5-2013

Disclosing climate change patterns using an adaptive markov chain pattern detection method

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

Gary LEE

Hoong Maeng CHAN

Reuben LI

Xiuju FU

See next page for additional authors

Published in 2013 International Conference on Social Intelligence and Technology (SOCIETY): May 8-10,
State College PA: Proceedings, pp. 72-79.
DOI: 10.1109/SOCIETY.2013.1

Disclosing climate change patterns using an adaptive Markov chain pattern detection method

Zhaoxia WANG, Gary LEE, Hoong Maeng CHAN, Reuben LI, Xiuju FU and Rick GOH
Department of Computing Science
Institute of High Performance Computing,
Singapore 138632
Email: wangz@ihpc.a-star.edu.sg

Pauline AW Poh Kim and Martin L. HIBBERD
Genome Institute of Singapore,
Singapore 138632
Email: hibberdml@gis.a-star.edu.sg

Hoong Chor CHIN
Department of Civil Environmental Engineering,
National University of Singapore
Singapore, 117578
E-mail: chin.hc@nus.edu.sg

***Abstract*—This paper proposes an adaptive Markov chain pattern detection (AMCPD) method for disclosing the climate change patterns of Singapore through meteorological data mining. Meteorological variables, including daily mean temperature, mean dew point temperature, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature and minimum temperature are simultaneously considered for identifying climate change patterns in this study. The results depict various weather patterns from 1962 to 2011 in Singapore, based on the records of the Changi Meteorological Station. Different scenarios with varied cluster thresholds are employed for testing the sensitivity of the proposed method. The robustness of the proposed method is demonstrated by the results. It is observed from the results that the early weather patterns that were present from the 1960s disappear consistently across models. Changes in temporal weather patterns suggest long-term changes to the climate of Singapore which may be attributed in part to urban development, and global climate change on a larger scale. Our climate change pattern detection algorithm is proven to be of potential use for climatic and meteorological research as well as research focusing on temporal trends in weather and the consequent changes.**

***Keywords*—Climate change, data mining, incremental Markov chain model, meteorological data, pattern detection, weather patterns of Singapore.**

I. Introduction and Related Work

Climate change is a widely recognized global environmental challenge [1]. A successful addressing of this challenge is essential to the sustainability of modern urban living, especially in domains such as environmental engineering, ecological management, human health, and global and regional economic systems. In this paper, we introduce techniques to understand and detect the patterns of climate change by using data from daily weather records. This, in turn, may benefit studies on public health (e.g. spread of dengue disease and incidences of respiratory diseases), energy consumption, while they also indirectly have an effect on urban problems such as mobility (e.g. flood events and traffic slowdown) which are affected by weather patterns.

Pattern detection, especially anomaly pattern detection, as a data mining task, refers to disclosing patterns that do not conform to expected behaviors in databases. These unusual patterns are often referred to as different terms in different application domains, such as rare patterns [2], outliers [3], [4], faults [5], peculiarities or contaminants, and etc [6].

Most existing pattern detection techniques resolve specific formulations of a problem. The formulations are induced by various factors such as the nature of the data, availability of the labelled data, and the type of anomalies to be detected, and etc. A considerable amount of data mining research on pattern detection has been conducted, and this stream has gained considerable interest owing to the realization that anomaly patterns can be detected from very large databases by data mining [4]. In addition, pattern detection is normally carried out through fault detection methods which analyze the current signals obtained from induction machines [5]. High-dimensional current signals are transformed to low-dimensional data by the mapping of the original signals into different clusters according to their characteristics [5].

Pattern detection methods in the field of data mining typically pick out different patterns (clusters), their changes, and rare/outlying events, and such changes are often the source of problems in impact studies [6]. They have been successfully applied to many fields. However, only a few studies have been adapted for environment, weather or climate applications.

