistent with the decision support focus of WARPLAM in providing suggested solutions, but also allowing users to experiment with many alternative preference weights, configurations, and parameters.

To test the efficiency of the modified DP optimal clustering algorithm, Coelho (2010) compared it to a genetic algorithm for solving the optimal clustering problem. Several test cases revealed that computer run times were comparable between the two methods, and both produced the global optimal solution for each case. Global optimality of the solutions was confirmed by performing time-consuming exhaustive enumeration procedures over all possible clustering alternatives for cases with a limited number of elements. However, application of the genetic algorithm, as well as other evolutionary-type algorithms, is considered less advantageous since the DP algorithm produces optimal clustering for any number of clusters without any additional expenditure of computer time since this is a routine result of the stage-wise recursive solution structure. In contrast, application of the genetic algorithm requires *a priori* specification of the desired number of clusters.

After completion of the optimal cluster analysis, WARPLAM DSS provides reports on important details about the results, such as the number of clusters created, the elements contained in each cluster, the area of each cluster, and the most significant aspects considered in defining each cluster. These results are also stored for providing future comparisons if further simulations are performed. In addition, fuzzy membership values of each element to the assigned cluster, as well as other adjacent clusters, are provided as a means of measuring the uncertainty associated with the clustering process. Considering the use of subjective criteria and preference weights, the elements have a continuous grade of membership to more than one cluster, indicating that cluster boundaries cannot be considered as precisely defined.

Providing an indication of the fuzzy logic-based uncertainty associated with defining an element as a member of a specific cluster is offered to the decision makers as a means of interpreting the results. The fuzzy membership values are calculated by evaluating the decrease in the objective function resulting from assignment of the respective element to a different cluster. The percentage decrease is used as a reduction factor for the *measure of homogeneity* associated with the respective element and each adjacent cluster. The measures of homogeneity of the respective elements are then balanced to represent the fuzzy membership values. For example, suppose element *a* is assigned to cluster *Y* and is adjacent to clusters *X* and *Z*. If element *a* is assigned to cluster *Z*, instead of cluster *Y*, then the maximum inter-cluster measure of homogeneity is reduced by *m*%. The amount *m* is used as the reduction factor for the *measure of homogeneity* between element *a* and the adjacent element in cluster *Z*. The reduced measure of homogeneity is then compared to the original distribution of measures of homogeneity in order to calculate a new distribution of importance, represented again in percentage. This resultant percentage is assigned as the membership function value of element *a* to cluster *Z*.

