Mining Spatio-Temporal Patterns of Periodic Changes in Climate Data

Corrado Loglisci, Michelangelo Ceci, Angelo Impedovo, Donato Malerba

Department of Computer Science, Universita' degli Studi di Bari "Aldo Moro", Italy
CINI- Consorzio Interuniversitario Nazionale per l'Informatica, Italy
{corrado.loglisci,michelangelo.ceci,donato.malerba}@uniba.it,
angelo.impedovo.89@gmail.com

Abstract. The climate changes have attracted always interest because they may have great impact on the life on Earth and living beings. Computational solutions may be useful both for the prediction of the climate changes and for their characterization, perhaps in association with other phenomena. Due to the cyclic and seasonal nature of many climate processes, studying their repeatability may be relevant and, in many cases, determinant. In this paper, we investigate the task of determining changes of the weather conditions which are periodically repeated over time and space. We introduce the spatio-temporal patterns of periodic changes and propose a computational solution to discover them. These patterns allows us to represent spatial regions with same periodic changes. The method works on a grid-based data representation and relies on a time-windows analysis model to detect periodic changes in the grid cells. Then, the cells with same changes are selected to form a spatial region of interest. The usefulness of the method is demonstrated on a real-world dataset collecting weather conditions.

1 Introduction

Climatology is a discipline essentially focused on the observation and study of weather conditions and it is one of the scientific fields characterized by a large variety of data-intensive and dynamic processes. Studying the evolution of the weather becomes thus determinant because might support the understanding of other processes, such as the industrialization and atmospheric changes. In this sense, a valid contribution is represented from the application of data-driven techniques [3], which open to the possibility to analyze climate observations in order to unearth empirical knowledge without demanding a-priori hypothesis, as the standard statistics method do instead. The proliferation of the technologies able to record and store multi-source and massive meteorological data has definitely confirmed the usefulness of the data analysis algorithms for several problems in Climatology.

One of the most scientifically and technologically challenging problems is that of finding changes and events of the weather conditions in order to build and refine models used for predictive tasks. Although in data mining we can find a long list of works on event and change detection [1], the identification of changes in climate data is challenging for several reasons. First, climate data tend to be noisy, thus we could have difficulty in distinguishing with an high degree of certainty the difference between significant changes and spurious outliers. Second, changes that persist over time and that cover relatively long intervals of time (e.g., days) can be originated from instantaneous deviations (e.g., rainfall extreme events which span few hours), which we could erroneously assess as meaningless. Third, the global models provide reliable indications for world-wide climate phenomena, while they could be no longer appropriate capture features of the regional weather conditions, where instead local models could be more effective [14].

In Climatology, many phenomena are cyclic in nature and can exhibit repetitive behaviors. Likewise, changes in weather conditions can be periodic because they are repeated at regular intervals of time. For instance, seasonal changes reflect the occurrence of the expected variations of the weather conditions and can recur up to one year of distance. The periodicity becomes thus a good indicator of the repeatability and meaningfulness of the changes since the variations which regularly recur may be considered more interesting than those episodic.

This paper focuses on the analysis of time-series describing the weather conditions recorded in geographically distributed locations and introduces the problem to discover spatio-temporal patterns able to relate periodic changes of the weather conditions with the spatial regions in which the changes occur. The geographic information of the weather conditions is used to determine the spatial component of the patterns, while the periodicity associated with the changes denotes the temporal component of the patterns. In this work, we propose a data mining framework which analyzes weather conditions data partitioned over a gridded data space and proceeds in two distinct steps in which the temporal dimension and spatial component are handled together. It first detects periodic changes at the level of individual cells of the grid and then it finds sequential patterns of the periodic changes over the cells in which the changes are present. The use of a technique of data partitioning is to not under-estimate the periodicity of local changes, which instead we could have whether we work on (global) statistical regularities.

More precisely, in the first step, we combine a model of analysis based on time-windows with a frequent pattern mining method to search for periodic changes in each grid cell. Changes are detected as significant variations of the frequency of the patterns mined from two different time-windows of data. The rationale in using the frequency is that it denotes regularity, therefore frequent patterns can provide empirical evidence about changes really happened. Building time-windows allows us to summarize the changes occurring at the level of time instants and model them at a higher level of temporal granularity, that is, intervals of time. Not all the changes are considered, but only those which are repeated over time-windows in several grid cells. The second step operates on the detected periodic changes and uses a sequential pattern mining method in order to find changes common to several cells. Mining sequential patterns allows us to find changes at a higher level of spatial granularity built with aggregations of grid cells.