From a meteorological viewpoint, research on climate change to disclose typical seasonal patterns of weather, such as the seasonal variability of thermal conditions in Singapore is very important [7], [8]. Natural variability in the climate of Singapore is influenced primarily by the monsoonal influences in various months of the year. Studies on Singapore data have shown a long-term increase in temperature in the past few decades [9], [10], in line with global warming studies over the Southeast Asian region [11]. The anomaly pattern or rare events may also be attributed to anomalous conditions such as the El Nino or Southern Oscillation (ENSO) phenomena [12]–[14].

On a larger scale, climate change is an unprecedented environment change that is affecting our planet [15], [16]. It is already having significant impacts on many aspects of our lives [15], [17]–[20]. Climate change projections on both the global and regional scales are characterized by multiple sources of uncertainty [17]. In order to characterize such uncertainties, global and regional climate model projections need to be based on probabilistic approaches using multi-model ensembles of experiments [17].

Kyung Soo Jun et al. discussed the impact of climate change on spatial water resources. The study was a new attempt to quantify hydrologic vulnerability that included the impacts of climate change [18]. The long term impact of climate change on the carbon budget of Lake Simcoe, Ontario were discussed by analyzing the relationship between temperature and dissolved inorganic carbon in some tributaries [20]. Anna Augustsson et al. discussed how the climate effect can be inserted in a commonly used exposure model, and how the exposure then changes compared to present conditions [19]. The results indicated that changes in climate are likely to affect the speciation, mobility, and risks associated with metals.

As mentioned previously, many research activities have been carried out to improve the understanding of climate change patterns by means of different techniques and made considerable contributions [20]–[23]. However, the introduction of data mining techniques into this research field has been limited. Therefore, this paper aims to develop a detection method based on data mining techniques for detecting and classifying weather patterns through a case study on weather data.

A Markov chain is often used to model the transition of real world events [24]–[27]. Each node in a Markov chain corresponds to a real world state. A Markov chain is viewed as the states in the model with probabilistic state transitions. Usually, the transition probabilities may be determined by learning or from domain expert [25]. Viewed as extensible variants of a static Markov chain, augmented Markov model [25], [26] and extensible Markov model [27] were proposed by allowing the number of the states grow with the input data.

In this paper, we propose a pattern detection data mining method based on an incremental Markov chain model. The proposed method, an adaptive Markov chain pattern detection (AMCPD) method, has a flexible structure for allowing the pattern grow and the transition probabilities of the Markov chain are adjusted adaptively. The proposed AMCPD method is applied to analyze the weather patterns of Singapore through meteorological data mining and different weather patterns in Singapore are disclosed. The results indicate that the early weather patterns disappear consistently across models and this suggests long-term climate changes. The proposed pattern detection algorithm will be of potential use for climatic and meteorological research as well as research focusing on pattern recognize or knowledge discovery in other research field.

II. Materials and Methods

A. Data collection

Meteorological data in the form of daily summary data were extracted from National Climate Data Center (NCDC), National Oceanic and Atmospheric Administration (NOAA) [28]. Climate variables studied in this paper include mean temperature, mean dew point temperature, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature and minimum temperature. Other variables were not included due to the presence of large amount of missing data in data sets such as mean sea level pressure, mean station pressure, maximum wind gust and precipitation amount, etc.

B. Proposed Method Description

It is noteworthy that the daily summary data is not continuously available for every day from 1962 to 2011. The total available data in each year are shown in Table I. In order to make the data appropriate for data mining, the meteorological data studied are pre-processed and converted to one vector for each day:

$$X_k = (\ x_k(1) \quad x_k(2) \quad ... \quad x_k(n)\ ) \tag{1}$$

where $x_k(1)$, $x_k(2)$, ...,$x_k(n)$ ($n = 7$ in this paper) represent the values of mean temperature, mean dew point temperature, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature and minimum temperature; $k(k = 1, 2, ..., N)$ is the input data index and N is the total number of available days ($N = 11499$).

In the previous pattern detection study [5], the high-dimensional data (number of dimensions>5000) was reduced to low-dimensional data (number of dimensions<15), and then the low-dimensional data was analyzed using a clustering algorithm. In this study, we propose a pattern detection method to detect patterns and outlying data from the meteorological data without dimension reduction since the dimensions available in this paper are less than 15.