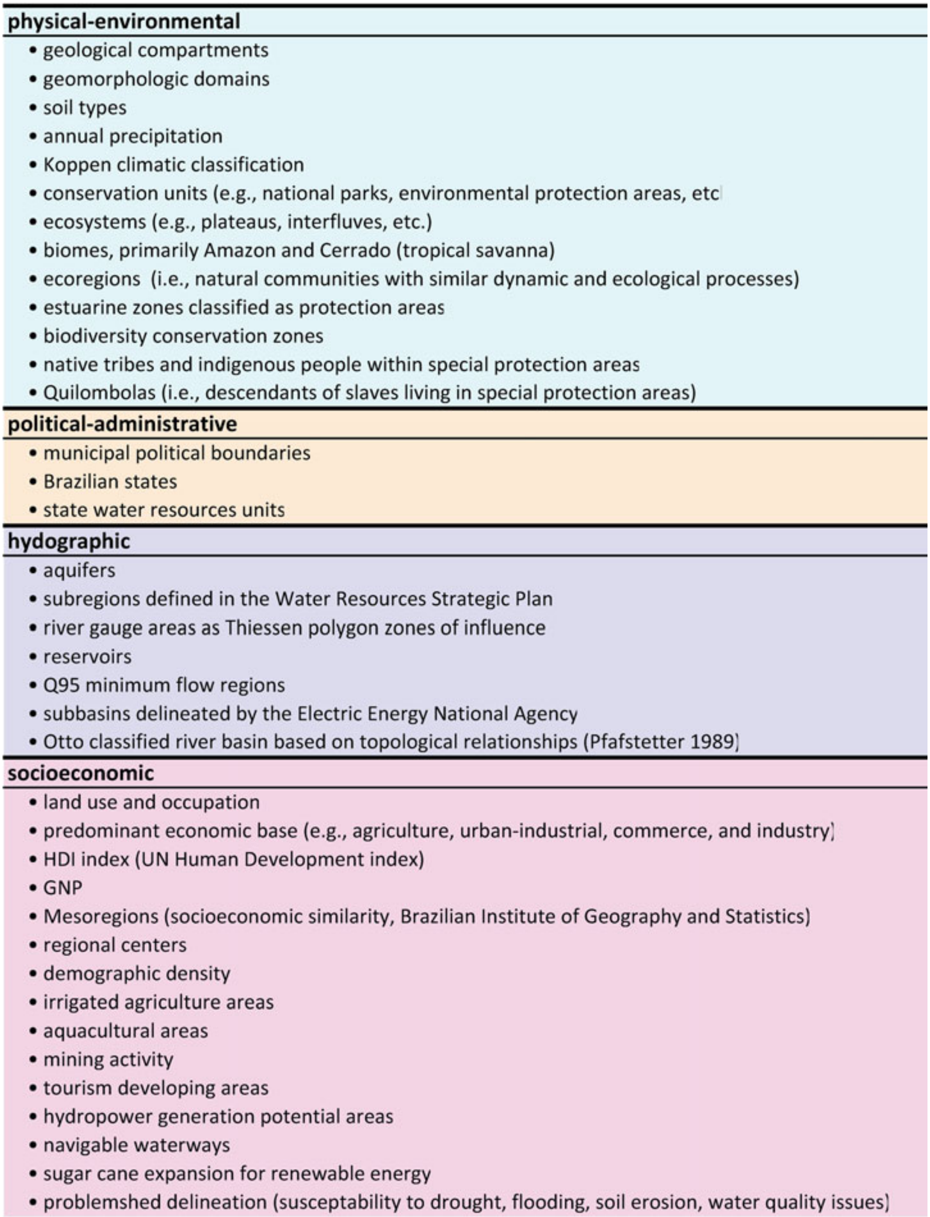
5 Application of WARPLAM DSS to the Tocantins-Araguaia River Basin, Brazil

Tocantins-Araguaia River Basin is the second largest in Brazil with a drainage area of 918,822 km^2 comprising 11% of the total area of Brazil, and a mean annual discharge of 13,800 m^3/s representing 8% of the country's total annual flow. The drainage basin of the Tocantins-Araguaia River is covered by the important Amazon Forest and Cerrado biomes, receives an average precipitation of 1,733 mm per year, and intersects six states: Pará, Tocantins, Goiás, Mato Grosso, Marahão and Distrito Federal (Fig. 6).

The Tocantins-Araguaia River Basin is designated as one of the 12 National Hydrographic Regions of Brazil, with the states further subdividing territories into 43 water resources units for planning and management purposes. The state units differ considerably in terms of scale, ranging from second or third level subbasins to small catchment areas. An example of the inconsistent scaling can be found in comparing the state of Pará comprised of three water resources units, with Tocantins which is divided into 30 water resources units, despite the fact that both states have approximately the same territorial area. This is indicative of a common problem in Brazil, with highly disparate scaling and non-harmonized delineation of water resources regions among the states and at the federal level.

The first step in the application of WARPLAM DSS is selection of a consistent planar partition representing the base units for clustering and aggregation analysis for delineating water resources planning and management regions. The Water Resources Strategic Plan of the Tocantins-Araguaia River Basin (ANA 2009) defined 17 base units for water resources planning considering hydrographic basin limits, available homogeneous hydrologic information, and existing hydropower generation plants. For this case study, it was assumed that these units represent relatively homogeneous partitions that are the appropriate building blocks for regionalization for the purpose of integrated water resources management

