Mining Spatio-Temporal Patterns of Periodic Changes in Climate Data

Corrado Loglisci[1,2(✉)], Michelangelo Ceci[1,2], Angelo Impedovo[1,2], and Donato Malerba[1,2]

[1] Department of Computer Science, Universita' degli Studi di Bari "Aldo Moro", Bari, Italy
{corrado.loglisci,michelangelo.ceci,angelo.impedovo, donato.malerba}@uniba.it

[2] CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy

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 study of the 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 [5], which opens 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 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 building and refining predictive models with changes and events of the weather

A. Appice et al. (Eds.): NFMCP 2016, LNAI 10312, pp. 198–212, 2017.
DOI: 10.1007/978-3-319-61461-8_13

conditions. Although in data mining we can find a long list of works on event and change detection [3], the identification of changes in climate data is challenging for several reasons. First, climate data tend to be noisy, therefore 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, while they could be no longer appropriate capture features of the regional weather conditions, where instead local models could be effective [17].

In Climatology, many phenomena are cyclic in nature and can exhibit repetitive behaviors. Likewise, changes in weather conditions can be periodic because they can be 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, in particular, 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. It proceeds in two subsequent steps, first detects periodic changes at the level of individual cells of the grid and then it finds sequential patterns of the periodic changes only 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 experience working on (global) statistical regularities. More precisely, in the first step, we combine a time windows-based analysis model with a frequent pattern mining method, in order 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 different cells. Sequential patterns allows us to find changes at a higher level of spatial granularity based on aggregations of cells.

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

2 Basics and Definitions

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

Let $\{t_1 \dots t_n\}$ be 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 $\{t_i, \dots, t_j\}$ $(t_1 \leq t_i, t_j \leq t_n)$, 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 pair is composed by a weather parameter and its value. It can have at most d pairs, which is the number of weather parameters. 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 $\frac{sup_{\tau_{h_1}}(P)}{sup_{\tau_{h_2}}(P)} \geq minGR \vee \frac{sup_{\tau_{h_2}}(P)}{sup_{\tau_{h_1}}(P)} \geq minGR$

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 [4]. In its classical formulation, it refers to the values of support of a pattern discovered on two different classes of data, while, in this work, we extend that notion to represent the differences between the data collected in two intervals of time, and therefore, we refer to the values of support of a pattern which has been discovered on 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* $\tau_{h_2}, \forall h \in \{i, \ldots, m\}$; $\langle GR_{\tau_{i_1},\tau_{i_2}}, \ldots, GR_{\tau_{m_1},\tau_{m_2}}\rangle$ *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* $GR_{\tau_{h_1},\tau_{h_2}}(P)$ *to a nominal 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* $(\tau_{l_1}, \tau_{l_2}), h < l$, *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. The function Θ is used to handle the numerical information associated to the growth-rate and allows us to crisply distinguish the magnitude of different growth-rate values. 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 represented by the nominal value $\psi \in \Psi$. We denote a periodic change PC with the notation $\langle P, T, \psi\rangle$. An example of periodic change is reported here. Consider the pattern

$$P : air\_temperature = [301; 307], pressure = [95; 100], relative\_humidity = [60; 70]$$

where $sup_{apr\_2011}(P) = sup_{apr\_2012}(P) = sup_{apr\_2013}(P) = 0.25$, $sup_{may\_2011}(P) = sup_{may\_2012}(P) = sup_{may\_2013}(P) = 0.5$, $sup_{nov\_2011}(P) = sup_{nov\_2012}(P) = sup_{nov\_2013}(P) = 0.5$, $sup_{dec\_2011}(P) = sup_{dec\_2012}(P) = sup_{dec\_2013}(P) = 0.1$. Here, the values of the support of the pattern P increase through the pairs of the windows $[apr\_2011, may\_2011]$, $[apr\_2012, may\_2012]$ and $[apr\_2013, may\_2013]$ respectively, indeed the values of growth-rate $GR_{apr\_2011,may\_2011}(P)$, $GR_{apr\_2012,may\_2012}(P), GR_{apr\_2013,may\_2013}(P)$ are equal to 2 (0.5/0.25). While, the values of the support of the pattern P decrease through the pairs of the windows $[nov\_2011, dec\_2011]$, $[nov\_2012, dec\_2012]$ and $[nov\_2013, dec\_2013]$ and the values of growth-rate $GR_{dec\_2011,nov\_2011}(P), GR_{dec\_2012,nov\_2012}(P)$, $GR_{dec\_2013,nov\_2013}(P)$ are equal to 5. By supposing $minGR = 1.5$, the pattern P is considered emerging over the windows $[apr\_2011, may\_2011]$, $[nov\_2011, dec\_2011]$, $[apr\_2012, may\_2012]$, $[nov\_2012, dec\_2012]$, $[apr\_2013, may\_2013]$ and $[nov\_2013, dec\_2013]$. However, in the windows $[nov\_2011, dec\_2011]$, $[nov\_2012, dec\_2012]$ and $[apr\_2013, may\_2013]$ its variation of support is different from the variation detected in the windows $[apr\_2011, may\_2011]$, $[apr\_2012, may\_2012]$, $[apr\_2013, may\_2013]$ both in terms of quantity (5 against 2) and in terms of growth

