A spatially constrained clustering program for river valley segment delineation from GIS digital river networks

T.O. Brenden a,*, L. Wang a, P.W. Seelbach a, R.D. Clark Jr. a, M.J. Wiley b, B.L. Sparks-Jackson b

a Institute for Fisheries Research, University of Michigan and Michigan Department of Natural Resources, 212 Museums Annex, Ann Arbor, MI 48109, USA
b School of Natural Resources and Environment, University of Michigan, G166 Dana Building, Ann Arbor, MI 48109, USA

Received 30 November 2006; received in revised form 4 September 2007; accepted 5 September 2007

Abstract

River valley segments are adjacent sections of streams and rivers that are relatively homogeneous in hydrology, limnology, channel morphology, riparian dynamics, and biological communities. River valley segments have been advocated as appropriate spatial units for assessing, monitoring, and managing rivers and streams for several reasons; however, methods for delineating these spatial units have been tedious to implement or have lacked objectivity, which arguably has limited their use as river and stream management units by natural resource agencies. We describe a spatially constrained clustering program that we developed for delineating river valley segments from geographic information system digital river network databases that is flexible, easy-to-use, and improves objectivity in the river valley segment delineation process. This program, which we refer to as the valley segment affinity search technique (VAST), includes a variety of options for determining spatial adjacency in stream reaches, as well as several data transformation methods, types of resemblance coefficients, and cluster linkage methods. The usefulness of VAST is demonstrated by using it to delineate river valley segments for river network databases for Michigan and Wisconsin, USA, and by comparing river valley segments delineated by VAST to an expert-opinion delineation previously completed for a Michigan river network database.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Spatially constrained clustering; Cluster affinity search technique; River valley segment; Digital river network; Stream management

Software availability

Program title: VAST
Developer: Travis Brenden
Hardware required: IBM-compatible computer system running Windows 2000 or higher
Software required: Microsoft Excel Version 10 or higher, Morefunc and Poptools Microsoft Excel Add-Ins
First available: July 2007

1. Introduction

One of the most important, yet challenging, aspects of river research and management is the identification of appropriate spatial units for sampling design, data interpolation, and formulation of management actions (Wang et al., 2006). Whereas lentic systems (e.g., lakes, ponds) have readily identifiable shoreline boundaries, the interconnectivity of stream reaches...
within a river network makes it difficult to identify distinct sampling or management units. While some might question whether distinct spatial units exist in rivers given our functional understanding of rivers (e.g., river continuum concept), the belief that rivers and streams consist of many small, relatively distinct, ecological units is common (Pringle et al., 1988; Maxwell et al., 1995; Malmqvist, 2002; Benda et al., 2004). It perhaps could even be argued that several aspects of stream and river management, such as communicating with the public regarding management actions or setting spatial boundaries for enacted regulations, necessitate a view that rivers and streams comprise a mosaic of distinct ecological units.

Because river networks comprise a hierarchical arrangement of ecosystems (Frissell et al., 1986), the most appropriate spatial unit for which to base stream and river management decisions can be a subject of debate (Dovciak and Perry, 2002; Fausch et al., 2002). Although sampling of stream habitat and biota is often conducted over relatively short distances (less than a few 100 m), management is considered most effective when decisions are conceptualized at spatial scales on the order of several 1000 m (Fausch et al., 2002; Seelbach et al., 2006). This length of stream is similar to the distance at which river valley segments are believed to exist (Frissell et al., 1986) and over which sufficient sampling for fish species richness or index of biotic integrity metrics is needed (Cao et al., 2001; Hughes et al., 2002; Hughes and Herlihy, 2007). River valley segments are adjacent sections of streams and river that are relatively homogeneous in hydrology, limnology, channel morphology, riparian dynamics, and biological communities (Frissell et al., 1986; Maxwell et al., 1995; Seelbach et al., 2006). River valley segments form as a result of streams and rivers flowing long distances across landscapes with abrupt boundaries in surficial geology, bedrock geology, landscape topography, and land cover characteristics. River valley segments also can form at junctures of unrelated hydrologic systems or from anthropogenic modifications to stream channels, each of which can cause distinct changes in chemical, thermal, and material load conditions in streams (Ward and Stanford, 1983; Minshall et al., 1985; Frissell et al., 1986). River valley segments have been used as spatial units for a variety of water resource monitoring and management purposes, including clarifying habitat requirements of both endangered and sport fish species (Stanfield et al., 2006; Wall and Berry, 2006), understanding how landscape characteristics influence local-scale habitat features (Burnett et al., 2006), and developing stream and river classifications (Seelbach et al., 2006).