The next step is specification of the criteria used for defining the homogeneity of alternative pairs of base units in the planar partition using the heuristic knowledge structure provided by WARPLAM DSS. For example, considering that Brazil is a Federative country,



physical-environmental

- geological compartments
- geomorphologic domains
- soil types
- annual precipitation
- Koppen climatic classification
- conservation units (e.g., national parks, environmental protection areas, etc
- ecosystems (e.g., plateaus, interfluves, etc.)
- biomes, primarily Amazon and Cerrado (tropical savanna)
- ecoregions (i.e., natural communities with similar dynamic and ecological processes)
- estuarine zones classified as protection areas
- biodiversity conservation zones
- native tribes and indigenous people within special protection areas
- Quilombolas (i.e., descendants of slaves living in special protection areas)

political-administrative

- municipal political boundaries
- Brazilian states
- state water resources units

hydographic

- aquifers
- subregions defined in the Water Resources Strategic Plan
- river gauge areas as Thiessen polygon zones of influence
- reservoirs
- Q95 minimum flow regions
- subbasins delineated by the Electric Energy National Agency
- Otto classified river basin based on topological relationships (Pfafstetter 1989)

socioeconomic

- land use and occupation
- predominant economic base (e.g., agriculture, urban-industrial, commerce, and industry)
- HDI index (UN Human Development index)
- GNP
- Mesoregions (socioeconomic similarity, Brazilian Institute of Geography and Statistics)
- regional centers
- demographic density
- irrigated agriculture areas
- aquacultural areas
- mining activity
- tourism developing areas
- hydropower generation potential areas
- navigable waterways
- sugar cane expansion for renewable energy
- problemshed delineation (susceptability to drought, flooding, soil erosion, water quality issues)

Fig. 7 Major criteria and associated sub-criteria for regionalization in the Tocantins-Araguaia River Basin, Brazil

the user is guided through an IF-THEN rule suggesting that limits on Federative States should be considered in the analysis. For political-administrative aspects, if the federative condition is true, then each State has its own regions for water resources planning and management that should be taken into account in the analysis. In addition, if municipalites or other administrative levels have viable competence in water resources planning and management, these boundaries should also be included. Heuristic rules may also guide the selection of a specific map classification as a single subcriterion. Figure 7 shows all sub-criteria under each of the major criteria classifications for this study, including physical-environmental, political-administrative, hydrographic, and socioeconomic. All of these sub-criteria are defined as polygon feature classes developed from datasets provided by ANA (2009).

The GIS intersection and polygon-to-line operations are next performed to provide the spatial data necessary for application of Eq. 1 for calculation of measures of homogeneity H_{jk}^{c} for all adjacent pairs (j,k) and for each criterion c. Equation 2 is then applied to calculation of the composite measures of homogeneity H_{jk} over all criteria c based on user-specified weights α_c , followed by the intra-cluster homogeneity B_i of each unique pair of base elements in cluster i. The DP clustering algorithm is then applied for various sets of weights, thereby providing a wide range of alternative regionalization scenarios for integrated water resources management.

The first scenario analyzed in this case study is based on assuming identical weights for all categories of criteria. Figure 8a shows the results of this analysis, with the optimum number of clusters as six and the set {6,3,2,2,2,2} representing the number of base unit elements in each cluster. Although this scenario produces the maximum inter-cluster measure of homogeneity of 0.347, there is an imbalance in the number of elements in each

