Multi-level association rules and directed graphs for spatial data analysis

B. Petelin a,b, I. Kononenko b, V. Malačič a, M. Kukar b

a Marine Biology Station Piran, National Institute of Biology, Fornače 41, 6330 Piran, Slovenia
b Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia

1. Introduction

Spatial data mining and geographic knowledge discovery have evolved rapidly over the last decade due to an enormous growth in geographical data. Those data were collected with location aware technologies, high-resolution remote sensing and surveys. In contrast to classical data mining these methods take into account the spatial–temporal properties, expressly geographic measurement frameworks, spatial dependency and heterogeneity, the complexity of objects/rules and diverse data types (Miller and Han, 2009). Although the field of spatial data mining is very large, we limit the scope of this paper to spatial–temporal association rules and multi-level directed graphs with different levels of space and time granularity. We propose a methodology that upgrades the methods of the Lagrangian analysis of surface sea-water parcels. This methodology includes data mining with efficient visualization techniques, namely, spatial–temporal association rules and multi-level directed graphs with different levels of space and time granularity. In the resulting multi-level directed graphs we can intertwine knowledge from various disciplines related to oceanography (in our application) and perform the mining of such graphs. We evaluate the proposed methodology on Lagrangian tracking of virtual particles in the velocity field of the numerical model called the Mediterranean Ocean Forecasting Model (MFS). We describe an efficient algorithm based on label propagation clustering, which finds cycles and paths in multi-level directed graphs and reveals how the number and size of the cycles depend on the seasons. In addition, we offer three interesting results of the visualization and mining of such graphs, that is, the 12 months periodicity of the exchange of water masses among sea areas, the separation of Mediterranean Sea circulation in summer and winter situations, obtained with the hierarchical clustering of multi-level directed graphs, and finally, with visualization with multi-level directed graphs we confirm the reversal of sea circulation in the Ionian Sea over the last decades. The aforementioned results received a very favorable evaluation from oceanographic experts.

© 2013 Elsevier Ltd. All rights reserved.

Keywords: Spatial data mining Lagrangian analysis Spatial–temporal association rules Multi-level directed graphs Oceanography

Abstract

We propose a methodology that upgrades the methods of the Lagrangian analysis of surface sea-water parcels. This methodology includes data mining with efficient visualization techniques, namely, spatial–temporal association rules and multi-level directed graphs with different levels of space and time granularity. In the resulting multi-level directed graphs we can intertwine knowledge from various disciplines related to oceanography (in our application) and perform the mining of such graphs. We evaluate the proposed methodology on Lagrangian tracking of virtual particles in the velocity field of the numerical model called the Mediterranean Ocean Forecasting Model (MFS). We describe an efficient algorithm based on label propagation clustering, which finds cycles and paths in multi-level directed graphs and reveals how the number and size of the cycles depend on the seasons. In addition, we offer three interesting results of the visualization and mining of such graphs, that is, the 12 months periodicity of the exchange of water masses among sea areas, the separation of Mediterranean Sea circulation in summer and winter situations, obtained with the hierarchical clustering of multi-level directed graphs, and finally, with visualization with multi-level directed graphs we confirm the reversal of sea circulation in the Ionian Sea over the last decades. The aforementioned results received a very favorable evaluation from oceanographic experts.
them represent various relationships between these areas, for example the probability of the transition of Lagrangian particles from one sea area to another in a given time interval. The “multi-levelness” of our graphs is therefore reflected in the spatial subdivision of the problem domain and the time interval in which we observe certain phenomena. In addition, we adapt the existing algorithms for graph mining and develop new ones for use on multi-level directed graphs.

The paper is organized as follows. Section 2 describes related work referring to the content of this article. Section 3 is divided into several parts: first we describe the formation of spatial-temporal association rules based on a large number of Lagrangian trajectories and provide a definition of a multi-level directed graph. Then we describe in detail the domain subdivision, the formation of spatial–temporal item sets, and the selection of an appropriate time interval. In the same section we evaluate the proposed methodology on results of the numerical model Mediterranean Ocean Forecasting System (MFS) (Tonani et al., 2008) and develop an algorithm for finding cycles and paths in resulting multi-level directed graphs. In Section 4 we provide some results of the aforementioned algorithm and, in addition, three interesting results of mining multi-level directed graphs. The first example indicates the periodicity of probability of the exchange of surface water masses among areas in the Mediterranean Sea. The second example illustrates the seasonality of the circulation in the Mediterranean Sea, obtained by using a hierarchical clustering of monthly multi-level directed graphs for the period 1999–2011. Finally, in the third example we show the reversal of anticyclonic circulation in the Ionian Sea to cyclonic in the period 1996–1997 and vice versa in 2006–2007 through a multilevel directed graph, which we constructed on the basis of measurements of drifters in same periods. Section 5 concludes this paper and gives some ideas for future work.