Two approaches primarily have been used to delineate river valley segments for river networks, an expert-opinion and an automated class-based approach. With an expert-opinion approach, aquatic ecologists familiar with river systems for a particular region use a geographic information system (GIS) to visualize river network maps in relation to thematic maps depicting landscape characteristics such as elevation, land cover, surficial geology, and soil texture. The aquatic ecologists then use their knowledge of what factors affect biological assemblages to make subjective decisions regarding the placement of river valley segment boundaries. An expert-opinion approach has been used to delineate river valley segments in both the upper (Baker, 2006) and lower (Seelbach et al., 2006) peninsulas of Michigan, as well as in several other Midwestern US states (M. DePhilip, The Nature Conservancy, personal communication). A perceived advantage of the expert-opinion approach for delineating river valley segments is that vast amounts of existing knowledge regarding relationships between biological assemblages and landscape characteristics can be incorporated in the delineation process. An expert-opinion approach also permits substantial flexibility in the delineation of river valley segments, which otherwise might not be possible with an automated approach. One of the disadvantages with this approach is the difficulty in replicating the delineation process for a particular region because of its subjective basis. Multiple sets of aquatic ecologists likely will interpret the same maps differently, which may lead to very different river valley segment partitions for an individual river network.

With an automated class-based approach, a GIS also is used to display river network maps in relation to thematic maps depicting landscape characteristics. The river network is then partitioned into river valley segments through an automated GIS function that inserts a breakpoint wherever the river network crosses a class boundary on one of the thematic maps. An automated class-based approach has been used to delineate river valley segments in Missouri (Sowa et al., 2007), South Dakota (Wall et al., 2004), and Ontario (Kilgour and Stanfield, 2006; Stanfield et al., 2006). For delineating river valley segments, an automated class-based approach will be quicker to implement and easier to replicate compared to an expert-opinion approach. The primary challenge associated with an automated class-based approach is in objectively defining the class boundaries for the landscape characteristics such that the class boundaries reflect real world thresholds for aquatic biota. Although conceptually appealing, real world identification of ecological thresholds can be complicated due to issues of scale, variable interactions, and nonlinear biotic–abiotic relationships (De’ath, 2002; Groffman et al., 2006).

An alternative to the expert-opinion and automated class-based approaches for delineating river valley segments is to use a statistical clustering procedure, such as K-means or model-based clustering, to group stream reaches with similar physicochemical and biological properties. The purpose of many clustering methods is to group objects within a data set, such that objects within a group are homogeneous and have distinct differences from the objects in other groups (Manly, 1994; Legendre and Legendre, 1998; Ben-Dor et al., 1999), which is similar to how river valley segments are defined. Using a statistical clustering procedure to identify river valley segments would be beneficial because of the efficiency, repeatability, and objectivity of the delineation process. The major challenge in using a statistical clustering procedure to delineate river valley segments is that most clustering routines included in statistical software packages assume that every object potentially can be grouped with every other object within a data set. For delineating river valley segments, though, such...
an assumption is inappropriate since it could result in groupings of river reaches that are not spatially adjacent and thus cannot actually constitute a river valley segment (Fig. 1). Rather, a spatially constrained clustering procedure, one that recognizes that only spatially adjacent objects can be grouped within clusters (Legendre and Legendre, 1998), is needed. Despite the perceived usefulness of spatially constrained clustering procedures (Legendre and Legendre, 1998), availability of computer programs for implementing this type of clustering is limited.

The purpose of this paper is to describe a spatially constrained clustering program that we developed for delineating river valley segments from digital river network databases. The clustering algorithm used in the program is based on the cluster affinity search technique. We illustrate the use of this clustering program by delineating river valley segments from digital river network databases for Michigan and Wisconsin, USA. Additionally, we compare the river valley segment partitions identified by our program to a delineation previously conducted by expert-opinion for a Michigan river network database (Baker, 2006; Seelbach et al., 2006).

2. Cluster affinity search technique (CAST)

The cluster affinity search technique (CAST) was originally developed by Ben-Dor et al. (1999) for clustering gene expression data. CAST is a non-hierarchical clustering method, similar to $K$-means, meaning that its intent is to partition a set of $n$ objects into $K$ groups such that objects within groups are more similar than objects in different groups (Ben-Dor et al., 1999). CAST is an agglomerative clustering routine, meaning all objects are initially considered separate from each other. Clusters are formed one at a time, with objects being both added to and removed from open clusters (an open cluster is the cluster currently being formed). An object is added to an open cluster if its affinity (i.e., similarity) to the cluster is within the bound set by an affinity threshold, which must be specified in advance by the user. An object is removed from an open cluster if its affinity to the cluster goes outside the bound set by the affinity threshold. A cluster remains open until no more objects can be added to or removed from the cluster. A cluster is then closed and formation of a new cluster begins. Once all objects have been assigned to a cluster, CAST checks whether the affinity of the objects to their current clusters is greater than their affinity to other clusters. If an object’s affinity to another cluster is greater than its affinity to its currently assigned cluster, then the cluster assignment of the object is changed to that of the other cluster. This reassignment of objects to clusters continues until all objects are assigned to their highest affinity cluster or until some maximum number of reassignments has been reached (Ben-Dor et al., 1999). Additional details regarding the CAST algorithm, example applications, and comparisons with other clustering methods can be found in Ben-Dor et al. (1999), Yeung and Ruzzo (2001), Bellaachia et al. (2002), and Tseng and Kao (2004).