(decrease against increase). This means that we could build different periodic changes from P. Indeed, by supposing a function Θ which maps the values of growth-rate 2 and 5 to the nominal values *weak_change* and *strong_change*, the values of $minREP$ and δ equal 2 and 365 days respectively, we can generate two PCs having the same conjunction of weather parameters.

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* v *periodic changes detected in* v *different geographic areal units,* P *is a spatio-temporal periodic change iff*

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

Intuitively, a SPC represents a periodic variation (quantified by ψ) of the frequency of weather parameters conjunction P. Such a variation is 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 over a geographic area on the sequence of time-points $\{t_1 \ldots t_n\}$. The method is structured in two steps performed consecutively (see Fig. 1). Initially, we build a gridded data space over the input geographic area in order to define the areal units as cells of equal size $\{c_{11}, \ldots, c_{\alpha,\beta}\}$. 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 [11] originally designed for data represented in relational logic, to the case of multi-dimensional time-series. In particular, it works on the succession $\langle(\tau_{1_1}, \tau_{1_2}), \ldots, (\tau_{h_1}, \tau_{h_2}), \ldots, (\tau_{z_1}, \tau_{z_2})\rangle$ of pairs of time-windows obtained from $\{t_1, \ldots, t_n\}$ (see Sect. 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 the technique of evaluation-generation of candidate patterns used in [11], which exploits the monotonicity property of the support. Obviously, the decision of using that specific technique does not exclude the possibility of considering alternative solutions based on evaluation-generation of patterns, which do not imply modifications neither to our proposal nor to the set of frequent patterns resulting from the current procedure.
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 efficiently perform this operation, we can act on the support of the 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}). Instead, we cannot apply no optimization on 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 emerging, namely its growth-rate could exceed the threshold $minGR$.

   The final EPs are stored in a 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 [11].
3. Detection of PCs from the EPs stored in the pattern base. To implement the function Θ_P (Definition 2) we resort to an equal-width discretization technique, which is able to return a set of ranges used here as nominal 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 a value of growth-rate to the range in which the value falls in. The choice of the equal-width discretization allows us to take the different magnitude orders into account and uniformly map the growth-rate values into different ranges, without making the distribution of the values unbalanced.

   The PCs are built with a procedure of generation-evaluation of candidates. In particular, we work on the EPs one at a time by generating as many candidates as the nominal values associated with the growth-rate of that EP. 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 nominal 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 of generation of PCs from one EP. Consider the time-points as years, $\Psi = \{\psi', \psi''\}$, $minREP = 3$, $\delta = 13$ and the lists $TWlist$ and $GRlist$ built as follows (the nominal value has the same position in $GRlist$ of the corresponding pair of time-windows in $TWlist$):

$TWlist$: ⟨([1970; 1972], [1973; 1975]) , ([1976; 1978], [1979; 1981]) , ([1982; 1984], [1985, 1987]) , ([1988; 1990], [1991; 1993]) , ([1994; 1996], [1997; 1999]) , ([2010; 2012], [2013; 2015])⟩
$GRlist$: $\psi', \psi', \psi'', \psi', \psi'', \psi'$⟩

By scanning the list $TWlist$, we can initialize the sequence 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 nominal value ψ'. The pair ([1982;1984], [1985;1987]) instead refers to a different nominal value (ψ'') and therefore it cannot be inserted into T of PC'. We use it to initialize the sequence 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 sequence T of a new candidate PC"'. The sequence 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.