The paper is organized as follows. In Section 2 we report necessary notions, while the method is described in Section 3. An application to the real-world dataset is described in Section 4. Then, we discuss the related literature (Section 5). Finally, conclusions close the paper (Section 6).

2 Basics and Definitions

Before formally describing the proposed method, we report some basic notions and definitions necessary for the paper.

Let $\{t_1 \ldots t_n\}$ be a a sequence of discrete time-points. For each time-point t_i, we have the values $A_i \in \Re^d$ of the weather parameters measured in geographically distributed areal units. A *time-window* τ is a sequence of consecutive time-points $\{t_i, \ldots, t_j\}$ $(t_1 \leq t_i,\ t_j \leq t_n)$ which we denote as $[t_i; t_j]$. The width w of a time-window is the number of time-points in τ, i.e. $w = j - i + 1$. We assume that all the time-windows have the same width w. Two time-windows τ and τ' defined as $\tau = [t_i, ; t_{i+w-1}]$ and $\tau' = [t_{i+w}; t_{i+2w-1}]$ are *consecutive.*

Let $\tau = [t_i; t_{i+w-1}]$, $\tau' = [t_{i+w}; t_{i+2w-1}]$, $\tau'' = [t_j; t_{j+w-1}]$, and $\tau''' = [t_{j+w}; t_{j+2w-1}])$ be time-windows, two pairs of consecutive time-windows (τ,τ') and (τ'',τ''') are *δ-separated* if $(j + w) - (i + w) \leq \delta$ ($\delta >0$, $\delta \geq w$). Two pairs of consecutive time-windows (τ,τ') and (τ'',τ''') are *chronologically ordered* if j>i. In the remaining of the paper, we use the notation τ_{h_k} to refer to a time-window and the notation (τ_{h_1}, τ_{h_2}) to indicate a pair of consecutive time-windows.

The following notions are crucial for this work.

A pattern P is a set of pairs, each pairs is composed by a weather parameter and its value. It can have at most d pairs. We say that P occurs at a time-point t_i if all pairs of P occur at the same time-point t_i. A pattern P is characterized by a statistical parameter, namely the *support* (denoted as $sup_{\tau_{h_k}}(P)$), which denotes the relative frequency of P in the time-window τ_{h_k}. It is computed as the number of the time-points of τ_{h_k} in which P occurs divided by the total number of time-points of τ_{h_k}. When the support exceeds a minimum user-defined threshold $minSUP$, P is *frequent* (FP) in the time-window τ_{h_k}.

Definition 1. Emerging Pattern (EP)

Let (τ_{h_1},τ_{h_2}) be a pair of consecutive time-windows; P be a frequent pattern in the time-windows τ_{h_1} and τ_{h_2}; $sup_{\tau_{h_1}}(P)$ and $sup_{\tau_{h_2}}(P)$ be the support of the pattern P in τ_{h_1} and τ_{h_2} respectively, P is an emerging pattern in (τ_{h_1}, τ_{h_2}) iff $\frac{sup_{\tau_{h_1}}(P)}{sup_{\tau_{h_2}}(P)} \geq minGR \quad \vee \quad \frac{sup_{\tau_{h_2}}(P)}{sup_{\tau_{h_1}}(P)} \geq minGR$

where, $minGR$ (> 1) is a user-defined minimum threshold.
The ratio $sup_{\tau_{h_1}}(P)/sup_{\tau_{h_2}}(P)$ ($sup_{\tau_{h_2}}(P)/sup_{\tau_{h_1}}(P)$) is denoted with $GR_{\tau_{h_1},\tau_{h_2}}(P)$ ($GR_{\tau_{h_2},\tau_{h_1}}(P)$) and it is called *growth-rate* of P from τ_{h_1} to τ_{h_2} (from τ_{h_2} to τ_{h_1}). When $GR_{\tau_{h_1},\tau_{h_2}}(P)$ exceeds $minGR$, the support of P decreases from

τ_{h_1} to τ_{h_2} by a factor equal to the ratio $sup_{\tau_{h_1}}(P)/sup_{\tau_{h_2}}(P)$, while when $GR_{\tau_{h_2},\tau_{h_1}}(P)$ exceeds $minGR$, the support of P increases by a factor equal to $sup_{\tau_{h_2}}(P)/sup_{\tau_{h_1}}(P)$.