A set of all possible clusters or states assumed by the process, $\{S_q, q = 1, 2, 3, ...\}$, is called the state space. A discrete-time random process forms a Markov chain if at any time the future behavior of the process depends only on the current state as the following property [27]:

$$\begin{aligned} &P\{S_{q+1} = j | S_q = i, S_{q-1} = i_{q-1}, ..., S_2 = i_2, S_1 = i_1\} \\ &= P\{S_{q+1} = j | S_q = i\} \end{aligned} \tag{2}$$

for all states $i_1, i_2, ..., i_{m-1}, i, j$, and all $q >= 1$.

In other words, the probability that the next state S_{q+1} is j, given the current state ($S_q = i$), and any past state ($S_1 = i_1, ..., S_{q-1} = i_{m-1}$), is dependent only upon the current state i.

As shown in Fig. 1, a node in Markov chain represents a weather pattern. In fact a node also represents a group or a cluster of weather data which belong to the same weather pattern. For instance, the first input data, which is daily weather information of a particular day, is placed as a center of a cluster. When more inputs are placed into this cluster, the cluster will include more weather data, and the centroid of the cluster will change over time.

The i^{th} cluster with $m - 1$ number of input data is represented as following:

$$C^{(i)}_{m-1} = \frac{\sum_{j=1}^{m-1} X_j}{m - 1} \tag{3}$$

where i represents the i^{th} cluster and $C^{(i)}_{m-1}$ is the centroid of the i^{th} cluster with $m - 1$ data in the cluster. $m \in [2, N]$. X_j is the j^{th} input data in the cluster.

When there is one more data added in this cluster, the updated centroid $C^{(i)}_m$ of the i^{th} cluster will be obtained by using the following:

TABLE I. THE NUMBER OF THE AVAILABLE DAILY SUMMARY DATA FROM 1962 TO 2011 (1^{st} JULY 2011)

Year	Number of the data	Year	Number of the data	Year	Number of the data	Year	Number of the data
1962	10	1963	96	1964	171	1965	274
1966	0	1967	0	1968	0	1969	0
1970	0	1971	0	1972	0	1973	0
1974	0	1975	0	1976	0	1977	0
1978	0	1979	0	1980	0	1981	181
1982	362	1983	365	1984	365	1985	365
1986	364	1987	365	1988	366	1989	365
1990	365	1991	365	1992	366	1993	365
1994	365	1995	365	1996	366	1997	365
1998	365	1999	365	2000	366	2001	365
2002	365	2003	364	2004	366	2005	365
2006	365	2007	364	2008	366	2009	365
2010	365	2011	182				

$$\begin{aligned} & C_m^{(i)} \\ = & \frac{\sum_{j=1}^{m} X_j}{m} \\ = & \frac{\sum_{j=1}^{m} X_j}{m-1}\cdot\frac{m-1}{m} \\ = & \frac{\sum_{j=1}^{m-1} X_j + X_m}{m-1}\cdot\frac{(m-1)}{m} \\ = & \left(\frac{\sum_{j=1}^{m-1} X_j}{m-1}+\frac{X_m}{m-1}\right)\cdot\frac{(m-1)}{m} \\ = & \left(C_{m-1}^{(i)}+\frac{X_m}{m-1}\right)\cdot\frac{(m-1)}{m} \\ = & \frac{(m-1)}{m}C_{m-1}^{(i)} + X_m/ni \end{aligned} \tag{4}$$

where $m \in [2, N]$.

The centroid $C_m^{(i)}$ of the i^{th} cluster is adaptive to the data that belongs to the cluster. This is the reason why we name it adaptive Markov chain model.