3. Material and methods

We propose a novel methodology for spatial-temporal data mining of Lagrangian trajectories, which consists of the following steps. First, we need to have available a sufficient number of trajectories. In doing so, we can generate a large number (usually several thousands) of trajectories in the domain of a numerical model. On the other hand, we can use all available observations of floats which cover the area of interest. After we have the trajectories, we derive the spatial-temporal association rules resulting from these trajectories. Then, we construct a multi-level directed graph based on these rules, and finally include additional attributes of various types, which represent the domain knowledge, in our case oceanographic knowledge. In the following subsections we provide guidance for the selection of the parameters required for the formation of multi-level spatial-temporal association rules and directed graphs. Eventually, the resulting graph is prepared for graph data mining using various methods and algorithms. In this study we develop an efficient algorithm which finds cycles in multi-level directed graphs. In addition we carry out a hierarchical clustering and Fourier analysis of graph weights in order to detect seasonal patterns in the time series of these graphs.

3.1. Spatial-temporal association rules

The basic problem that we discuss here can be formulated as follows: “If a particle in the current time period is located in area A of a given spatial domain, then the same particle in the next time period will be located in area B with a certain probability”. In order to tackle this problem, we first subdivide the given spatial domain into smaller areas and choose an appropriate time interval with the help of heuristics which we describe later. Then, we determine the probability of displacement of particles from one area to another using spatial association rules (Koperski and Han, 1995). The mea-
sure of confidence of these rules denotes that probability. Fig. 1 illustrates the proposed concept.

3.2. Multi-level directed graphs

We represent the spatial–temporal association rules, described in the previous subsection, in the form of a multi-level directed graph. In the following we define the multi-level directed graph, which is used for the visualization of our spatial–temporal association rules:

**Definition 1 (Multi-level directed graph).** Let \( N \) be a set of nodes which represent areas in the model domain and \( E \) a set of edges between these nodes. The edges have weights \( w \) which represent the confidence of spatial–temporal association rules, i.e. the probability of the transition of the particles between nodes (areas) in a given time interval through an edge. A multi-level directed graph for Lagrangian analysis \( G = (N,E,w) \) has the following properties:

1. \( G \) is a Markov graph, i.e. a graph where for all nodes the weights of outgoing edges including a reflexive edge sum up to 1.0. \( G \) is associated with a transition matrix of the Markov process where graph nodes represent the states and the weights of edges representing the probability of the transition of particles between nodes (sea areas). The current states depend only on states from the previous time interval.
2. The temporal attributes are associated with graph \( G \) itself and they represent its time span.
3. Spatial and non-spatial attributes are associated with nodes \( N \). The spatial attributes define volume and location, and the non-spatial define the domain (oceanographic) and other properties.
4. In addition to the mandatory attribute \( w \) (the probability of transition), the edges \( E \) have any number of other attributes of any types that represent the relationships between the probability of transition and domain specific (oceanographic) and other quantities.
5. \( G \) is a multi-level graph, which means that different spatial subdivisions of the domain and corresponding time intervals can be used.

An example of such a graph is shown in Figs. 5 and 6 (right), where the weights of the reflexive edges are denoted by the saturation of red color of sea areas and the thickness of the edges in the graph is proportional to the probability of transition between the areas. On such a graph we can apply various data mining methods and algorithms. In developing such algorithms, which we describe later, we found inspiration in label propagation clustering, which is frequently used to find groups or communities in social networks (Raghavan et al., 2007; Šubelj and Bajec, 2011). In our work, we also considered algorithms for finding frequent subgraphs (Lakshmi and Meyyappan, 2012; Washio and Motoda, 2003).

3.3. Domain subdivision

Before the construction of a multi-level directed graph it is necessary to determine a spatial subdivision of the domain, which is the size and shape of the areas that belong to individual vertices. These can be arranged in a uniform rectangular grid (Fig. 2(A)), a non-uniform rectangular grid (Fig. 2(B)) or an arbitrary grid, composed of polygons (Fig. 2(C)). The subdivision of domain in areas depends on the problem we wish to solve with multi-level directed graphs. If we want to have a very fine subdivision, it is best to generate a uniform rectangular grid (Fig. 2(A)) or a non-uniform rectangular grid (Fig. 2(B)), when we wish to have certain areas divided more finely and the others roughly. The arbitrary subdivision is used for a small number of large areas (seas), where it is possible to define the subdivision manually. The user should define the subdivision which is best tailored for his/her needs.

After defining the subdivision, it is necessary to assign to each node (area) a unique label. In our study, we use a uniform rectangular grid and label the nodes with sequential numbers from 0 to \( n - 1 \), where \( n \) is the number of nodes. We start in the bottom left corner and move from left to right column-wise from bottom to top (see Fig. 2(A)). In order to reduce the number of nodes, we take into account only those which are of interest i.e. the nodes which partly or wholly contain the marine (“wet”) areas. The nodes in the case of a non-uniform rectangular subdivision can be labeled in a similar fashion. For an arbitrary subdivision, where we usually deal with a small number of areas, we suggest labeling areas with their geographical names. The use in continuation of such labeled areas is the same as for each of the previously described approaches.