The concept of emerging pattern is not novel in the literature [2]. In its classical formulation, it refers to the values of support of the same pattern which has been discovered in two different classes of data, while, in this work, we extend it to represent the differences between the data collected in two intervals of time, and therefore, we refer to the values of support of the a pattern P which has been discovered in two time-windows.

Definition 2. Periodic Change (PC)

Let $T : \langle(\tau_{i_1}, \tau_{i_2}), \ldots, (\tau_{m_1}, \tau_{m_2})\rangle$ be a sequence of chronologically ordered pairs of time-windows; P be an emerging pattern between the time-windows τ_{h_1} and τ_{h_2}, $\forall h \in \{i, \ldots, m\}$; $\langle GR_{\tau_{i_1},\tau_{i_2}}, \ldots, GR_{\tau_{m_1},\tau_{m_2}}\rangle$ be the values of growth-rate of P in the pairs $\langle(\tau_{i_1}, \tau_{i_2}), \ldots, (\tau_{m_1}, \tau_{m_2})\rangle$ respectively; $\Theta_P : \Re \rightarrow \Psi$ be a function which maps $GR_{\tau_{h_1},\tau_{h_2}}(P)$ into a discrete value $\psi_{\tau_{h_1},\tau_{h_2}} \in \Psi, \forall h \in \{i, \ldots, m\}$, P is a periodic change iff:

1. $|T| \geq minREP$
2. (τ_{h_1}, τ_{h_2}) *and* (τ_{k_1}, τ_{k_2}) *are δ-separated* $\forall h \in \{i, \ldots, m-1\}$, *k=h+1 and there is no pair* (τ_{l_1}, τ_{l_2}), $h < l$, *s.t.* (τ_{h_1}, τ_{h_2}) *and* (τ_{l_1}, τ_{l_2}) *are δ-separated*
3. $\psi = \psi_{\tau_{i_1},\tau_{i_2}} = \ldots = \psi_{\tau_{m_1},\tau_{m_2}}$

where $minREP$ is a minimum user-defined threshold and Θ can be implemented as any discretization technique. A PC is a frequent pattern whose support increases (decreases) at least $minREP$ times with an order of magnitude greater than $minGR$. Each change (increase/decrease) occurs within δ time-points and it is quantified by a discrete value $\psi \in \Psi$. Intuitively, a PC represents a variation, manifested with a particular periodicity, of the frequency of the conjunction P of weather parameters. We denote a periodic change PC with the notation $\langle P, T, \psi\rangle$.

Definition 3. Spatio-temporal Periodic Change (SPC)

Let $T : \langle(\tau_{i_1}, \tau_{i_2}), \ldots, (\tau_{u_1}, \tau_{u_2})\rangle$ be a sequence of chronologically ordered pairs of time-windows, let $\Pi : \{PC_1 : \langle P, T_1, \psi\rangle, \ldots, PC_v : \langle P, T_v, \psi\rangle\}$ be a set of periodic changes detected in v different geographic areal units, P is a spatio-temporal periodic change iff

1. $|\Pi| \geq minUNITS$
2. $(\tau_{h_1}, \tau_{h_2}) \in T_k \forall h \in \{i, \ldots, u\}, k = 1, \ldots, v$
3. (τ_{h_1}, τ_{h_2}) *and* (τ_{k_1}, τ_{k_2}) *are δ-separated* $\forall h \in \{i, \ldots, u-1\}, k = h+1$

Intuitively, a SPC represents a periodic variation (quantified by ψ) of the frequency of the conjunction P of weather parameters, which has been observed in v different geographic areal units.

3 The Method

In this section we propose a method to mine SPCs from the measurements of the weather parameters $A_1, ...A_d$ recorded by sensors equally displaced in a geographic zone over the sequence of time-points $\{t_1 \dots t_n\}$. The method is structured in two steps performed consecutively (see Figure 1). Initially, we build a gridded data space over the input geographic zone in order to define the areal units as cells of equal size $\{c_{11}, \dots, c_{\alpha,\beta}\}$. This means that the cells comprise the same number of sensors. The first step works on the values of the weather parameters of each cell c_{rs} and mines PCs in accordance with the definition 2. The second step inputs the PCs detected on all the cells, it selects the PCs which are present in at least $minUNITS$ cells and then mines SPCs in accordance with the definition 3. The details of these two steps are reported in the following.