The main description of the proposed adaptive Markov chain model data mining method can be described by using the following parameters:

a. X, input data which is a set of observation symbols and in this paper represents daily summary data $\{x_k(1),\ x_k(2),\ ...,x_k(n)\}(n = 7$ in this paper), and each X is associated with at least one state.

b. S, a set of states and each has three attributes $< i, m_{i,k}, C_{i,k} >$:
   - i, the symbol that the state recognizes
   - $m_{i,k}$, the size of the cluster S_i at the moment when the k^{th} input data is the current input; in other words, the number of time steps that the system remains in S_i when k^{th} input data is current input.
   - $C_{i,k}$, the centroid of the S_i state when the k^{th} input data is the current input.

c. P, an $N \times N$ matrix represents state transition probabilities. Each element, p_{ij}, is the probability of a transition from state S_i to S_j, with $\sum_j{}^N p_{ij} = 1$.

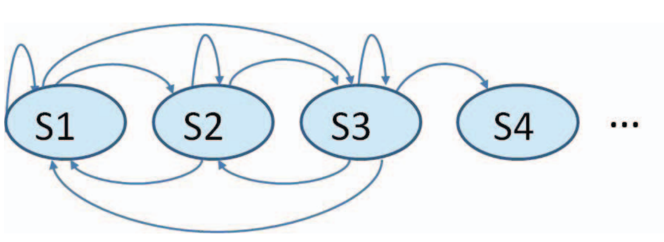

Fig. 1. A simple example of Markov chain incremental models

The main process of the proposed adaptive Markov chain model data mining method is as follows:

1) Initialize the system by creating an initial state, S_ϕ. This state provides a starting point for model construction. No transitions are ever made to S_ϕ. After initialization, X, and P are empty.
2) Designate the input X_k to the system and the current state as S_C
3) Calculate the similarity of X_k to the existing clusters, put it in a cluster according to the similarity obtained. Readjust all of the transition probabilities and update the whole system. If X_k is not similar to any existing cluster, a new state will be created, expand matrix P to account for the new states, and calculate the new transition probabilities, and the system will be updated.
4) Go to step 2, repeating until there are no more input data.

Interested readers are encouraged to refer to the research work [25]–[27] for more details about incremental Markov chain models.

The Euclidean distance [29], [30] is employed to calculate the distance between the two input data vectors $X_m = (\ x_m(1)\ \ x_m(2)\ \ ...\ \ x_m(n)\ )$ and $C^{(i)} = (\ c^{(i)}(1)\ \ c^{(i)}(2)\ \ ...\ \ c^{(i)}(n)\ )$ in n-dimensional space:

$$D(X_m, C^{(i)}) = \sqrt{\sum_{j=1}^{n}(x_m(j) - c^{(i)}(j))^2} \tag{5}$$

The state transition probability is calculated using the following:

$$p_{i,j} = n_{ij}/m_i \tag{6}$$

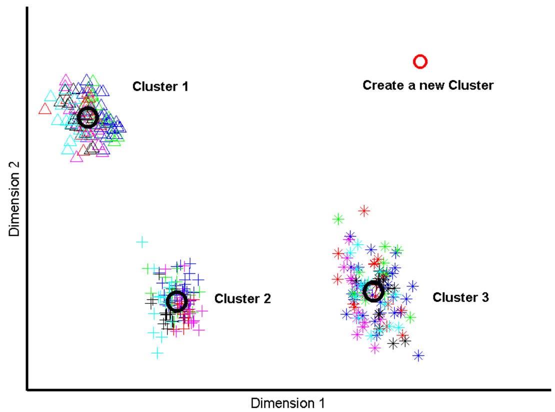

Fig. 2. Explanation of the principle of the proposed method by using 2-dimensional data as an example

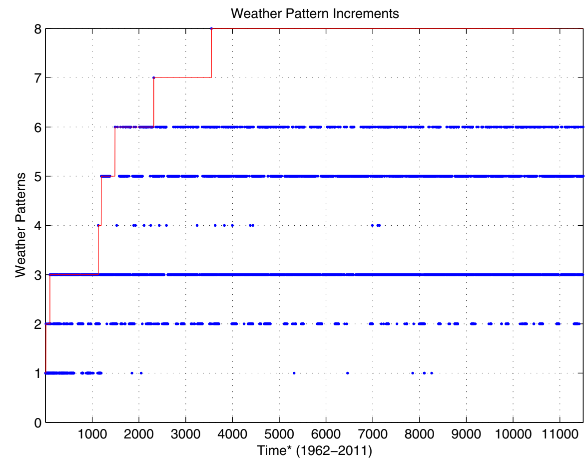

Fig. 3. Weather patterns obtained with a high cluster threshold (0.59) by using the normalized weather variables