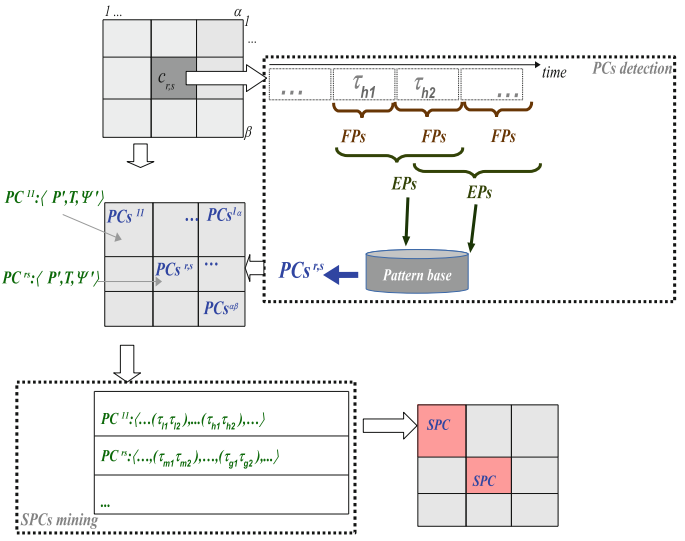


Fig. 1. The block-diagram of the two-step method for mining spatio-temporal patterns of periodic changes.

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, PC': $\langle P', T', \psi \rangle$ and PC": $\langle P'', T'', \psi \rangle$, P' is redundant if (i) the conjunction of weather parameters of P'' includes the conjunction of weather parameters of P' ($P' \subset P''$); (ii) the pairs of time-windows of PC" comprise those of PC' ($T' \subset T''$); (iii) they have the same nominal value 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 techniques, for instance those implementing the dynamic-time-warping distance, could be ineffective because, although able to handle sequences of different lengths, they return groups of sequences with similar/close time-windows, whilst we are interested in obtaining sequences with identical time-windows. Our proposal is investigating this problem with a sequence mining approach, which naturally handles sequences of different lengths and takes the chronological order of the time-windows into account [14]. 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 Fig. 1. So, we take the set of the sequences T associated with a specific emerging pattern P' having a specific nominal 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 for the sequence mining algorithm but only those found in at least $minUNITS$ cells.

Here, we could experience the problem of redundant patterns, so 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 additionally maximize both the number of cells in which the change occurs and number of repetitions of the change in each cell. We exploit the algorithm CloSpan [18], which implements a candidate maintenance-and-test approach. 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 decision of using the algorithm CloSpan does not exclude the possibility of considering alternative solutions. Indeed, other algorithms of closed sequential patterns mining do not imply modifications to the method, considering that our purpose here is the generation of the minimal set of frequent sequences of pairs of the windows for each periodic change.

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

We applied the proposed method to real-world climate data generated from the NCEP/NCAR Reanalysis project and available on the data bank NOAA [15]. The climate data were recorded every day from January 1997 to December 1999 by 697 sensors uniformly distributed over a grid of 41×17 points (41 sensors by longitude, 17 sensors by latitude). So, totally we have 1094 daily measurements (1094 time-points). The distribution of the sensors delimits a specific geographic area 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".

Experimental Setup. We pre-processed the time-series by using an equal-frequency discretization technique, which guarantees a uniform distribution on the five (discretized) ranges of the same parameter when generating patterns with the ranges of different parameters. In particular, for each parameter, we considered 5 ranges. To implement the function Θ_P, we applied an equal-width discretization technique to the values of the growth-rate experimentally obtained, which fall in the interval $[1.2, 5]$. The number of the ranges generated is 6, namely $\{[1.5, 2), [2, 2.5), [2.5, 3), [3, 3.5), [3.5, 4), [4, 4.5)\}$, to which we manually assign the nominal values $very\_weak\_change, weak\_change, middle\_weak\_change$, $middle\_strong\_change, strong\_change, very\_strong\_change$. So, the function Θ_P maps values from the interval $[1.2, 5]$ to the set $\{very\_weak\_change, weak\_change, middle\_weak\_change$, $middle\_strong\_change, strong\_change, very\_strong\_change\}$.