CAST is an appealing clustering algorithm for delineating river valley segments for several reasons. First, unlike some other popular clustering procedures, the number of groups to which objects are assigned does not need to be specified in advance, which can be a difficult task particularly for data sets consisting of large numbers of poorly-separated objects. CAST does require advance specification of the affinity threshold, which also requires initial insight into cluster structure (Bellaachia et al., 2002). However, this feature also may be useful when processing multiple data sets, as it will help to ensure similar levels of within cluster variability across data sets. An additional strength of the CAST clustering algorithm is its dropping of objects from open clusters during cluster formation and its reassignment of objects to other clusters at the end of the clustering procedure. Many other clustering methods are considered “greedy” because once objects are assigned to clusters the associations are permanent (Ben-Dor et al., 1999). Because CAST allows objects to be removed from clusters, this helps to ensure that objects are associated with their most alike clusters and that the final partition of objects is not overly influenced by cluster formation order.

3. Valley segment affinity search technique (VAST)

Because our method for delineating river valley segments is based on CAST, we refer to the clustering algorithm as the valley segment affinity search technique (VAST). VAST is very

Fig. 1. Example results of a regular cluster analysis approach for delineating river valley segments for the Ausable river system in Michigan. $K$-means clustering (number of clusters $= 100$) was used to identify grouped objects. The stream reaches in bold were one of the groupings found by the clustering procedure. Because the stream reaches are not spatially adjacent, this grouping of reaches does not constitute a river valley segment.
similar to CAST: river valley segments (i.e., clusters) are formed one at time, stream reaches (i.e., objects), defined herein as interconfluence stretches of water, are added to open river valley segments based on the affinity of the reaches in relation to an affinity threshold, stream reaches can be removed from open river valley segments, and stream reaches are reassigned to different river valley segments at the end of the routine. There are two major differences between CAST and VAST. First, with VAST, only stream reaches that are spatially adjacent to the open river valley segment are added to the open cluster. The other major difference is the timing as to when stream reaches are dropped from open river valley segments. With CAST, objects can be removed whenever a cluster is open. When delineating river valley segments, though, the removal of stream reaches from open valley segments may be problematic as the remaining stream reaches may not be spatially adjacent and thus may not actually compose a river valley segment. After multiple removal steps, it is likely that several river valley segments could be open at the same time. To prevent such a situation from occurring, removal of stream reaches from open river valley segments in VAST occurs only after no additional stream reaches can be added to the river valley segment. Once no additional stream reaches can be added to the open river valley segment, VAST checks the affinity of each stream reach to the open river valley segment. If the affinity of a stream reach to the open river valley segment is beyond the boundary designated by the affinity threshold, then that stream reach is flagged. Formation of the open river valley segment then begins anew. As the open river valley segment is reformed, flagged stream reaches are prevented from joining the river valley segment. This also prevents reaches that are either downstream or upstream from the flagged stream reach (depending on the location of the flagged stream reach in relation to the cluster starting point) from being added to the open river valley segment. Although preventing a flagged reach from joining an open river valley segment may be undesirable as the stream reach may not be initially assigned to its highest affinity river valley segment, the reassignment of stream reaches to river valley segments at the end of the clustering process should help to ensure that stream reaches are ultimately assigned to their most similar river valley segments.

The reassignment of stream reaches among river valley segments at the end of the clustering process occurs according to several rules. A stream reach will be reassigned to an adjacent river valley segment if the stream reach’s affinity to the adjacent river valley segment is greater than its affinity to its currently assigned river valley segment, and as long as its addition to the adjacent river valley segment does not expel other stream reaches from the cluster. In other words, the only way for a stream reach to be expelled from its currently assigned river valley segment is if it has a greater affinity to an adjacent river valley segment. If the stream reach to be reassigned is its own river valley segment, then it will be reassigned to an adjacent river valley segment if its affinity to the adjacent river valley segment is within the bound set by the affinity threshold and so long as it does not expel a stream reach already assigned to the river valley segment. When a stream reach is reassigned to a different river valley segment, there does exist the possibility that the stream reaches that remain in the river valley segment may not all be adjacent. In such case, the remaining stream reaches are clustered using the same process originally used to delineate the river valley segments.

4. VAST program design and operation

We programmed the VAST clustering algorithm in Microsoft Excel’s Visual Basic for Applications (VBA). We chose this programming environment because of its accessibility and the routine use of Microsoft Excel, so those interested in using VAST should have little difficulty in running or modifying the program if needed. VAST does make use of several functions that are packaged in two free Microsoft Excel Add-Ins, Morefunc (http://xcel105.free.fr/english/) and Pootools (http://www.cse.csiro.au/pootools/). Before using VAST, these Add-Ins must be downloaded, installed, and loaded into Excel.