3.1 Detection of periodic changes

To detect PCs, we adapt the algorithm proposed in [9], which we originally designed for data represented in relational logic, to the case of multi-dimensional time-series. In particular, it works on the succession $\langle(\tau_{1_1}, \tau_{1_2}), \dots, (\tau_{h_1}, \tau_{h_2}), \dots, (\tau_{z_1}, \tau_{z_2})\rangle$ of pairs of time-windows obtained from $\{t_1, \dots, t_n\}$ (see Section 2). Each time-window τ_{u_v} (except the first and last one) is present in two consecutive pairs, so, given two pairs (τ_{h_1}, τ_{h_2}) and $(\tau_{(h+1)_1}, \tau_{(h+1)_2})$, we have that $\tau_{u_v} = \tau_{h_2} = \tau_{(h+1)_1}$. This is done to capture the changes of support of the patterns from τ_{h_1} to τ_{u_v} and from τ_{u_v} to $\tau_{(h+1)_2}$. The algorithm performs three main procedures.

1. Discovery of frequent patterns for each time-window. Frequent patterns are discovered from each time-window with a technique of evaluation-generation of candidate patterns based on the monotonicity property of the support [8]. Obviously, the use of this technique does not exclude the possibility of using alternative solutions, which does not imply modifications to our proposal.
2. Extraction of the EPs from the frequent patterns discovered on τ_{h_1} against the frequent patterns discovered from τ_{h_2} in accordance with the definition 1. To perform this efficiently operation, we can act on the evaluation of some patterns. Indeed, we avoid the evaluation of a pattern $P2$, which is super-set of a pattern $P1$ ($P1 \subset P2$), if $P1$ is frequent in the time-window τ_{h_1} (τ_{h_2}) but it is not frequent in the time-window τ_{h_2} (τ_{h_1}). However, we cannot apply no optimization with respect to the growth-rate because, unfortunately, the monotonicity property does not hold. In fact, given two frequent patterns $P1$ and $P2$ with $P1 \subset P2$, if $P1$ is not emerging, namely $GR_{\tau_{h_1},\tau_{h_2}}(P1) < minGR$ ($GR_{\tau_{h_2},\tau_{h_1}}(P1) < minGR$), then the pattern $P2$ may or may not be an EP, namely its growth-rate could exceed the threshold $minGR$.
   The final EPs are stored in the pattern base, which hence contains the frequent patterns that satisfy the constraint set by $minGR$ on at least one pair of time-windows. Each EP is associated with two lists, named as $TWlist$ and $GRlist$. $TWlist$ is used to store the pairs of time-windows in which the growth-rate of the pattern exceeds $minGR$, while $GRlist$ is used to store

the corresponding values of growth-rate. The technical details can be found in the paper [9].

3. Detection of PCs from the EPs stored in the pattern base. To implement the function Θ_P (Definition 2) we use an equal-width discretization technique which returns a set of ranges, which we use as set of discrete values Ψ. The discretization technique is applied to the set of values of the lists *GRlist* of all the stored EPs. Thus, we can map the numeric values of the lists *GRlist* into their associated ranges in Ψ. This means that an individual EP may have different discrete values and may generate several PCs since one discrete value is assigned to one PC. Thus, we find PCs by working on the EPs one at a time. A PC is built incrementally by examining the pairs of time-windows of *TWlist* in chronological order and joining those pairs that have the same discrete value ψ on the condition that they are δ-separated.
   In order to clarify how the detection of PCs works, we report an explanatory example. Consider the time-points as years, $\Psi = \{\psi', \psi''\}$, $minREP$=3, δ=13 and the lists *TWlist* and *GRlist* built as follows:

$$\begin{aligned}
&TWlist : \langle ([1970;1972],[1973;1975]), ([1976;1978],[1979;1981]), ([1982;1984],[1985,1987]),\\
&\quad ([1988;1990],[1991;1993]), ([1994;1996],[1997;1999]), ([2010;2012],[2013;2015])\rangle\\
&GRlist : \langle \qquad \psi', \qquad \psi', \qquad \psi'',\\
&\qquad \psi', \qquad \psi'', \qquad \psi'\rangle
\end{aligned}$$