We built three different configurations of the grid from the geographic area. In each configuration, the grid cells cover the same number of sensors and therefore have the same size. Specifically, the distribution of the sensors in each cell is 10×8, 5×8, 8×4, respectively, so the three configurations have 8 cells, 16 cells, 20 cells. Experiments are performed by tuning $minGR, \delta$ 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 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 values of PCs averaged by 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 $minUNITS$. In the following, we discuss the influence of the input thresholds $minGR, \delta$ and $minREP$ on these results.

Discussion. In the boxes (a), (b) and (c) of Table 1, 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 a quite expected behavior, that is, the spatial regions with greater extent show there higher variability of the weather conditions compared with the smaller regions. As to the influence of $minGR$, we observe that there not are PCs and SPCs when it is higher than 6. 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 greater sets of PCs. Indeed, at higher values of δ the method detects both the changes which are replicated more frequently (that is, at $\delta = 60$ days) and the changes which are replicated less frequently, that is, with distant repetitions ($\delta = 365$). Consequently, the sets of the PCs (which are the input of the step of SPCs mining) are greater and this implies the discovery of greater sets of SPCs.

By increasing only the threshold $minREP$, we obtain smaller sets of PCs. In fact, when setting higher values of $minREP$, we require climate changes with a relatively high number of repetitions, which is a requirement that only the PCs with longer sequences of T can satisfy. Consequently, the number of PCs that feeds the second step (SPCs mining) is lower and the set of SPCs is smaller but it is composed by the longer SPCs since generated with longer PCs. This is evident whether comparing the tables in the box (a) against those in the boxes (b) and (c). A concrete example is when $minREP$ is 5 ($w = 90$ days). In that case, we search changes repeated at a distance of even 5 semesters, that is, almost the whole dataset (6 semester long).

Table 2 reports a quantitative evaluation of the SPCs. We see that the better coverage is almost three-quarters of the requested cells and it is reached 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 $minGR = 1.5$, which corresponds to SPCs with "very weak changes". Instead, when $minGR > 4$, we have SPCs with "very strong changes" but replicated in a smaller set of cells. By considering only δ, we note that there is a discrete coverage of the cells at relatively low values δ. 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.

Table 1. Results obtained by tuning a parameter at time on three grid configurations, that is, 8 cells (a), 16 cells (b) and 20 cells (c). When tuning $minGR, \delta$ is 365 and $minREP$ is 3. When tuning $\delta, minGR$ is 2 and $minREP$ is 3. When tuning $minREP, minGR$ is 2 and δ is 365. 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.

$minGR$			
1.5	2	4	6
102–47	71–32	26–17	0-0

δ				
60	90	120	180	365
9–2	20–12	21–12	58–17	71–32

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

(a)

$minGR$			
1.5	2	4	6
54–2	32–6	16–4	0–0

δ				
60	90	120	180	365
2–0	2–0	9–2	10–2	32–6

$minREP$			
3	4	5	6
32–6	14–3	3–0	0–0

(b)

$minGR$			
1.5	2	4	6
42–11	27–7	9–0	0–0

δ				
60	90	120	180	365
2–0	2–0	5–3	11–4	27–7

$minREP$			
3	4	5	6
27–7	19–6	9–0	0–0

(c)

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

$minGR$			
1.2	2	4	6
0.72	0.71	0.52	–

δ				
60	90	120	180	365
0.55	0.53	0.59	0.66	0.71

$minREP$			
3	4	5	6
0.71	0.56	0.51	–

Interpretation of the Spatio-Temporal Patterns. Here we present some examples of SPCs mined from the real-word climate data and report the pairs of windows over which they are repeated and the modelled change. The grid cells are graphically drawn on the geographic map for ease of the interpretation.