VAST was developed specifically for delineating river valley segments from GIS digital river network databases. Within a GIS, river networks consist of multiple line features, which are sometimes referred to as reaches or arcs. Each reach has an associated from- and to-node, which indicates direction of water flow and can be used to identify spatial adjacency of the reaches. Areal features (i.e., lakes, ponds, and reservoirs) also may have centerline representations, so spatial adjacency of stream reaches to lakes, reservoirs, and ponds also can be determined through from- and to-node information. Methods for attributing stream reaches with environmental data in a GIS are described in Brenden et al. (2006).

A VBA UserForm provides the graphical user interface for the VAST program (Fig. 2). Data contained in an Excel worksheet can be read into VAST through several VBA RefEdit controls, which mimic the behavior of Excel’s reference edit boxes. VAST requires the following information to delineate river valley segments from a river network database: a unique integer value identifier for the reaches, from- and to-node information for the reaches, Strahler stream order and link number (variables that approximate stream reach size in a river network), and the environmental attribute data. Depending on which options are selected by the user, Strahler stream order and link number may not actually be used by VAST. In such cases, users can input artificial data for these fields. VAST requires at least two environmental attributes for the stream reaches to delineate river valley segments.

Worksheet data ranges for the requisite field are loaded into VAST by clicking the “Load Data” button. Upon loading the data, VAST conducts several quality-control checks to ensure the data are properly formatted. Users are prompted to verify the correct number of environmental attributes have been read into the program. VAST also verifies that equal numbers of observations have been read into the fields, and that none of the fields contain missing entries.

If the data pass the initial quality-control check, the user is then asked whether adjacency of stream reaches should be
based only on linear adjacency. For two stream reaches to be considered linearly adjacent, the from-node for a stream reach must equal the to-node for another stream reach. Two stream reaches that share the same from- or to-node are not considered linearly adjacent (Fig. 3). Once a stream reach adjacency type has been selected, the user is then prompted as to whether stream reach adjacency should be constrained by stream linkages. If this option is selected, for each spatially adjacent pair of stream reaches VAST calculates the proportional difference in stream reach link numbers

\[ P_{ij} = \frac{\text{Link}_{\text{Max}} - \text{Link}_{\text{Min}}}{\text{Link}_{\text{Min}}} \times 100\% \]  

where \( P_{ij} \) is the proportional difference in stream reach link numbers for reaches \( i \) and \( j \), and \( \text{Link}_{\text{Max}} \) and \( \text{Link}_{\text{Min}} \) are the maximum and minimum stream reach link numbers for reaches \( i \) and \( j \), respectively. For example, two neighboring stream reaches with stream reach link numbers of 10 and 8 would have a proportional difference of 25%. For reaches with stream reach link numbers of 10 and 5, the proportional difference would be 100%. Users must then designate what maximum \( P_{ij} \) is needed for stream reaches to be considered neighbors. The use of proportional differences in stream reach links for limiting which stream reaches are considered neighbors may be helpful for limiting the occurrence of branching in river valley segments and thus ultimately restrict river valley segment sizes. Once users have selected how neighboring stream reaches will be determined, VAST constructs an adjacency table based on the stream reach from- and to-node information.

After the stream reach adjacency table has been built, VAST prompts the user to select among several data transformation methods, processing orders, resemblance coefficients, and linkage methods. VAST includes two methods of data transformations, rank transformation and Z-scores transformation. Rank transformation may be useful for environmental attribute data where there is a concern about outliers, while Z-score transformation may be helpful for environmental attribute data measured in different units (Romesburg, 1984). Earlier versions of VAST also included an option for variable transformation through principal components analysis, but we no longer support this transformation method given research indicating that principal component scores can degrade cluster quality (Yeung and Ruzzo, 2001).

Processing order refers to the order in which stream reaches are grouped during river valley segment formation. VAST includes four possible processing orders: “from headwaters”, “from outflow”, “randomly”, and “based on similarity”. If “from headwaters” processing order is selected, stream
attribute data are non-negative, which is a requirement when calculating Bray–Curtis distances. If necessary, VAST adds a constant to each of the variables to ensure that the attribute data are positive values. It is beyond the scope of this paper to discuss either the advantages or disadvantages of these resemblance coefficients; rather, users should consult sources such as Romesburg (1984), Legendre and Legendre (1998), and Krebs (1999) for information regarding these resemblance coefficients. McKenna (2003) also provides a useful and concise summary of several of these resemblance coefficients. We only note that some of these resemblance coefficients are intended for qualitative (presence/absence) attributes, while others are intended for quantitative attributes. Additionally, some of the resemblance coefficients intended for qualitative (presence/absence) data cannot be used when conjoint absences occur in the data set. We thus highly encourage a review of resemblance coefficients before using VAST to delineate river valley segments.