3.4. Generation of spatial–temporal association rules

In order to generate spatial–temporal association rules we use the APriori algorithm (Agrawal and Srikant, 1994) that was originally used to discover frequent itemsets in a finite set of transactions. First we transform the movements of the particles from one area to another in a given time interval \( \Delta t \) into transactions, which are then put in use by the algorithm A Priori. We define the transactions in the form of tuples \((\text{year}, \text{month}, \text{position}(t), \text{depth}(t), \text{position}(t+\Delta t))\) where the \( \text{position}(t) \) is of type \( (x,y) \) with spatial coordinates \( x \) and \( y \), year and month represent temporal information of the time \( t \) and \( \Delta t \) is the time interval (usually in days) in which the particles either remain in the same area or move to another. The optimal choice of the time interval \( \Delta t \) is described in Section 3.5. We extract the examples from the trajectory database, that is, we sample the pairs of points that belong to trajectories and are of length \( \Delta t \) apart. Thus, the number of examples \( N_t \) that we can obtain in this manner is:

\[
N_t = n_i(l_i - \Delta t)
\]

where \( n_i \) is the number of trajectories and \( l_i \) is the length of the individual trajectory. The length \( l_i \) and \( \Delta t \) are both expressed in time units (usually in days). Then we map the positions in the examples to the areas and thus obtain the tuples \((\text{year}, \text{month}, \text{area}(t), \text{area}(t+\Delta t))\) using the procedure which is known in computational geometry as “point location”, i.e. given a partition of the space into disjoint areas, we have to determine the area where a given point lies (de Berg et al., 2000). We can perform this task with a “brute force” approach by searching sequentially all areas to find the one to which the given point belongs. This is useful for a small number of subdivisions with high complexity, for example large sea areas bounded by complex coastlines. In the case, where the subdivision of the domain is formed by multiple polygons (Fig. 2(C)), the “point location” problem can be solved using slab decomposition (Dobkin and Lipton, 1976), monotone subdivisions (Edelsbrunner et al., 1986), triangulation refinement and trapezoidal decomposition. All those algorithms have computational time complexity \( O(\log n) \), where \( n \) denotes the number of polygons in a given domain. For non-uniform rectangular decomposition (Fig. 2(B)) a binary search (or bisection) on both spatial axes (time complexity \( O(\log n) \)) can be used. Finally, for a uniform rectangular decomposition this reduces to a mapping (see Fig. 2(A)):

\[
i = \frac{|x - x_0|}{\Delta x} \quad j = \frac{|y - y_0|}{\Delta y} \quad L_k = n_i i + j
\]
where $x$ and $y$ are spatial coordinates of the given point in horizontal and vertical directions, respectively, $x_0$ and $y_0$ are the coordinates of the bottom left corner of the domain, $i \in 0, \ldots, n_x - 1$ is the index of the area in the horizontal direction, $j \in 0, \ldots, n_y - 1$ the index in the vertical direction, $\Delta x$ and $\Delta y$ are dimensions of areas in horizontal and vertical directions, respectively and finally, $L_k$ is the label (sequence number) of the area, assuming the ordering of nodes, shown in Fig. 2(A), where $L_k k \in 0, \ldots, n - 1$, denotes the labels of areas. Thus, the computational time complexity becomes $O(1)$. The overall time complexity for obtaining $N$ samples is therefore $NO(f(n))$, where $f(n)$ is either $n$ for the rough arbitrary subdivision, $\log(n)$ for the non-uniform rectangular subdivision or 1 for the uniform rectangular subdivision.

After obtaining the transactions we generate the spatial–temporal association rules, selecting $10^{-4}$ as an appropriate value for minimum support. From numerous association rules we extract only those from which we construct the multi-level directed graphs that cover months, years and periods of several years. The extracted spatial–temporal association rules are therefore in one of the following forms:

$$ year(t) \land month(t) \land area(t) \Rightarrow area(t + \Delta t) \quad (s(\%), c(\%)) $$

$$ year(t) \land area(t) \Rightarrow area(t + \Delta t) \quad (s(\%), c(\%)) $$

$$ area(t) \Rightarrow area(t + \Delta t) \quad (s(\%), c(\%)) $$

where $s(\%)$ and $c(\%)$ are the support and confidence of the association rule, respectively. Assuming that certain particles are located in area $A$ at time $t$ and that some of these particles move from area $A$ to area $B$ in a given interval $\Delta t$, then the support and confidence is given with:

$$ s(\%) = N_A/N \times 100\% \quad c(\%) = N_B/N_A \times 100\% $$

where $N_A$ is the number of particles which reside in area $A$ at time $t$, while $N_N$ is the number of particles which moved from area $A$ to area $B$ in time interval $\Delta t$, and $N$ is the total number of particles in the model domain at time $t$.