   By scanning the list *TWlist*, we can initialize the set T of a candidate PC' by using the pairs ([1970;1972],[1973;1975]) and ([1976;1978],[1979;1981]) since they are δ-separated (1979-1973$< \delta$) and they have the same discrete value ψ'. The pair ([1982;1984],[1985;1987]) instead refers to a different discrete value (ψ'') and therefore it cannot be inserted into T of PC'. We use it to initialize the set T of a new candidate PC", which thus will include the time-windows referred to ψ''. Subsequently, the pair ([1988;1990],[1991;1993]) is inserted into T of PC' since its distance from the latest pair is less than δ (1991-1979$< \delta$). Then, T of PC" is updated with ([1994;1996],[1997;1999]) since 1997-1985 is less than δ, while the pair ([2010;2012],[2013;2015]) cannot be inserted into T because the distance between 2013 and 1997 is greater than δ. Thus, we use the pair ([2010;2012],[2013;2015]) to initialize the set T of a new candidate PC"'. The set T of PC' cannot be further updated, but, since its size exceeds $minREP$, we consider the candidate PC' as valid periodic change. Finally, the candidate PC" cannot be considered as valid since its size is less than $minREP$. The candidate PC"' is not even considered since its sequence T has less than $minREP$ elements.

3.2 Mining spatio-temporal periodic changes

As result of the first step, we have a set of PCs for each cell. A preliminary operation we perform is the removal of redundant PCs. Indeed, the invalidity of the property of monotonicity of the growth-rate and the procedure of detection of PCs do not allow us to exclude the presence of redundancies, that is, PCs whose information is expressed also by other PCs. For instance, given two PCs,

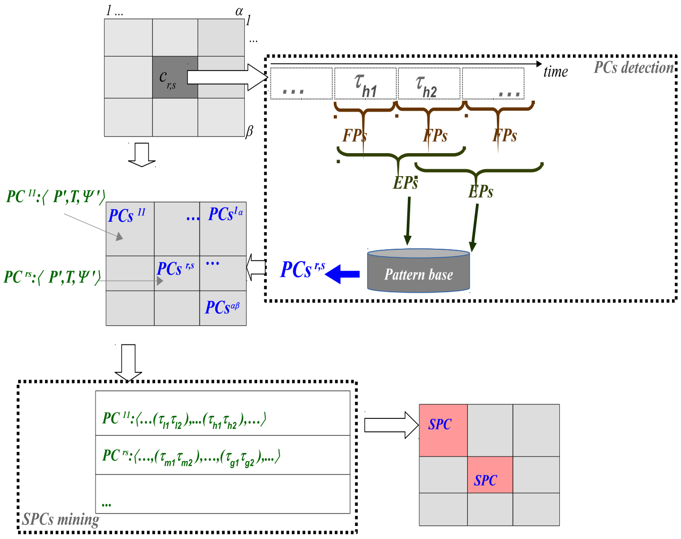

Fig. 1. The block-diagram of the two-step method for mining spatio-temporal pattern of periodic changes.

PC':$\langle P', T', \psi \rangle$ and PC":$\langle P'', T'', \psi \rangle$, P' is redundant if P'' has at least the same conjunction of weather parameters of P', $P' \subset P''$ and it changes at the same time-windows of P', $T' \subset T''$, with the same magnitude ψ.

After having removed the redundant PCs, to find SPCs we should act on the sequences T. Different alternatives can be considered, which we discuss briefly in the following. Using a grouping/clustering algorithm could turn out to be inapplicable because the lengths of T can be different. This is also the reason for which we cannot adopt algorithms for the generation of frequent itemsets. The distance-based technique, for instance those implementing the dynamic-time-warping distance, could be ineffective because, although they can handle sequences with different lengths, they return groups of sequences with similar time-windows, whilst we search for sequences having identical time-windows. We propose to face this point with a sequence mining approach, which naturally handles sequences of different lengths and satisfies the chronological order of the time-windows [11]. Here, the input data of the sequence mining problem is the set of the sequences T of one PC in common to several cells, for instance $\{PC^{11}, \ldots, PC^{rs}\}$ in Figure 1. More precisely, we take the set of the sequences T associated with a specific emerging pattern P' having a specific discrete value ψ'. The output is the complete set of SPCs in form of sequential patterns whose elements are pairs of time-windows. By considering that there are different PCs, the algorithm of sequence mining is applied to one collection of sets of sequences, whose cardinality is equal to the total number of PCs. Not all the PCs are used but only those found in at least $minUNITS$ cells.