When calculating the affinity of a stream reach to an open valley segment, three types of linkage methods can be selected: complete linkage (CLINK), single linkage (SLINK), or unweighted pair-group method using arithmetic averages (UPGMA). With CLINK, the affinity of an individual stream reach to an open river valley segment equals its similarity to its most dissimilar stream reach within the cluster. With SLINK, the affinity of an individual stream reach to an open river valley segment equals its similarity to its most similar stream reach within the cluster. With UPGMA, the affinity of an individual stream reach to an open river valley segment equals its mean similarity to all stream reaches within the cluster. As with the resemblance coefficients, users should refer to Romesburg (1984), Legendre and Legendre (1998), and Krebs (1999) for discussions regarding the advantages and disadvantages of these linkage methods.

The final user input required by VAST is the affinity threshold value. As previously stated, the affinity threshold sets the boundary for which stream reaches are added to and dropped from open river valley segments. It is important to keep in mind when specifying the affinity threshold that some resemblance coefficients are bounded, while others are not. Thus, specification of the affinity threshold should be at least partly based on which resemblance coefficient the user has selected. It also may be advantageous to use several affinity threshold values and to compare or combine partitioning results.

The delineation of river valley segments proceeds by clicking the “Run” button. The graphical user interface for VAST includes several progress meters that allow users to monitor the status of the delineation process. Once river valley segment delineation is completed, several worksheets are added to the open Microsoft Excel workbook, including an “Output” worksheet that lists the stream reach identifier along with an integer value river valley segment identifier. Other worksheets that may be of interest also are added, including the stream reach adjacency table, a listing of the transformed environmental attribute data, and a listing of what stream reaches were reassigned to different river valley segments at the end of the clustering process.

Fig. 3. Illustration of the concept of linear adjacency in streams from the Ausable river system in Michigan. For this system, reaches 553 and 559, 658 and 601, 565 and 529, and 468 and 452 are (among others) linearly adjacent because the from-node (nodes are shown as dark filled circles) of one of the reaches equals the to-node of the other reach. Reaches 566 and 553, 615 and 658, 542 and 529, and 465 and 468 are not linearly adjacent even though these reach pairs share the same to-node.
5. Example applications

5.1. Application of VAST to Michigan and Wisconsin river network databases

We used VAST to identify river valley segments from digital river network databases for Michigan and Wisconsin, USA. The 1:100,000 scale National Hydrography Dataset (NHD; http://nhd.usgs.gov/) was the river network database used for this delineation. Identification of river valley segments was based on seven physicochemical stream attributes that were believed to be important determinants of fish distribution in Michigan and Wisconsin streams: loge transformed network catchment area, percent non-forested wetland land type in network catchments, percent lacustrine surficial geology in reach catchments, percent moraine surficial geology in reach catchments, mean reach catchment slope, predicted July mean reach water temperature, and predicted loge transformed 90th percentile reach base flow yield. Prior to partitioning the river network database into river valley segments, we standardized the stream attribute data at a statewide scale using Z-score standardization.

To delineate river valley segments, we subdivided the NHD for the states by 8-digit Hydrologic Units (Seaber et al., 1987) to form processing units. In some instances, boundaries of the 8-digit Hydrologic Units had to be modified manually to prevent adjacent river reaches from occurring within different processing units. We identified a total of 46 processing units for the Michigan NHD. Twenty-three processing units were identified for the Wisconsin NHD. The number of interconfluence stream reaches within these processing units ranged from fewer than 100 to more than 10,000. Weighted Euclidean distance (equal weights assigned to all variables) and UPGMA linkage were the resemblance coefficient and linkage method used to delineate river valley segments. Reach processing order was based on “average similarity”. Only linearly adjacent stream reaches with proportional differences in stream links less than 60% were considered spatially adjacent, although an exception was made for neighboring stream reaches with stream links of one and two. We used a range of affinity thresholds (0.6, 1.0, 1.5, and 2.0) to delineate river valley segments so that the partitions resulting from these different affinity thresholds could be compared.

For each river valley segment partition identified for the Michigan and Wisconsin NHD river networks, we calculated the Calinski and Harabasz (1974) index (CH index) as a measure of stream attribute homogeneity within the identified river valley segments. The CH index is sometimes used as a stopping rule for identifying the “optimal” number of clusters and is a function of between and within cluster sum of squares. Milligan and Cooper (1985) found the CH index performed well relative to other stopping rules in a simulation study. The CH index is calculated as:

\[ CH = \frac{SSB / k - 1}{SSW / n - k} \]  

where SSB is the between cluster sum of squares, SSW is the within cluster sum of squares, \( k \) is the number of identified clusters, and \( n \) is the number of objects. The CH index increases as within cluster variability decreases and between cluster variability increases. Although river valley segment partitioning was conducted separately for each processing unit, the calculation of the CH index was done for each of the statewide databases.