For instance, the following SPC has been mined with $minGR = 2$, $\delta = 365$, $minREP = 3$

$$SPC_1 : [P : air\_temperature = [301.5; 307.2], pressure = [96, 99; 100],$$
$$relative\_humidity = [82.75; 89.75], precipitable\_water = [0.46; 10.89];$$
$$T : \langle([june\_1997, july\_1997], [may\_1998, june\_1998]), ([may\_1999, june\_1999])\rangle;$$
$$\Psi = middle\_weak\_change]$$

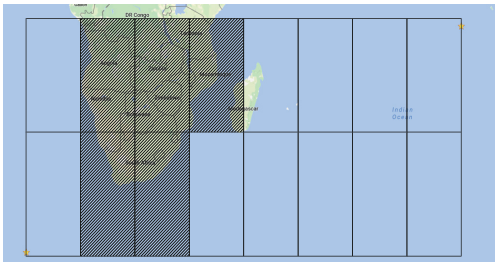
SPC_1 represents a change of frequency denoted as $middle\_weak\_change$, which corresponds to the range [2.5;3]. This variation recurs three times,

specifically over the pairs of windows $\langle([june\_1997, july\_1997], [may\_1998, june\_1998]), ([may\_1999, june\_1999])\rangle$ and it is replicated on the five cells drawn Fig. 2a. Intuitively, we see that such a change recurs with a periodicity of at most 12 months and covers the land of the geographic area under examination.

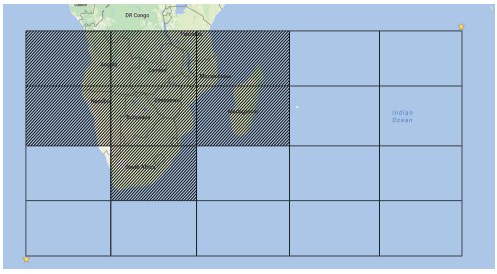
Another SPC the method discovered with $minGR = 2$, $\delta = 365$, $minREP = 3$ is the following

$$SPC_2 : [P : air\_temperature = [301.5; 307.2], pressure = [96, 99; 100], \\ relative\_humidity = [82.75; 89.75]; \\ T : \langle([june\_1997, july\_1997], [may\_1998, june\_1998]), ([may\_1999, june\_1999])\rangle; \\ \Psi = middle\_strong\_change]$$

It exhibits a frequency variation included in the range [3;3.5) ($middle\_strong\_change$) over the pairs of windows $T : \langle([june\_1997, july\_1997], [may\_1998, june\_1998]), ([may\_1999, june\_1999])\rangle$ and on the five cells drawn in dashed in Fig. 2b. We see that SPC_2 has a conjunction of weather parameters, which is a subset of SPC_1, additionally, it appears in the same sequence and same spatial region of SPC_1, but it denotes a stronger change. This means that the support of SPC_2 is higher than the support of SPC_1 either in the windows $june\_1997$, may_1998, may_1999 or in the windows $july\_1997$, $june\_1998$, $june\_1999$. This is explained with the monotonicity property of the support.



(a)



(b)

Fig. 2. Visualization of the grid obtained by collecting 5×8 (a) and 8×4 (b) sensors per cell. The spatio-temporal periodic changes SPC_1 and SPC_2 occur in the dashed cells of the grid (a). The spatio-temporal periodic changes SPC_2 occurs also in the dashed cells of the grid (b).

The same periodic change represented by SPC_2 has been mined with $minGR = 2$, $\delta = 365$, $minREP = 3$ by using the grid configuration with 20 cells (Fig. 2b). In this case, it has the same pairs of windows T that had within the grid at 16 cells and clearly covers a different subset of cells. We see that the spatial region of SPC_2 in the first grid (Fig. 2a) greatly overlaps the spatial region of SPC_2 in the second grid (Fig. 2b).

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. Günnemann et al. [6] 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. [8] 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. [13] 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 and McGuire [9] 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. In [1], the authors investigate a problem of change analysis with a descriptive method aiming at summarizing evolving data streams in spatial domains. They propose a clustering-based technique to detect groups of georeferenced data which vary according to a similar trend, which is determined over time-windows.

The periodicity has been often seen as a disturbance effect to be removed from the climate data because makes the applicability of the classical methods unfeasible. Tan et al. [16] 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. The study presented in [10] focused on the periodic variation of phenotype data and applied the solution to seasonal diseases. However, as our knowledge, very few attempts have been done to investigate the periodicity of the change over space and no attempt focused on the use of patterns. Boriah et al. [2] proposed a recursive merging algorithm that exploited the seasonality to distinguish between locations that experienced a land cover

change and locations that did not. However, it does provide no information on the change and on the spatial and temporal components associated to it. In [12] the authors investigated the effect of the periodicity in form of temporal autocorrelation for regression problems on time-stamped networks. Spatio-temporal patterns are the main subject of study in trajectory mining. In [7] 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 Climatology, 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. We performed experiments to test the sensibility of the method with respect to input thresholds. We plan to explore different future directions: (i) automatic determination of the input parameters, (ii) qualitative evaluation the discovered patterns against ground-truth on weather changes (iii) study of the usefulness of the patterns for predictive problems.