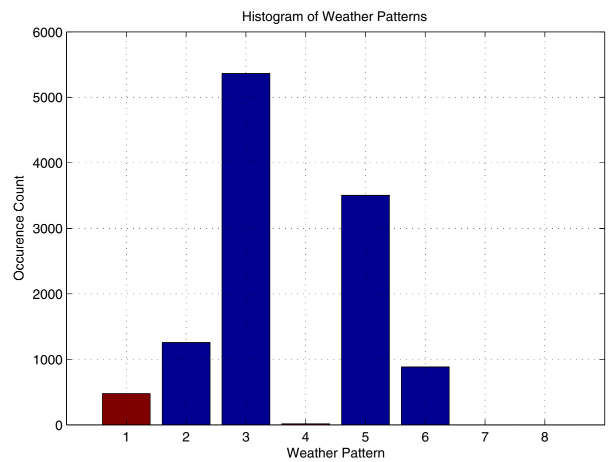

Fig. 4. Histogram of the 8 weather patterns shown in Fig. 3

where m_i represent the size of the cluster (node S_i); n_{ij} is the number of times the transition from S_i to S_j has occurred.

We describe the principal idea of the proposed method by using two-dimensional data and a simple example graph as shown in Fig. 2. If a new input is not sufficiently similar to any existing clusters, it will form a new node or cluster which represents a new pattern. Through the procedure of data processing, each input data is either placed in existing clusters or treated as a new pattern according to its similarity to existing clusters. A simple example is shown in Fig. 2 wherein the fourth cluster is treated as a new pattern. The advantage of the proposed detection method is that the similarity threshold or resolution can be tuned. More defined weather patterns can be discovered by setting a lower cluster threshold which represents a higher resolution.

III. Results and Discussions

In this paper, the original data is not used directly as the scales and units of the measurements of the various meteorological variables are different. Normalization is used in this paper to ensure that all meteorological variables are given equal levels of importance. In addition, by using the normalized values, different cluster threshold settings (0.59, 0.5 and 0.3) are used in the delineation of clusters/patterns for comparative purposes. These three different scenarios (threshold settings 0.59, 0.5 and 0.3) are specified in the application of our proposed approach for detecting weather patterns. This allows for a comparison between results obtained using different threshold variations. Comparing the results is equivalent to a sensitivity analysis that is able to indicate the robustness of the algorithm.

A. Scenario 1: Model with a high cluster threshold

It is evident from Fig. 3, that the first input data is considered as the first weather pattern. There are 10 sets of input data that were made available in 1962 from records of the Changi meteorological station, each set representing a specific day. All the 10 sets of input data are classified into the first weather pattern. Table II shows the details of the weather patterns and the time of the formation of the weather patterns. According to Table II and Fig. 3, the 11th input data (1963-01-23) is classified as the second weather pattern. The second weather pattern is considered as a common pattern as it is constantly present till 2011 (See Fig. 3), and a larger number of data sets is placed in this weather pattern, as can be seen in Fig. 4.

According to Fig. 3, the early weather pattern gradually disappears. The changes in temporal weather patterns indicate long-term changes to the climate of Singapore which may be attributed in part to global climate change on a larger scale [7], [9], [10], [16]. It can be seen in Fig. 3 that the third, fifth and the sixth weather patterns are also considered as normal weather patterns because these three weather patterns are constantly present till 2011. This can be seen in Fig. 4, which shows that many of the input data sets are classified into these four weather patterns (i.e., the second, the third, fifth and the sixth weather patterns).