A total of 30,845 and 34,308 stream reaches (excluding shoreline and lake centerline reaches) were identifiable from the 1:100,000 scale NHD for Michigan and Wisconsin, respectively. With affinity thresholds ranging from 0.6 to 2.0, VAST identified between 15,107 and 18,542 river valley segments for Michigan (Fig. 4) and between 15,928 and 19,176 river valley segments for Wisconsin (Table 1). Mean lengths of identified river valley segments ranged from 4.48 to 5.49 km for Michigan and from 4.28 to 5.14 km for Wisconsin (Table 1). The use of VAST resulted in fairly large increases in the frequency of long stream units in both Michigan and Wisconsin (Fig. 5). With the original NHD, stream reaches longer than 4 km in length comprised less than 19.3 and 14.4% of all stream units for Michigan and Wisconsin, respectively. After clustering with VAST, river valley segments that were longer than

Fig. 4. River valley segments identified for the Ausable river system in Michigan using VAST with an affinity threshold of 0.6. Linear adjacent stream reaches with the same line colors form the river valley segments. River valley segments of the same color are simply an artifact of the limited number of unique color combinations in the software used to generate the map, and should not be interpreted as meaning the river valley segments are of the same type.
The CH index values decreased as affinity thresholds increased for both the Michigan and Wisconsin river network databases (Fig. 5).

4 km in length comprised between 35 and 40% of the stream units in Michigan and Wisconsin, respectively (Fig. 5).

The CH index values decreased as affinity thresholds increased for both the Michigan and Wisconsin river network databases. The maximum CH index values were 36.41 for Michigan and 44.22 for Wisconsin at an affinity threshold of 0.6 (Table 1). The minimum CH index values were 7.84 for Michigan and 8.76 for Wisconsin at an affinity threshold of 2.0 (Table 1).

### 5.2. Comparison of VAST to an expert-opinion delineation for Michigan rivers

To determine how well river valley segments identified by VAST agreed with those identified by expert-opinion, we used VAST to delineate river valley segments for a Michigan river network database that previously was partitioned into river valley segments through an expert-opinion approach (Baker, 2006; Seelbach et al., 2006). The river valley segments delineations by Baker (2006) and Seelbach et al. (2006) were conducted on the 1:100,000 scale US Environmental Protection Agency’s Reach File 3 (RF3) hydrography data set. River valley segments were identified for the RF3 data set using the following landscape and river channel attributes: surficial geology, catchment slope, catchment land use, valley width, valley wetlands, channel sinuosity, and potential groundwater influx to river channels (Seelbach et al., 2006). Not all stream reaches on the RF3 river network database were assigned to a river valley segment by Baker (2006) and Seelbach et al. (2006); small, headwater streams were generally excluded from the river valley segment delineation process.

Because some of the landscape characteristic databases that were used in the expert-opinion river valley segment delineations were no longer available to us, we transferred the river valley segment boundaries identified for the RF3 river network database to the attributed NHD river network database for Michigan described in Section 5.1. This transfer of river valley segment boundaries was conducted in a GIS by converting the RF3 river network map to a 30 m pixel raster map in which individual pixels were assigned the river valley segment identifier for the reaches they overlaid. We then overlaid the NHD river network map on the RF3 raster map and used a GIS to transfer the RF3 river valley segment identifiers for the pixels to the NHD stream reaches. If an NHD stream reach overlaid several grid pixels with different river valley segment identifiers, then a majority rule was used to assign a single river valley segment identifier to that stream reach. Only those NHD stream reaches that were assigned an RF3 river valley segment identifier were used to delineate river valley segments with VAST. The same stream reach attributes and VAST configurations described in Section 5.1 were used to delineate river valley segments for this comparison of delineation approaches.

Comparisons between the river valley segment delineations were based on adjusted (chance-corrected) and unadjusted Rand indices of agreement (Rand, 1971; Hubert and Arabie, 1985). The unadjusted Rand index of agreement compares the results of two clustering methods \( V_1 \) and \( V_2 \) based upon how often object pairs are grouped by none, one, or both methods. Specifically, the unadjusted Rand index is calculated as:

\[
R = \frac{a + d}{a + b + c + d},
\]

where:

- \( a \) is the number of objects classified in the same groups by both methods;
- \( b \) is the number of objects classified in one group by one method and in another group by the other method;
- \( c \) is the number of objects classified in different groups by the two methods;
- \( d \) is the number of objects that are unclassified by both methods.

![Fig. 5. Frequencies of occurrence of stream units (reaches or river valley segments) of various lengths for the raw (unclustered) Michigan (top panel) and Wisconsin (bottom panel) NHD river network databases and for river valley segments identified using VAST with affinity thresholds of 0.6, 1.0, 1.5, and 2.0.](image)
where $R$ is the index of agreement between $V_1$ and $V_2$, $a$ is the number of pairs of objects within the same cluster in both $V_1$ and $V_2$, $b$ is the number of pairs of objects within the same cluster in $V_1$ but not $V_2$, $c$ is the number of pairs of objects within the same cluster in $V_2$ but not $V_1$, and $d$ is the number of pairs of objects that are not within the same cluster in either $V_1$ or $V_2$. The unadjusted Rand index of agreement ranges from 0 to 1, with 1 indicating perfect agreement between the clustering methods. Hubert and Arabie’s (1985) adjustment to Rand index accounts for clustering results agreeing between the two methods purely by chance, and typically results in a much lower agreement rate than the unadjusted Rand index. When calculating agreement between the river valley segment delineation approaches, we limited our analysis to only those streams reaches that were spatially adjacent in order to prevent inflation of the $d$ counts. The same VAST configurations that were used to delineate river valley segment for the Michigan and Wisconsin NHD river networks were used for the RF3 river network database. For each river valley segment partition identified for the RF3 Michigan river network database, we again calculated the CH index as a means for measuring stream attribute homogeneity for the identified river valley segments.