### 3.5. Time interval selection

The next task that arises is choosing the appropriate time interval $\Delta t$ for which we determine the probability of the transitions of particles between sea areas. The time interval is dependent on the dimensions of areas and the velocity of particles in those areas. For a multi-level directed graph, it is desirable to have as many edges to the spatially neighboring nodes as possible and as few as possible to remote nodes. Ideally, the resulting multi-level directed graph would be similar as much as possible to the “graph of neighbors” in which we define the edges as follows:

$$ e_{ij} = \begin{cases} 
1 & i \neq j \text{ or touches } (v_i, v_j) \\
0 & \text{otherwise} 
\end{cases} $$

where the predicate “$\text{touches}(v_i, v_j)$” means that the nodes $v_i$ and $v_j$ share either a side or a corner of the corresponding polygon. The rule of thumb for determining the suitable time interval would be $\Delta t = d_{\text{avg}}/v_{\text{avg}}(3600 \times 24)$, where $d_{\text{avg}}$ is the average distance between pairs of neighboring nodes in meters and $v_{\text{avg}}$ is the average velocity (m/s) of particles or sea currents in the domain, and $\Delta t$ is given in days. Another way would be the exhaustive generation of all possible multi-level directed graphs with the same spatial subdivision with different $\Delta t$, which would be intractable. Here we use a different approach. First, we select a suitable interval $(\Delta t_{\text{min}}, \Delta t_{\text{max}})$ and for each $\Delta t$ from that interval we randomly select a smaller subset of examples and calculate the number of particles which: (1) remain in the same area, (2) move to the neighboring nodes and (3) move beyond the neighboring nodes. By increasing $\Delta t$, the proportion of particles that remain in the same area decreases, while the number of particles that pass into the neighboring areas and beyond increases. The number of particles that pass into adjacent areas and stay there increases up to a certain maximum and then begins to fall. The position of this maximum is the optimal choice for $\Delta t$, which in our case equals six days (see Fig. 3).

### 3.6. Evaluation

We studied the velocity fields of ocean currents from the numerical model of the Mediterranean Ocean Forecasting System (MFS)\(^1\) (Tonani et al., 2008) for the period 1999–2011 and calculated the Lagrangian trajectories using the software tool Ariane.\(^2\) In simulation we released virtual (numerical) particles from 239 equally horizontally distributed starting positions (every 16th cell in $x$ and $y$ direction of the numerical model) at a depth of 1 m, and at the beginning of each month up to and including January 1st, 2011. Thus, over 145 months, we obtained a total of 34,655 trajectories with a length of 365 days each. We set vertical velocity in the model to 0, so the virtual particles move continuously at the same depth (1 m), which is similar to oceanographic surface floats (drifters), which travel consistently at the same depth. Figs. 4 and 6 (left) show the resulting set of trajectories, from which we cannot determine the essential circulation in the Mediterranean Sea and, therefore, it is necessary to use methods of data mining with multi-level directed graphs.

We subdivided the domain in Fig. 4 in $60 \times 30 = 1800$ equally sized areas and among them 848 areas are fully or partially covered by the sea. In this case, the optimal “graph of neighbors” (see Eq. 7) contains 848 nodes and 7006 edges. Using the method described in Section 3.5, we selected the optimal time interval $\Delta t$ of six days, in which 40.3% of particles (maximum) move into the neighboring areas, 50% remain in same areas and 9.7% go beyond the neighborhood (Fig. 3). Thus we obtained 12,415,413 examples and generated 156 monthly graphs (199901, 199902, ..., 201112), 13 yearly graphs (1999, 2000, ..., 2011) and one multi-level

\(^1\) <http://gnoo.bo.ingv.it/mfs/ >.

\(^2\) <http://stockage.univ-brest.fr/grima/Ariane>.
directed graph, which covers the whole period 1999–2011, using spatial–temporal association rules of the form denoted by Eqs. (3)–(5), respectively. Table 1 summarizes the number of edges of the obtained graphs. From the table it is evident that by increasing the time span covered by the graph, the number of edges also increases, that is, a yearly graph has more than twice as many edges compared to monthly graphs. This can be attributed to different edges with low support and confidence, pointing to non-neighboring (remote) areas. The multi-level directed graph for the whole period 1999–2011 (Figs. 5 and 6) has 18,770 edges.

Table 1
Descriptive characteristics of edges of multi-level directed graphs with different time granularity (monthly, yearly and overall).

<table>
<thead>
<tr>
<th>Time granularity</th>
<th>Number of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± SD</td>
</tr>
<tr>
<td>Monthly</td>
<td>4224 ± 766</td>
</tr>
<tr>
<td>Yearly</td>
<td>11193 ± 758</td>
</tr>
<tr>
<td>Overall</td>
<td>18770 ± 0</td>
</tr>
</tbody>
</table>

Fig. 3. The percentage of the number of particles: (1) remaining in the same areas (red), (2) moving to neighboring areas (green), and (3) moving to remote areas (blue), depending on the time interval. The quantity $d$ denotes the number of steps over graph edges which the particles perform. The optimal choice for a given spatial subdivision $60 \times 30$ (see subSection 3.6) for $\Delta t$ is six days, which is the maximum of percentage of particles moving to neighboring areas (green curve). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Trajectories of surface virtual particles released in the velocity field of the numerical model Mediterranean Ocean Forecasting System (MFS) in the period 1999–2011. Black dots represent the initial positions of particles. Trajectories are of different colors for ease of presentation.
3.7. Mining of multi-level directed graphs