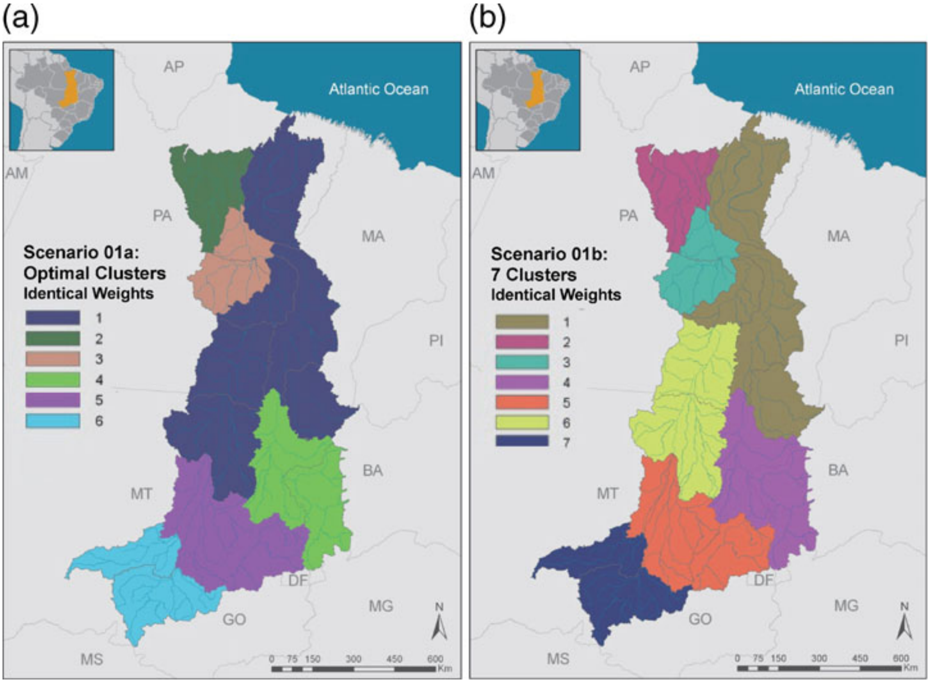
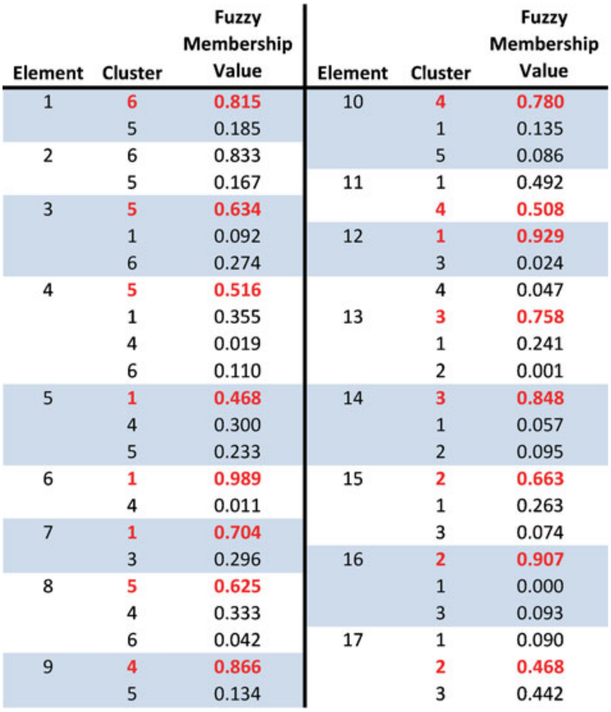


Fig. 8 **a** Optimal clustering for identical weights; **b** optimal clustering with specification of seven clusters

cluster that may not be considered ideal by WARPLAM users. Figure 8b shows the results of the same scenario with the equal-weighting assumption, but with extraction of the optimal solution of the DP optimal clustering algorithm corresponding to seven clusters. Although the optimum inter-cluster benefit is slightly reduced to 0.337 in this case, the distribution of the number of elements in each cluster {4,3,2,2,2,2} may be preferable to water planners. Again, the advantage of the DP-based clustering algorithm is that these alternative scenarios are automatically produced as a part of the DP recursive solution procedure.

As a type of sensitivity analysis, WARPLAM DSS also calculates fuzzy membership functions representing the "degree of truth" that a base unit element should belong to a particular cluster. This provides a measure of the uncertainty associated with the clustering decision, considering the subjective criteria and the qualitative specification of relative weights to each criterion. It is clear, in this context, that the elements have a continuous grade of membership within clusters, representing situations that do not completely fulfill the quantitative results or have no sharp boundaries. For the 17 element example under equal weighting, the table shown in Fig. 9 is generated following execution of the clustering algorithm, providing the fuzzy membership values associated with assigning elements to alternate clusters. It is interesting that Fig. 9 suggests the alternate distribution {4,3,2,3,3,2} of the number of elements in each cluster by clustering elements based on the highest fuzzy membership value, even though this gives a lower overall inter-cluster measure of homogeneity.