A total of 10,714 stream reaches for Michigan were assigned a river valley segment identifier from the expert-opinion approach. Baker (2006) and Seelbach et al. (2006) identified 2632 river valley segments using an expert-opinion approach. In comparison, VAST identified between 3466 and 4734 river valley segments with affinity thresholds ranging from 0.6 to 2.0 (Table 2). Adjusted and unadjusted Rand indices of agreement ranged from 45.4 to 64.2% and from 76.9 to 87.7%, respectively (Table 3). The commission error rate, meaning that VAST clustered stream reaches that were not clustered by the expert-opinion approach, ranged from 2.7 to 3.7%. Conversely, the omission error rate, meaning that VAST did not cluster stream reaches that were clustered by the expert-opinion approach, ranged from 16.1 to 25.0% (Table 3).

Mean lengths of river valley segments identified by VAST ranged from 6.78 to 9.28 km (Table 2). In comparison, mean length of river valley segments identified by expert-opinion approach. In Fig. 6, the percent of delineated river valley segments longer than 4 km in length (Fig. 6). In comparison, the percent of river valley segments longer than 4 km in length ranged from 58 to 65% when delineations were conducted using VAST (Fig. 6). For the expert-opinion approach, 87% of delineated river valley segments were longer than 4 km in length (Fig. 6). Using VAST with an affinity threshold of 0.6 resulted in the largest CH index (46.57) of all the river valley segment partitioning methods. The second largest CH index (16.25) was from the expert-opinion approach to delineating river valley segments. For the other affinity thresholds that were used with the VAST program, the CH index declined as affinity thresholds increased (Table 2).

There are several factors that likely affected similarity in river valley segments identified by VAST and expert-opinion. First, when delineating river valley segments by expert-opinion, the aquatic ecologists that placed river valley segment boundaries allowed segments to be comprised of both rivers and lakes (Baker, 2006; Seelbach et al., 2006). Because the VAST delineation used variables such as predicted 90% base flow yield and July mean water temperature lakes could not be clustered with streams and rivers as reliable estimates for such variables were not available for lakes (lakes were assigned values of $-9999$ for these variables). This limited the lengths of river valley segments that could be identified by

<table>
<thead>
<tr>
<th>Affinity threshold</th>
<th>Unadjusted agreement (%)</th>
<th>Chance-corrected agreement (%)</th>
<th>Commission error rate (%)</th>
<th>Omission error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>76.9</td>
<td>45.4</td>
<td>19.4</td>
<td>3.7</td>
</tr>
<tr>
<td>1.0</td>
<td>84.1</td>
<td>57.1</td>
<td>11.1</td>
<td>4.6</td>
</tr>
<tr>
<td>1.5</td>
<td>87.2</td>
<td>63.3</td>
<td>7.9</td>
<td>4.9</td>
</tr>
<tr>
<td>2.0</td>
<td>87.7</td>
<td>64.2</td>
<td>7.3</td>
<td>5.0</td>
</tr>
</tbody>
</table>

The commission and omission error rates are also shown.
VAST in addition to limiting agreement with the expert-opinion approach. Additionally, although the river valley segment delineations were based on similar types of landscape characteristics, the databases used to represent and summarize these characteristics differed between the two approaches. Despite the factors that limited similarity in river valley segment delineations, we believe comparing VAST with the expert-opinion approach was useful for determining how well VAST can mimic expert-opinion. As previously stated, an advantage of the expert-opinion approach is that vast amounts of existing knowledge regarding relationships between biological assemblages and environmental conditions can be incorporated in the delineation process. The agreement rates that we observed between VAST and the expert-opinion approach (unadjusted Rand index of agreement ranging from 77 to 88%) suggest that VAST, when supplied with the correct type of environmental attribute data, can provide fairly similar results to expert-opinion and in a fraction of the time needed to manually delineate the segments.