As a basis for mining the multi-level directed graphs we consider label propagation clustering, which is frequently used for mining social networks (Raghavan et al., 2007; Šubelj and Bajec, 2011) and frequent subgraphs mining with algorithms using the A Priori principle (Lakshmi and Meyyappan, 2012; Washio and Motoda, 2003). We have adopted these principles and developed a label propagation algorithm for clustering nodes of multi-level directed graphs. The algorithm examines all nodes and assigns to each node a label of the neighboring node, from which the incoming edge has the maximum weight (the probability of transition). The resulting algorithm has a time complexity of $O(m + n)$ for our specific type of graph, where $n$ is the number of nodes and $m$ the
number of edges in the graph. From that algorithm we deduce another algorithm for finding frequent paths and cycles in multi-level directed graphs, which is presented in Algorithm 1 and described in detail in the following paragraphs. After obtaining paths and cycles, we find the frequent cycles using the Apriori algorithm.

Initially all vertices are candidates to be part of either cycles or paths and none of the vertices has been visited yet. We add a sequence number of cycle (cycle.number) and type of cycle (cycle.type) as edge attributes which are undefined at the beginning. The type of cycle can be cyclonic (rotating in the counter-clockwise sense), anticyclonic (clockwise) or “undefined”. The latter is reserved for a cycle that contains only two nodes, and the algorithm cannot determine the sense of rotation. The rotation sense is determined by calculating the pseudo-area of the polygon that the cycle surrounds. If it is positive, then the cycle is cyclonic, otherwise it is anticyclonic. The algorithm visits each of the nodes, marks it as visited and pushes it onto a stack. Then it chooses its successor which leads the edge with maximum weight (the probability of transition). If the successor exists and has not yet been visited, it is marked as visited and pushed onto the stack. The algorithm repeats this procedure until it encounters a node that has already been visited, which means that it has found a cycle. It then pops the nodes (predecessors) from the stack and marks them with the number of the cycle that was found. In doing so, the algorithm walks on the cycle in the opposite direction, until it encounters again the node cycle.begin and completes the cycle. It calculates the pseudo-area of a polygon which is enveloped with that cycle by the formula (Braden, 1986):

\[
A = \frac{1}{2} \left[ (x_1 - x_2)(y_1 + y_2) + (x_2 - x_3)(y_2 + y_3) + \cdots + (x_n - x_1)(y_n + y_1) \right]
\]

where points \(P_1(x_1,y_1), P_2(x_2,y_2), \ldots, P_n(x_n,y_n)\) denote the mass centers of nodes (areas) contained in the cycle, and \(x_i\) and \(y_i\) are their spatial coordinates. The point \(P_1(x_1,y_1)\) is the beginning and ending point of the cycle (cycle.begin). If the resulting pseudo-area is positive, then the cycle is cyclonic, if it is negative, the cycle is anticyclonic. If the pseudo-area equals to zero, the cycle is “undefined”. This happens when forming a cycle of two spatially close nodes between which there are mutually two edges, which have the highest probability of transition. The edges of the remaining nodes in the stack form a path and not a cycle.

For a given candidate node, if the algorithm cannot find a cycle in this way, then one of the following events occurred: either it has encountered a node that has no outgoing edges, or has encountered a node that has already been processed.

The algorithm visits each node only once and finds the edges with maximum weights pointed to successors. The average time complexity of the algorithm is \(O(nd) = O(n)\), where \(n\) is the number of nodes, \(d\) the average number of outbound edges of nodes and \(m\) the number of edges. In addition, all \(n\) nodes are stored exactly once on the stack and taken from it and in the same time the attributes cycle.number and cycle.type are assigned to the edges. Thus the total expected time complexity is \(O(m + n)\). The previously mentioned algorithms for label propagation clustering (Raghavan et al., 2007; Šubelj and Bajec, 2011) are more general and much more demanding on multi-level directed graphs. They perform this task in \(t\) iterations and have the time complexity \(O(nt)\), where \(n\) is the number of nodes in the graph.

3 This is true for the northern hemisphere, in the southern hemisphere a cyclonic gyre is in a clockwise direction while anticyclonic is counter-clockwise.

4 This is not a proper area, because it is calculated directly from the geographical coordinates instead of length units.