The problem of redundant patterns could also affect the SPCs, therefore we decide to use an algorithm able to mine *closed* sequential patterns. A sequential pattern S' is closed if there exists no sequential pattern S'' such that $S' \subset S''$ and S'' occurs in the same sequences of S'. The use of closed sequential patterns allows us to maximize both the number of cells in which the change occurs and number of repetitions of the change in each cell.

In this work, the closed SPCs are discovered with an implementation of the algorithm CloSpan [15]. It follows the candidate maintenance-and-test approach. More precisely, it first generates a set of closed sequence candidates, which is stored in a hash-based tree structure and then performs a post-pruning operation on that set. The post-pruning operation exploits search space techniques. Obviously, the use of the algorithm CloSpan does not exclude the possibility of alternative solutions, which does not imply modifications to our proposal.

Finally, not all the closed sequential patterns are considered but only those that meet two conditions: i) the pairs of time-windows are δ-separated and ii) the grid cells associated to the patterns are adjacent. These cells denote together the spatial region in which a periodic change occurs.

4 Experiments

Dataset description. We apply the proposed method to real-world climate data generated from the NCEP/NCAR Reanalysis project and available on the data bank NOAA [12]. The climate data were recorded every day from January 1997 to December 1999 by 697 sensors uniformly distributed over a grid of 41x17 points (41 sensors by longitude, 17 sensors by latitude). So, totally we have 1094 daily measurements, therefore 1094 time-points. The distribution of the sensors delimits a specific geographic zone localized between Atlantic Ocean and Indian Ocean and covers almost 36,000,000 km^2. The weather parameters are "Air temperature", "Pressure", "Relative humidity", "Eastward Wind", "Northward Wind" and "Precipitable Water". We pre-processed the time-series by using an equal-frequency discretization technique with 5 ranges. The number of the ranges used for the function Θ_P is 5.

Experimental Setup. We built three different configurations of the grid from the geographic zone. In each configuration, the grid cells cover the same number of sensors and therefore have the same area. Specifically, the distribution of the sensors in each cell is 10x8, 5x8, 8x4, respectively, so the three configurations have 8 cells, 16 cells, 20 cells. Experiments are performed by tuning $minGR$, δ and $minREP$. The value of minimum support for the step of PCs detection is fixed to 0.1, while the value of minimum support for the step of SPCs mining is fixed to 0.5 in order to find patterns which cover at least the half of the minimum number of cells fixed by $minUNITS$. The value of $minUNITS$ equals to the half of the total number of cells for each grid configuration, that is, 4, 8, 10 respectively. The value of the width w of the windows is 30 (days).

Results. We collected three kinds of quantitative results. Specifically, Table 1 illustrates the average values of PCs computed on the number of cells and the total number of SPCs. Table 2 reports the evaluation of the SPCs in form of average portion of cells in which the final SPCs occur. More precisely, the evaluation considers the number of cells covered by the SPCs divided by the minimum number of requested cells ($minUNITS$) and has values in [0;1], where 1 refers the best coverage and indicates that the SPCs cover all the cells provided by $minUNITS$. We report the average computed on the three grid configurations. In the following, we discuss the influence of the input thresholds $minGR$, δ and $minREP$ on these results.