6. Conclusions

Loss of aquatic biodiversity in rivers and streams has been globally pervasive and has been caused by a number of factors, including intensive land use practices, construction of dams, habitat degradation, pollution, and nonnative species invasion (Benke, 1990; Allan and Flecker, 1993; Rinne et al., 2005; Rose, 2005; Reed and Czech, 2005). As a result, a number of programs have been enacted to identify and preserve remaining vestiges of aquatic biodiversity in running waters (Groves et al., 2002; Sowa et al., 2007). The most appropriate spatial unit upon which to conceptualize assessment, monitoring, and management of rivers and streams for the purpose of preservation or restoration has been a matter of question (Dovciak and Perry, 2002; Fausch et al., 2002). River reaches are not suitable management units because of their small sizes (Fausch et al., 2002). Larger systems (e.g., catchments, hydrologic units) also may not be appropriate as rivers within these systems can exhibit remarkable amounts of complexity in environmental attributes and thus may not respond similarly to management actions (Hawkins and Norris, 2000; Omernik, 2003). River valley segments, being intermediate in scale to river reaches and catchments, are appealing as management units for rivers and streams for several reasons (Seelbach et al., 2006). First, they are similar in scale to which rivers are believed to react to heterogeneity in the landscape. Second, river valley segments typically are large enough to contain the multiple habitats required by some stream fishes to complete their entire life cycles. Third, given our understanding of how river valley segments are formed, it is conceivable for these units to be cost-effectively identified from landscape-scale GIS databases, without the need for expensive field visitations.

Borrowing a term from landscape ecology, river valley segments can be regarded as medium scale habitat patches for river networks. A number of stream ecologists have advocated for landscape ecological principles to play a larger role in the formulation of stream and river management decisions (Pringle et al., 1988; Wiens, 2002). Identification of habitat patches (i.e., river valley segments) is a requisite first step in adopting such principles. Once habitat patches have been identified, it is then possible to address issues related to patch quality, patch boundaries, patch density, and patch juxtaposition (Pringle et al., 1988; Wiens, 2002). In order for landscape ecological principles to provide information critical to management of streams and rivers, we believe it is important for appropriately sized habitat patches (i.e., river valley segments) to be identified objectively.

We believe that spatially constrained clustering is a promising approach for identifying river valley segments, and that the use of such methods will yield significant advantages, namely in efficiency, repeatability, and objectivity, over either an expert-opinion or automated class-based approach. Our goal in developing VAST has been to provide an easy-to-use spatially constrained clustering program for the purpose of delineating river valley segments from GIS river network databases. We intentionally have tried to make it a stand-alone program, thus VAST creates its own adjacency tables, conducts its own data transformations, and calculates its own resemblance coefficients. Other software programs capable of performing spatially constrained clustering, such as Knorr-Held and Rasser’s (2000) nonparametric Bayesian clustering method (www.statistik.lmu.de/index_e.html) or Casgrain and Legendre’s (1999) K-means clustering method (www.bio.umontreal.ca/casgrain/en/labo/R/v4/telecharger.html), also could be used to objectively identify river valley segments, although with these other methods users must develop their own spatial adjacency tables or calculate their own similarity or dissimilarity estimates for the stream reaches. The need to specify an affinity threshold in VAST may be viewed by some as a disadvantage; however, this also may be beneficial when identifying river valley segments for multiple data sets as it ensures a consistent level of within river valley segment variability. It also may be advantageous for users to control the level of variability when identifying river valley segments as different stream restoration or preservation scenarios may necessitate fewer numbers of river valley segments. When using VAST, it may be helpful to try several different combinations of affinity thresholds, processing orders, and resemblance coefficients to determine sensitivity of results to changes in these program options. It also may be beneficial to use cluster ensemble procedures to combine cluster partitions developed using these different program options to help derive the most stable cluster structure (Strehl and Ghosh, 2002; Fred and Jain, 2005).

There are several areas of research pertaining to the delineation of river valley segments through VAST, as well as other spatially constrained clustering methods, that we believe would be useful to explore. First, research into what variables should be used to delineate river valley segments should be a high priority. Our selection of variables that we used to delineate river valley segments was based on a combination of multivariate analyses of fish-habitat relationships, as well as our prior experiences studying fish assemblages in Michigan and Wisconsin streams. It is not our contention that these
are the best set of variables to use for delineating river valley segments in Midwestern US streams. Further, it is very likely that variables that best delineate river valley segments in one region may not be the best choice for other regions. A more rigorous validation of VAST’s ability to identify river valley segments also needs to be conducted. An appropriate validation will need to extend beyond simply comparing the results of VAST with other methods of delineating river valley segments. Rather, an independent data set that accurately reflects real world river valley segment partitioning in river networks will be needed. Finally, we believe it would be useful to begin exploring landscape ecological principles as they apply to river valley segments (e.g., patch quality, patch juxtaposition, and patch boundaries), particularly with respect to how such information might be used for protection and preservation of stream and river habitat and biodiversity.

Acknowledgements

The authors thank S. Aichele, E. Bissell, A. Cooper, A. Holtrop, J. Lyons, J. McKenna, Jr., D. Passino-Reader, C. Rieng, and J. Stewart for assisting with the development of the stream network databases used in this study and for participating in discussions regarding development of VAST. This publication was partially developed under STAR Research Assistance Agreement No. R-83059601-0 awarded by the US Environmental Protection Agency. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication. This project was also supported by Federal Aid in Sport Fishery Restoration Program, Project F-80-R-6, through the Fisheries Division of the Michigan Department of Natural Resources. This is publication 2007-07 of the Quantitative Fisheries Center at Michigan State University.

References