Further it is observed that the fourth weather pattern

TABLE II. THE DETAILS ON THE INPUT DATA INDEX AND THE TIME OF FORMING NEW WEATHER PATTERNS OF THE FIG. 3 & 4

No.	Data index	YEARMODA	TEMP	DEWP	VISIB	WDSP	MXSPD	MAX	MIN
1	1	19620512	81	75	19.13	3.7	12	86	75
2	11	19630123	77.5	73.5	8.4	13.5	15	79	77
3	97	19631201	78	74	20	0	4.27	81	73
4	1131	19830206	80.6	72.8	7.1	6.3	42.9	88.7	74.5*
5	1195	19830411	86.7	78.9	2.6	5.4	11.8	93.2	81*
6	1488	19840129	72.8	71.6	5.7	2.2	8	75.2	71.6
7	2317	19860507	84.3	76.9	7.3	13.3	49.9	89.1	78.8*
8	3550	19890922	76.8	74.7	5.4	1.9	11.1	86	59.2

Note: YEARMODA = timestamp; TEMP = mean temperature; DEWP = mean dew point temperature; VISIB = mean visibility; WDSP = mean wind speed; MXSPD = maximum sustained wind speed; MAX = maximum temperature; MIN = minimum temperature;* represents El Nino occurrence and ** represents La Nina occurrence

appeared at input data index 1131 (1983-02-06), while the seventh weather pattern appeared at input data index 2317 (1986-05-07), and the eighth weather pattern appeared at input data index 3550 (1989-09-22) respectively. These three weather patterns were detected as rare events by the algorithm and there is less data classified into each of these weather patterns. Evidently, the seventh and eighth weather patterns are deemed as rare events as they contained less than four data points each. Some of these rare events coincided with ENSO phenomena [12]–[14] such as the El Nino and La Nina events (weather patterns 4 and 7) or may have formed the initial part of a new pattern (e.g. weather pattern 5) as is evident in Tables II, III and IV(indicated with asterisks).

B. Scenario 2: Model with a lower cluster threshold than the Model in Scenario 1

It can be seen in Fig. 5 that the first input data was also treated as the first weather pattern, and all the 10 sets of the input data formed the first weather pattern. Table III shows the input data index and the formation times of the new weather patterns. According to Table III, Fig. 5 and Fig. 6, it can be seen that more detailed weather patterns were detected when using models with higher resolutions. The general trends were consistent with the results obtained by using models with lower resolutions. Both models detected the emergence of weather patterns at input data index 1131 (1983-02-06), input data index 2317 (1986-05-07), and input data index 3550 (1989-09-22). These three weather patterns were also rare events detected by the algorithm since too few input data sets were classified into these three weather patterns. Similar to the results in Scenario 1, the first weather pattern disappeared after a period of time. The weather patterns in the long term are consistent with global climate change on weather patterns [7], [9], [16]. As with the previous scenario, several of the new patterns coincided with ENSO events [12]–[14] according to Table III (indicated with asterisks).

C. Scenario 3: Model with a lower cluster threshold than Scenario 2

According to Figs. 7, 8 and Table IV, more details about weather patterns are obtained when a model with an even higher resolution is used. The results are consistent with the previous results obtained by using models with lower resolutions where new weather patterns appeared at input data

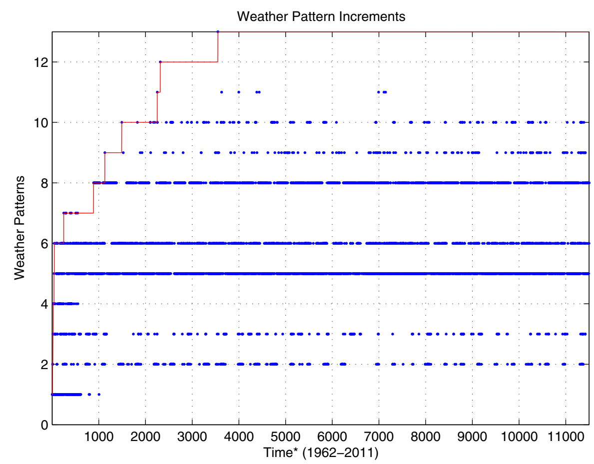

Fig. 5. Weather patterns obtained with a lower cluster threshold (0.5) by using the normalized weather variables