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. Appice, A., Ciampi, A., Malerba, D.: Summarizing numeric spatial data streams by trend cluster discovery. Data Min. Knowl. Discov. **29**(1), 84–136 (2015)
2. Boriah, S., Kumar, V., Steinbach, M., Potter, C., Klooster, S.: Land cover change detection: a case study. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 857–865. ACM, New York (2008)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection for discrete sequences: a survey. IEEE Trans. Knowl. Data Eng. **24**(5), 823–839 (2012)
4. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52 (1999)

5. Faghmous, J.H., Kumar, V.: Spatio-temporal data mining for climate data: advances, challenges, and opportunities. In: Chu, W.W. (ed.) Data Mining and Knowledge Discovery for Big Data. Studies in Big Data, vol. 1, pp. 83–116. Springer, Heidelberg (2014). doi:10.1007/978-3-642-40837-3_3
6. Günnemann, S., Kremer, H., Laufkötter, C., Seidl, T.: Tracing evolving subspace clusters in temporal climate data. Data Min. Knowl. Discov. **24**(2), 387–410 (2012)
7. Hai, P.N., Poncelet, P., Teisseire, M.: GeT_Move: an efficient and unifying spatio-temporal pattern mining algorithm for moving objects. In: Hollmén, J., Klawonn, F., Tucker, A. (eds.) IDA 2012. LNCS, vol. 7619, pp. 276–288. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34156-4_26
8. Kleynhans, W., Salmon, B.P., Wessels, K.J.: 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, 13–18 July 2014, pp. 4208–4211. IEEE (2014)
9. Lian, J., McGuire, M.P.: Mining persistent and dynamic spatio-temporal change in global climate data. In: Latifi, S. (ed.) Information Technology: New Generations. AISC, vol. 448, pp. 881–891. Springer, Cham (2016). doi:10.1007/978-3-319-32467-8_76
10. Loglisci, C., Balech, B., Malerba, D.: Discovering variability patterns for change detection in complex phenotype data. In: Esposito, F., Pivert, O., Hacid, M.-S., Raś, Z.W., Ferilli, S. (eds.) ISMIS 2015. LNCS (LNAI), vol. 9384, pp. 9–18. Springer, Cham (2015). doi:10.1007/978-3-319-25252-0_2
11. Loglisci, C., Malerba, D.: Mining periodic changes in complex dynamic data through relational pattern discovery. In: Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) NFMCP 2015. LNCS (LNAI), vol. 9607, pp. 76–90. Springer, Cham (2016). doi:10.1007/978-3-319-39315-5_6
12. Loglisci, C., Malerba, D.: Leveraging temporal autocorrelation of historical data for improving accuracy in network regression. Stat. Anal. Data Min. **10**(1), 40–53 (2017)
13. McGuire, M.P., Janeja, V.P., Gangopadhyay, A.: Mining trajectories of moving dynamic spatio-temporal regions in sensor datasets. Data Min. Knowl. Discov. **28**(4), 961–1003 (2014)
14. Mooney, C.H., Roddick, J.F.: Sequential pattern mining - approaches and algorithms. ACM Comput. Surv. **45**(2), 19:1–19:39 (2013)
15. Simons, R.A.: ERDDAP - the environmental research division's data access program. NOAA/NMFS/SWFSC/ERD, Pacific Grove (2011). http://coastwatch.pfeg.noaa.gov/erddap
16. Tan, P., Steinbach, M., Kumar, V., Potter, C., Klooster, S., Torregrosa, A.: Finding spatio-temporal patterns in earth science data. In: Proceedings of KDD Workshop on Temporal Data Mining (2001)
17. Wilby, R.L., Wigley, T.M.L.: Downscaling general circulation model output: a review of methods and limitations. Prog. Phys. Geogr. **21**(4), 530–548 (1997)
18. Yan, X., Han, J., Afshar, R.: CloSpan: mining closed sequential patterns in large databases. In: Proceedings of the Third SIAM International Conference on Data Mining, CA, USA, 1–3 May 2003, pp. 166–177 (2003)