Discussion. In the boxes (a), (b) and (c) we report the results obtained with the three grid configurations. We see that the smaller the area of the cells the lower the number of PCs and SPCs, meaning that the method is able to capture an expected behavior, that is, there is higher variability in spatial regions with greater extent. As to the influence of $minGR$, we observe that there not are PCs and SPCs when it is higher than 6 in all the three grid configurations. This indicates that there is no conjunction of weather parameters whose frequency increases or decreases by an order of magnitude higher of 6. By increasing only the threshold δ, we have an higher average of PCs. Indeed, at higher values of δ the method detects both the changes which are replicated more frequently (that is, at δ=90) and the changes with more distant repetitions (that is, at δ=365). Consequently, the sets of distinct PCs (which are the input to the step of SPCs mining) are greater, with the result of having an higher number of distinct SPCs. By increasing only the threshold $minREP$, we obtain smaller sets of PCs. This is explained with the fact that higher values of $minREP$ require climate changes with a relatively high number of repetitions. This is requirement that not all the PCs can satisfy but only those with longer sequences of T. It should be noted that the threshold $minREP$ influences both the number and length of the SPCs. Specifically, at higher values of $minREP$, we have a smaller set of final SPCs because there are less distinct PCs as input of the step of SPCs mining. Moreover, the SPCs are longer because the sequences T of the PCs are longer.

Table 2 reports a quantitative evaluation of the SPCs. We have the better coverage (more than the two third of the requested cells) at the lowest values of $minGR$ and $minREP$ and highest value of δ. By considering only $minGR$, we observe that the better result is obtained at $minGR$=2, which corresponds to SPCs with mild changes. Instead, when $minGR$ >4, which corresponds to stronger changes, the SPCs are present in less cells. By considering only δ, we note that there is a discrete coverage of the cells with δ=90 or δ=120. This can be explained by the lower number of SPCs. Finally, by increasing only the threshold $minREP$, the coverage decreases because of the combined effect of the number of the SPCs and their length. This is not surprising because weather changes with less repetitions occur in larger spatial regions, while those with more repetitions are present in smaller regions.

$minGR$			
2	4	6	8
71–32	26–17	0-0	0–0

δ			
90	120	180	365
20–12	21–12	58–17	71–32

$minREP$			
3	4	5	6
71–32	29–12	17–3	4–0

(a)

$minGR$			
2	4	6	8
32–6	16–4	0–0	0–0

δ			
90	120	180	365
2–0	9–2	10–2	32–6

$minREP$			
3	4	5	6
32–6	14–3	3–0	0–0

(b)

$minGR$			
2	4	6	8
27–7	9–0	0–0	0–0

δ			
90	120	180	365
2–0	5–3	11–4	27–7

$minREP$			
3	4	5	6
27–7	19–6	9–0	0–0

(c)

Table 1. Results obtained when tuning $minGR$, δ and $minREP$. Each slot of the tables reports the average values of PCs and the total number of SPCs. The average values of PCs are computed on the number of the cells. The boxes (a), (b) and (c) refer to the grid configurations with 8 cells, 16 cells and 20 cells respectively. For the tables $minGR$, we set δ=365, $minREP$=3. For the tables δ, we set $minGR$=2, $minREP$=3. For the tables $minREP$, we set $minGR$=2, δ=365.

5 Related Work

The analysis of climate data has always attracted interest by different disciplines and the study of the dynamics is considered particularly relevant for the effects on the Earth. Gunnemann et al. [4] work on the hypothesis that the changes can regard subspaces of the descriptive attributes. Then, they describe a clustering technique based on the similarity which tracks the changes of subspaces in time-variable climate data and associates a type of climate behaviour with each cluster. Kleynhans et al. [6] propose a method to detect and evaluate land cover change by examining at each point in time for a specific pixel neighborhood the spatial covariance of a hyper-temporal time series. McGuire et al. [10] introduce the problem of mining moving dynamic regions. Their solution is based on spatial auto-correlation and finds dynamic spatial regions across time periods and dynamic time periods over space. Finally, moving dynamic regions are identified by determining the spatio-temporal connectivity, extent, and trajectory for groups of locally dynamic spatial locations whose position has shifted from one time period to the next. Lian et al [7] propose an algorithm to detect high change regions based on quadtree-based index and classify heterogeneous and homogeneous change. Finally, spatio-temporal changes are analyzed at long time scales to find high change persistent regions and high change dynamic regions.

$minGR$			
2	4	6	8
0.71	0.52	–	–

δ			
90	120	180	365
0.53	0.59	0.66	0.71

$minREP$			
3	4	5	6
0.71	0.56	0.51	–

Table 2. A quantitative evaluation of the SPCs in terms of average portion of distinct cells covered by the final SPCs.