The next scenario again applies equal weighting for all criteria, but allows user specified limits to the number of elements per cluster to four (Fig. 10a) and five (Fig. 10b). For these cases, it is evident that reducing the maximum number elements in any cluster produces



Element	Cluster	Fuzzy Membership Value
1	6	0.815
	5	0.185
2	6	0.833
	5	0.167
3	5	0.634
	1	0.092
	6	0.274
4	5	0.516
	1	0.355
	4	0.019
	6	0.110
5	1	0.468
	4	0.300
	5	0.233
6	1	0.989
	4	0.011
7	1	0.704
	3	0.296
8	5	0.625
	4	0.333
	6	0.042
9	4	0.866
	5	0.134
10	4	0.780
	1	0.135
	5	0.086
11	1	0.492
	4	0.508
12	1	0.929
	3	0.024
	4	0.047
13	3	0.758
	1	0.241
	2	0.001
14	3	0.848
	1	0.057
	2	0.095
15	2	0.663
	1	0.263
	3	0.074
16	2	0.907
	1	0.000
	3	0.093
17	1	0.090
	2	0.468
	3	0.442

Fig. 9 Fuzzy membership values of each element to alternate clusters

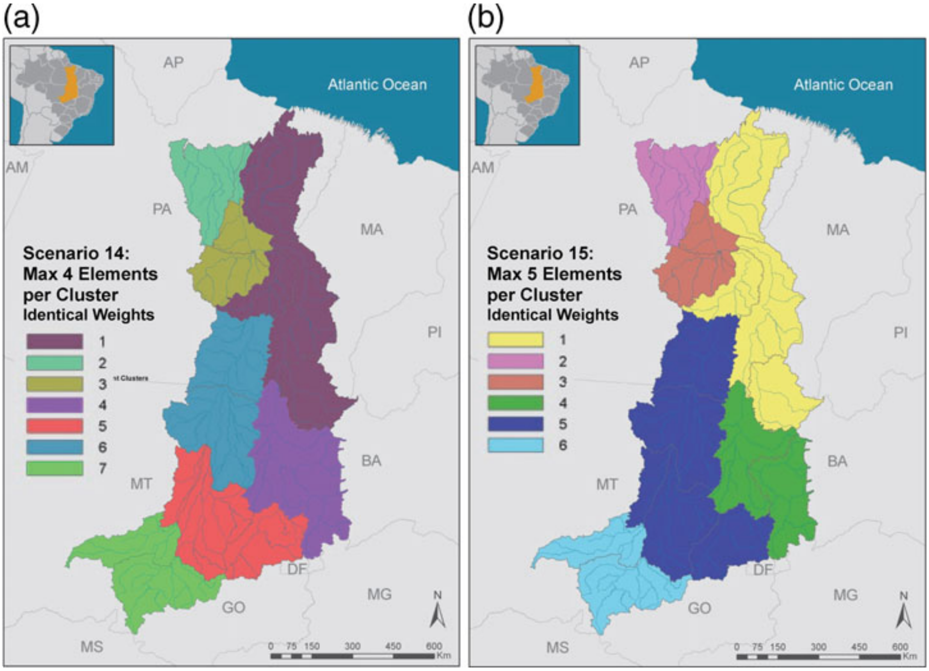


Fig. 10 **a** Maximum number of elements per cluster=4; **b** maximum number of elements per cluster=5

more uniformly sized clusters. In these cases, the inter-cluster homogeneity measure is only slightly reduced by limiting the maximum number of clusters to four (reduced from 0.347 to 0.344), as well as for a maximum of five elements per cluster (reduced from 0.347 to 0.337).

In order to test the influence of changes in the weighting factors reflecting the relative importance of each primary criterion, weights are assigned for the next scenario that emphasize socioeconomic aspects, as shown in Fig. 11a. In this case, the optimal number of clusters is seven, but with a relatively low inter-cluster homogeneity measure of 0.284 occurring for a distribution of {3,2,2,4,2,2,2} of the number of elements in each cluster. Emphasizing political-administrative aspects results in the clustering shown in Fig. 11b, with a distribution of elements {6,5,2,2,2} and an optimum inter-cluster measure of homogeneity of 0.403.