Algorithm 1. Finds paths and cycles in multi-level directed graphs

Require: Graph \(G(N,E)\) with weights \(w\)
Ensure: Graph \(G_{cycle}(N,E_{out})\) with weights \(w_{out}\) (cycles and paths)
1: for all \(n \in N\) do
2: \(visited_n \leftarrow false\)
3: \(candidate_n \leftarrow true\)
4: end for
5: for all \(e \in E\) do
6: \(cycle.number_{e} \leftarrow NA\)
7: \(cycle.type_{e} \leftarrow NA\) (“cyclonic”, “anticyclonic”, “undefined”, “path”)
8: end for
9: \(s \leftarrow stack()\)
10: cycle.seq \leftarrow 1
11: for all \(n \in N\) do
12: if \(candidate_n\) then
13: \(visited_n \leftarrow true\)
14: \(found.visited \leftarrow false\)
15: \(found.blind \leftarrow false\)
16: \(begin.cycle \leftarrow NA\)
17: \(push(s, n)\)
18: \(curr \leftarrow n\) (\(n\) is the current node)
19: while not (found.visited or found.blind) do
20: \(succ \leftarrow argmax(w_{out}), j \neq curr\) (successor with max. \(w\))
21: if \(exists_{succ} \) then
22: if \(candidate_{succ} \) then
23: \(visited_{ succ} \leftarrow true\)
24: \(found.visited \leftarrow true\)
25: \(begin.cycle \leftarrow succ\)
26: else
27: \(visited_{succ} \leftarrow true\)
28: \(push(s, succ)\)
29: end if
30: \(curr \leftarrow succ\)
31: else
32: \(found.blind \leftarrow true\) (successor already used)
33: \(push(s, succ)\)
34: end if
35: else
36: \(found.blind \leftarrow true\) (successor not found)
37: end if
38: end while
39: \(found.begin \leftarrow false\)
40: \(area \leftarrow 0\)
41: \(prepred \leftarrow begin.cycle\) (pre-predecessor is the beginning of cycle)
42: while not empty(s) do
43: \(pred \leftarrow pop(s)\)
44: if not (found.blind or found.begin) then
45: cycle.number_{prepred} \leftarrow prepred \leftarrow cycle.seq
46: segment \leftarrow calculate area of segment
47: \(area \leftarrow addSegment(area, segment)\) (adds segment to area)
48: if \(pred \leftarrow begin.cycle\) then
49: \(found.begin \leftarrow true\)
50: if area > 0 then
51: \(cycle.type \leftarrow “cyclonic”\)
52: else if area > 0 then
53: \(cycle.type \leftarrow “anticyclonic”\)
54: else

(continued on next page)
4. Results

4.1. Cycles in multi-level directed graphs

In one of the recent and most complete works about the circulation of the Mediterranean Sea, which is grounded on results of the numerical model the Mediterranean Ocean Forecasting System (MFS) (Tonani et al., 2008), there are many phenomena and circulation features (e.g. interannual variability of circulation, transport through straits) which are revealed. It appears, however, that the capacities of the applied numerical model to reveal closed circulation features have not been explored to date and this is the motivation for the development of an algorithm for finding circulation cycles in a “space” sense and paths in the resulting multi-level directed graphs.

We applied the algorithm for finding cycles on 156 monthly graphs obtained from the numerical model of the Mediterranean Ocean Forecasting System (MFS) for the period 1999–2011 (see Section 3.7), and we found a total of 1361 cyclonic, 774 anticyclonic and 5197 “undefined” cycles. Figs. 7 and 8(C) give an example of the cycles found in the graph for September 2000, which was selected because it contains some interesting larger cyclonic cycles.

Fig. 8(A) and (B) show the coverage of cyclonic and anticyclonic cycles. Both were obtained by stacking either all cyclonic or anticyclonic cycles. The edges that appear more frequently are in more saturated color. This gives the user a first impression about how cycles are spread and where the most frequent are.

We also analyzed the time evolution of the number and size of cycles in the period 1999–2011 on a monthly scale. The time series of both number and size of cycles exhibit interesting seasonal variations. From Fig. 9 we can see that the number of cyclonic cycles in the winter is almost twice as high as in the summer. The minimum number appears in July and the maximum in January. The number of anticyclic cycles grows from December to the warmer seasons and reaches the maximum in September. The total number of cycles is prominent in winter months, then starts to decay and reaches a global minimum in July. After that it begins to grow rapidly and achieves a significant peak in September and then again decays and reaches a local minimum in November.

Figs. 8(D) and 10 show frequent cycles which were found by the APriori algorithm (see the last paragraph in Section 3.7), with a minimum support 0.0005. In other words, each of the resulting frequent cycles must occur at least twice. The total number of cycles, excluding the “undefined” ones, yields 2135 (1361 cyclonic and 774 anticyclic). Thus the requested minimum number of occurrences must be 2 (2 \* (2135 \* 0.0005 = 1.07)). The number of occurrences of the resulting frequent cycles in the period 1999–2011 ranges from 2 to 13 on a monthly scale. Therefore identical cycles recur once a year at most, or even less frequently.

Many cycles which were obtained using the described algorithm are addressed in several studies, for example (Hamad et al., 2006; Millot, 1999; Robinson et al., 2001). For the Adriatic Sea the method clearly reveals the most frequent cyclonic (red) gyre in the southern Adriatic, less frequent in the central Adriatic, where, near the coastlines, anticyclonic gyres are also present. Between the central and northern Adriatic the cyclonic gyre is present, while in the northern Adriatic anticyclonic circulation is most frequent in the northern-most corner, with cyclonic gyres south of it. Other examples of cycles that were obtained by the algorithm conform with known cycles: the wind induced cyclonic Northern Tyrrhenian Gyre, the cyclonic Cretan Gyre, the cyclonic Rhodes Gyre, the anticyclonic Mersa-Matrhu Eddy at the south of Cyprus, etc.