The periodicity has been often seen as effect to be removed from the climate data because makes the applicability of the classical methods unfeasible. Tan et al [13] present a comprehensive study based on classical pattern discovery algorithms to find spatio-temporal patterns from spatial zones over time. Preliminarily, seasonal variation is removed from data with data transformation techniques, like discrete Fourier transform. Patterns denote regularities within individual zones, among different zones, within the same time-interval or along a series of time-intervals. Spatio-temporal patterns are the main subject of study in trajectory mining. In [5] the authors propose unifying incremental approaches to automatically extract different kinds of spatio-temporal patterns by applying frequent closed item-set mining techniques.

6 Conclusions

The research presented in this paper has two main contributions. First, we extend a previous method in order to identify different occurrences of the same periodic changing behavior. Second, we explore the possibility to identify periodic changing behaviors in the climate data domain, which is typically characterized by temporal and spatial component. We have introduced the notion of spatial-temporal pattern of periodic changes to denote the spatial extent of variations repeated on the temporal axis. The proposed method relies on the frequent pattern mining framework, which enables us to *i)* capture the changes in terms of variations of the frequency, *ii)* estimate the regularity over time of these changes, and *iii)* identify contiguous areal units in which the change can be tracked. The application to a real dataset highlights the viability and usefulness of the proposed method to a real-world problem. The results show the appropriateness of some decision algorithms, for instance the integration of techniques to remove uninteresting patterns allows us to have sets of SPCs which humans can interpret without overwork. As future work, we plan to apply the approach to other evolving domains with spatio-temporal components.

Acknowledgements. The authors would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

1. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):823–839, 2012.
2. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, 1999.
3. J. H. Faghmous and V. Kumar. *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities*, chapter Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities, pages 83–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
4. S. Günnemann, H. Kremer, C. Laufkötter, and T. Seidl. Tracing evolving subspace clusters in temporal climate data. *Data Min. Knowl. Discov.*, 24(2):387–410, 2012.
5. P. N. Hai, P. Poncelet, and M. Teisseire. Get_move: An efficient and unifying spatio-temporal pattern mining algorithm for moving objects. In J. Hollmén, F. Klawonn, and A. Tucker, editors, *Advances in Int. Data Analysis XI - 11th Inter. Symp., IDA 2012, Helsinki, Finland, 2012. Proc.*, volume 7619 of *LNCS*, pages 276–288, 2012.
6. W. Kleynhans, B. P. Salmon, and K. J. Wessels. A novel spatio-temporal change detection approach using hyper-temporal satellite data. In *2014 IEEE Geoscience and Remote Sensing Symposium, IGARSS 2014, Quebec City, QC, Canada, July 13-18, 2014*, pages 4208–4211. IEEE, 2014.
7. J. Lian and M. P. McGuire. *Mining Persistent and Dynamic Spatio-Temporal Change in Global Climate Data*, pages 881–891. Springer International Publishing, Cham, 2016.
8. F. A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. *Machine Learning*, 55(2):175–210, 2004.
9. C. Loglisci and D. Malerba. Mining periodic changes in complex dynamic data through relational pattern discovery. In M. Ceci, C. Loglisci, G. Manco, E. Masciari, and Z. W. Ras, editors, *New Frontiers in Mining Complex Patterns - 4th Inter. Work., NFMCP 2015, Held with ECML-PKDD 2015, Porto, Portugal, September 7, 2015, Revised Selected Papers*, volume 9607 of *LNCS*, pages 76–90. Springer, 2015.
10. M. P. McGuire, V. P. Janeja, and A. Gangopadhyay. Mining trajectories of moving dynamic spatio-temporal regions in sensor datasets. *Data Min. Knowl. Discov.*, 28(4):961–1003, 2014.
11. C. H. Mooney and J. F. Roddick. Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv.*, 45(2):19:1–19:39, Mar. 2013.
12. R. A. Simons. Erddap - the environmental research division's data access program. *http://coastwatch.pfeg.noaa.gov/erddap . Pacific Grove, CA: NOAA/NMFS/SWFSC/ERD.*, 2011.
13. P. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Finding spatio-temporal patterns in earth science data. In *Proc. of KDD Workshop on Temporal Data Mining*, 2001.
14. R. L. Wilby and T. M. L. Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548, 1997.
15. X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large databases. In D. Barbará and C. Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003*, pages 166–177. SIAM, 2003.