6 Summary and Conclusions

The lack of uniform and integrated water resources regions that support integrated water resources planning and management (IWRM) within a river basin is a critical issue. Presented herein is the water resources planning and management (WARPLAM) DSS that has been developed with recognition of the multidimensional character of regionalization in support of IWRM. Although it is important to define appropriate territorial units with consideration of the capacity, articulation, and needs of the existing institutional structure, there is also the necessity of incorporating more comprehensive hydrographic, physical-environmental, socioeconomic, political-administrative, and cultural-historical criteria in the regionalization decision process.

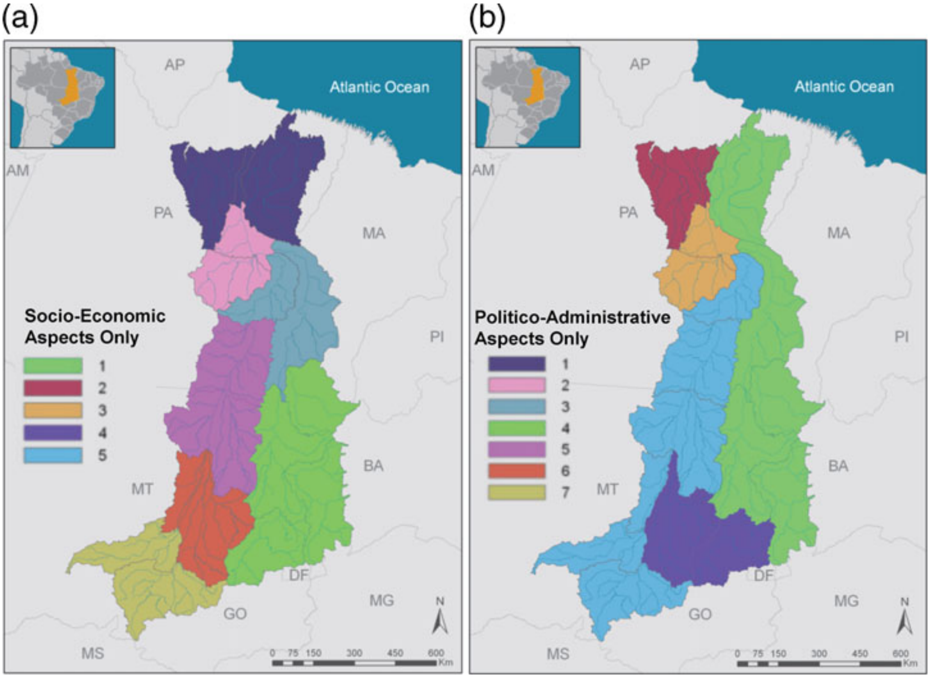


Fig. 11 **a** Optimal clustering under socioeconomic emphasis; **b** optimal clustering under political-administrative emphasis

The spatial extent and distribution of the selected criteria as to how they overlap adjacent pairs of base units in the planar partition (e.g., delineated catchments) is efficiently processed in WARPLAM using a geographic information system. The GIS processed information provides the basis for defining measures of homogeneity with respect to all relevant criteria for alternative clustering decisions of the base planar partition. Measures of homogeneity for all possible clustered pairs of base elements are defined as membership values in a fuzzy set, which are than combined as composite measures of homogeneity over all criteria as ranked by the planners and decision makers in order of importance to the regionalization process. A modified dynamic programming algorithm is then applied to providing multiple families of clustering solutions based on these subjective preferences, including determining the optimal number of clusters under specified limits on the number of elements assigned to any cluster. The term “optimal” is used in a limited sense as reflecting the best solutions under subjective preferences and priorities associated with user specified criteria governing regionalization decisions.