4.2. Transition of surface waters between the areas of the Mediterranean

Although there were many previous analyses of circulation through and between domains (regions) in the Mediterranean Sea (Hamad et al., 2006; Millot, 1999; Robinson et al., 2001) there is a lack of a simple view showing the exchange of a surface water mass among areas in the Mediterranean within a month – this seems to be an important issue for the spread of pollutants (Coppini et al., 2011) and for the wandering of wrecks. This is the motivation for the application of multi-level directed graphs.

Although the seasonality of Mediterranean circulation dates back into early systematic experimental explorations of the Mediterranean Sea (Ovchinnikov, 1966) and was also explored later (Millot, 1991; Roussenov et al., 1995; Tziperman and Malanotte-Rizzoli, 1991), we have found that there is a lack in the analysis
of the surface circulation from the simple point of view of a surface water exchange among areas on a monthly scale. It is useful to know the percentage of the seasonal amplitude of probability of the exchange of water mass among domains versus the 'steady' (annual) component, where the movements of artificial Lagrangian parcels on a monthly scale are applied. To explore this seasonality is a motivation for the application of multi-level directed graphs and the Fourier analysis method on the results of the numerical model MFS.

This example shows the periodicity of the transition of surface waters between sea areas of the Mediterranean Sea in the period 1999–2011. For this purpose we subdivided the domain of the Mediterranean Forecasting System numerical model into areas which roughly correspond to geographical seas (Adriatic Sea, Ionian Sea, etc). For the time interval \( \Delta t \) we have chosen 30 days in order to compare the results with other monthly oceanographic analyses. Therefore, in this case we omitted the procedure for finding the optimal \( \Delta t \), described in Section 3.5.

From the spatial–temporal association rules we obtained 156 (13 years by 12 months) directed graphs for each pair (year, month) in the period 1999–2011 with the nodes and weighted edges that represent sea areas and the probabilities of transitions between these areas, respectively. By examining sequentially in time 156 monthly directed graphs from the period 1999–2011, we visually observed the oscillation of their weights (transition probabilities) and thus we carried out a Fourier analysis of weights of the whole time series of graphs. We found the first period of 12 months, which suggests that the Lagrangian 'monthly mean' velocity field in the model has a seasonal characteristic. Fig. 11 shows the resulting directed graph. We observe that the maximum annual fluctuation (4.3%) is reached between the Ligurian and Balearic Seas. The maximum probabilities that water parcels remain in the same areas (95.4–95.6%) are shown for the Adriatic and Levantine basins. From the graph in Fig. 11 we observe that the annual fluctuations of the probabilities of transition between sea areas for the Western Mediterranean (1.5% on average) are small compared to their “mean” values (12.9% on average). For the Eastern Mediterranean the “mean” values are smaller (4.2% on average) and the values of fluctuations are still present (0.7% in average). These two facts confirm the steady surface circulation in the West Mediterranean and a stronger seasonal signal in the surface circulation with respect to the steady signal of the Ionian and eastern Levantine basins of the Eastern Mediterranean (see Tziperman and Malanotte-Rizzoli, 1991, p. 432).

4.3. Hierarchical clustering of multi-level directed graphs

Several studies were conducted on seasonality which is a dominant signal in the Mediterranean Sea (Millot, 1991; Ovchinnikov, 1966; Rousenov et al., 1995; Tziperman and Malanotte-Rizzoli, 1991). Our goal is to confirm this seasonality through the hierarchical clustering of 156 monthly multi-level directed graphs obtained from the results of the numerical model MFS over the period 1999–2011.

In this case, we applied agglomerative hierarchical clustering on transition matrices, which contain the weights (probability of the transition of particles) of 156 monthly multi-level directed graphs from 848 areas with a time interval of six days, obtained from the numerical model of MFS in the period 1999–2011 (see Section 3.6 and Fig. 11). For the distance measure between clusters we used the Ward criterion, which, unlike other measures (single, complete and average linkage clustering), minimizes the total within-cluster variance. The other aforementioned measures use the Euclidean distance between clusters. Hierarchical clustering, which uses the Ward criterion, divides the 156 graphs into two classes which are comparable in size. The other aforementioned methods gave two classes, one of which contained almost all graphs, while the other only a few while we obtained 98 graphs in cluster1 and 58 graphs in cluster2.

Fig. 12 shows the monthly distribution of the number of graphs which are classified into one of the two clusters. From the figure we can see that cluster1 represents the graphs, which mainly belong to cooler months, and cluster2 is mainly com-
posed of graphs from warmer periods. Therefore, hierarchical clustering into two main clusters separates the general circulation in the Mediterranean Sea into winter and summer situations, confirming the seasonality of the circulation in the Mediterranean Sea. The same experiment on a rougher graph with 235 nodes showed a similar separation, which confirms the robustness of this method.