Results of the Tocantins-Araguaia River Basin case study application provide clear evidence of GIS as an essential spatial processing element of WARPLAM allowing an interdisciplinary focus considering physical-environmental, political-administrative, hydrographic, and socioeconomic criteria with numerous associated subcriteria. The application of fuzzy set theory was advantageous in representing the uncertainty associated with the clustering decisions stemming from the qualitative ranking of the subjective criteria without requiring application of probability measure theory. The contribution of the modified dynamic programming algorithm for optimal clustering solutions offers a distinct advantage over other alternative clustering methods in providing consistent, global optimal solutions for multiple clustering scenarios based on user preferences.

Finally, it is believed that in addition to development of new regionalization scenarios in support of IWRM, WARPLAM DSS can be useful as a *critiquing* DSS in that existing regionalization plans can be analyzed as a means of evaluating the underlying logic behind the decision processes employed. This can provide a better understanding of which criteria were implicitly weighted the highest in the regionalization decisions, and afford a degree of confirmation of the validity of those decisions. On the other hand, it might pinpoint certain important criteria that were neglected in the analysis, or perhaps identify less important criteria that played an inordinate role. Future work on WARPLAM DSS will include development of a more robust user interface that is fully integrated with ArcGIS™.

Acknowledgements The authors would like to thank the Fulbright International Educational Exchange Program of the U.S. Department of State, the Brazilian National Water Agency, and the Brazilian Agency for Improvement of Higher Education Personnel (CAPES) for their financial support through grant 2276/05-4 of the first author in her research program leading to development of WARPLAM DSS.

References

Allan JA (2005) Water in the environment/socio-economic development discourse: sustainability, changing management paradigms and policy responses in a global system. Government and opposition. Blackwell Publishing, Oxford

Allende TC, Mendoza ME, Lopez-Granados E, Morales-Manilla L (2009) Hydrogeographical regionalization: an approach for evaluating the effects of land cover change in watersheds. A case study in the Cuitzeo Lake Watershed, Central Mexico. Water Resources Management 23:2587–2603

ANA (2009) Plano Estratégico de Recursos Hídricos da Bacia Hídrográfica dos Rios Tocantins e Araguaia: Relatório Síntese. Brasília: Agência Nacional de Águas, Ministério do Meio Ambiente

Bellman R (1973) A note on cluster analysis and dynamic programming. Math Biosci 18:311–312

Coelho AC (2010) Multicriteria decision support system to delineate water resources planning and management regions. PhD Dissertation. Colorado State University, Fort Collins

Coelho AC, Gontijo WC, Cardoso A (2005) Unidades de planejamento e gestão de recursos hídricos: Uma proposta metodológica. In: Anais 7o Simpósio de Hidráulica e Recursos Hídricos dos Países de Língua Oficial Portuguesa, Silusba, Portugal

Esogbue AO (1986) Optimal clustering of fuzzy data via fuzzy dynamic programming. Fuzzy Set Syst 18:283–298

Everitt BS, Landau S, Leese M (2001) Cluster analysis. Oxford University Press, New York

Global Water Partnership (2000) Towards water security: a framework for action. Stockholm, Sweden

Haunert JH, Wolff A (2010) Area aggregation in map generalisation by mixed-integer programming. Int J Geogr Inf Sci 24(12):1871–1897

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York

Labadie JW (2003) Generalized dynamic programming package: CSUDP: documentation and user guide, Version 2.44. Department of Civil Engineering, Colorado State University, Fort Collins

Lee J-H, Labadie JW (2007) Stochastic optimization of multireservoir systems via reinforcement learning. Water Resources Research 43(W11408)

Maulik U, Bandyopadhyay S (1999) Genetic algorithm-based clustering technique. Pattern Recogn 33:1455–1465

Mostert E, Craps M, Pahl-Wostl C (2008) Social learning: the key to integrated water resources management? Water Int 33(3):293–304

Wiering M, Verwijmeren J, Lulofs K (2010) Experiences in regional cross border co-operation in river management. Comparing three case studies at the Dutch-German border. Water Resour Manag 24:2647–2672

Xu R, Wunsch D (2009) Clustering. IEEE Press, Piscataway