4.4. The reversal of the circulation in the Ionian Sea

Borzelli et al. (2009) and Gačić et al. (2010) performed in depth analyses of experimental observations, satellite data, trajectories of floats and temperature and salinity vertical profiles. These analyses lead to the conclusion that the circulation of the Ionian Sea experiences a shift from an anticyclonic (clockwise) to a cyclonic (anti-
A hypothesis for the mechanism is given by which one developed type of circulation starts providing the conditions for the development of another type of circulation and that this happens with a period of ten years. The authors suggest that a reversal from an anticyclonic to a cyclonic regime of the circulation took place in 1997 and vice versa in 2006–2007.

This hypothesis was also explored by applying the multi-level directed graphs method in the analysis of measurements of drifters (Poulain et al., 2012). To show a possible circulation reversal in the Ionian Sea, we used the data of available observations with surface floats – drifters (Poulain et al., 2012) at depths from 0 to 15 m in the period 1994–2007. In accordance with the findings in the previously mentioned oceanographic literature, we...
divided the period 1994–2007 into three sub-periods: from 1994 to June 1997, from July 1997 to 2006, and the entire year 2007, which belong to anticyclonic, cyclonic and anticyclonic regimes in the Ionian Sea, respectively. For each period we constructed a multi-level directed graph with 246 nodes (uniform rectangular subdivision) with dimensions of sea areas, which were about twice as coarse as those described in Section 3.6. Fig. 13 shows the resulting multi-level directed graphs for each sub-period. In order to make the circulation more visible, we show in the figure only the edges between neighboring areas $i$ and $j$, which have a minimum confidence 10% and minimum support 0.01% for the first and the third sub-periods and 0.1% for the second sub-period.

In Fig. 13(A) we can identify a distinct anticyclonic regime before mid-1997, a little less pronounced cyclonic regime in the period from mid-1997 to 2006 (Fig. 13B) and a clear anticyclonic
regime in 2007 (Fig. 13(C)), which confirms studies about circulation reversal in the Ionian Sea over the last two decades (Borzelli et al., 2009; Gačić et al., 2010).

5. Conclusions

We presented a new approach that adds value to existing methods of Lagrangian analysis (Griffa et al., 2007) and helps the user to discover and visualize hidden patterns and regularities in the results of measurements and numerical models. For this purpose, we used spatial–temporal association rules and multi-level directed graphs, which have proven to be a promising choice for this kind of task. We developed methods and algorithms for the analysis and mining of multi-level directed graphs. These algorithms are effective in detecting seasonal patterns in the transitions of water masses between sea areas. This approach enabled us to find that the probabilities of the transition of water masses between different areas of the Mediterranean Sea have a period of 12 months, and using hierarchical clustering we separated the sea circulation into winter and summer situations. In addition, the proposed methodology is also suitable for the visualization of long term transient phenomena, such as circulation reversal in the Ionian Sea, described in the oceanographic literature (Borzelli et al., 2009; Gačić et al., 2010; Poulain et al., 2012). The multi-level directed graph method applied on Lagrangian trajectories of surface floats confirmed the circulation reversal in the Ionian Sea in 1997 and 2006–2007.

We developed an algorithm for finding cycles and paths in multi-level directed graphs, which allows the detection of specific cycles which rarely occur. Using this algorithm, we found that the number and size of cycles have strong seasonal dependence. The algorithm has a linear time complexity depending on the number of edges in the graph. For this type of graph the time complexity is comparable to or even more efficient than the existing algorithms for social network analysis (Raghavan et al., 2007; Subelj and Bajec, 2011). Using the A PRIORI algorithm we extracted from the resulting cycles those occurring most frequently.

Although the described methodology and algorithms look promising, there are still many unanswered questions which are left for future work. First, what should be the spatial and temporal granulation of multi-level directed graphs that will most effectively visualize the results of mining and have maximum interest for the user? Second, to convert a multitude of individual cycles and paths, which differ from each other in a small number of nodes, in a more general and persistent picture of circulation in the domain, it would be necessary to use a “fuzzy” approach and construct frequent and persistent “soft” cycles and paths.

The results, together with the search for cycles and paths, clearly demonstrate the usefulness of the proposed methodology. We can conclude that the proposed methodology and mining algorithms developed so far gave results that are very promising in extracting useful information from vast modeled data of numerical simulations or measurements of ocean circulation.

Acknowledgments

This research is supported by the Slovenian Research Agency (Contract No. P1–0237), EC project PERSEUS (Contract No. 287600) and EC project MyOcean2 under the 7th Framework Programme. It is a part of the Doctoral Program in Computer and Information Science at the University of Ljubljana, Faculty of Computer and Information Science. Special thanks go to Pierre-Marie Poulain and Milena Menna from Istituto Nazionale di Oceanografia e di Geofisica Sperimentale – OGS, Trieste, Italy, who kindly provided the measurement data of drifters from the Mediterranean Surface Velocity Programme (MedSVP).

References