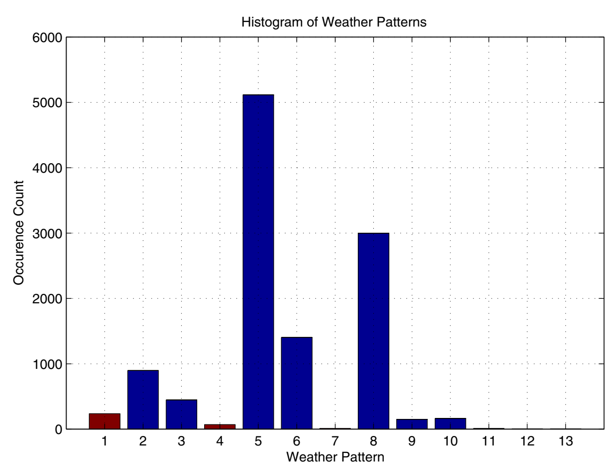

Fig. 6. Histogram of the 13 weather patterns shown in Fig. 5

TABLE III. THE DETAILS ABOUT THE INPUT DATA INDEX AND THE TIME OF FORMING NEW WEATHER PATTERNS AS SHOWN IN FIG. 5 & 6

No.	Data index	YEARMODA	TEMP	DEWP	VISIB	WDSP	MXSPD	MAX	MIN
1	1	19620512	81	75	19.13	3.7	12	86	75
2	11	19630123	77.5	73.5	8.4	13.5	15	79	77
3	18	19630316	82.2	72.6	19.1	13.2	20	90	75
4	25	19630329	86.7	73	25.6	9	13	97	77
5	41	19630604	80	75	17.7	0	18.45	91	79
6	49	19630620	76.5	73.3	12.4	4.8	7	81	73
7	251	19641114	79.7	69	23.9	6.5	8.9	93	70**
8	889	19820609	87	78.3	7.5	7.1	9.9	91.4	82.6*
9	1131	19830206	80.6	72.8	7.1	6.3	42.9	88.7	74.5*
10	1492	19840202	73.1	71	6.2	4.1	6	75.2	69.8
11	2252	19860303	78.3	73.2	6.3	12.9	34	86	73.4
12	2317	19860507	84.3	76.9	7.3	13.3	49.9	89.1	78.8*
13	3550	19890922	76.8	74.7	5.4	1.9	11.1	86	59.2

Note: YEARMODA = timestamp; TEMP = mean temperature; DEWP = mean dew point temperature; VISIB = mean visibility; WDSP = mean wind speed; MXSPD = maximum sustained wind speed; MAX = maximum temperature; MIN = minimum temperature;* represents El Nino occurrence and ** represents La Nina occurrence

TABLE IV. THE DETAILS ABOUT THE INPUT DATA INDEX AND THE TIME OF FORMING NEW WEATHER PATTERNS AS SHOWN IN FIG. 7 & 8

No.	Data index	YEARMODA	No.	Data index	YEARMODA	No.	Data index	YEARMODA
1	1	19620512	26	176	19640629**	51	1135	19830210*
2	2	19620516	27	234	19641020**	52	1203	19830419*
3	3	19620605	28	235	19641021**	53	1249	19830604*
4	6	19621012	29	251	19641114**	54	1488	19840129
5	11	19630123	30	271	19641221	55	1660	19840719
6	14	19630228	31	280	19650104*	56	2252	19860303
7	15	19630301	32	281	19650113*	57	2317	19860507*
8	18	19630316	33	346	19650429*	58	2419	19860817*
9	24	19630328	34	368	19650524*	59	2672	19870428*
10	25	19630329	35	374	19650530*	60	3243	19881119**
11	40	19630603	36	491	19651014*	61	3550	19890922
12	41	19630604	37	495	19651019*	62	3631	19891212
13	42	19630607	38	514	19651108*	63	3793	19900523
14	74	19630914	39	555	19810706	64	4255	19910828*
15	80	19630929	40	556	19810707	65	4356	19911207*
16	83	19631015	41	598	19810819	66	4648	19920924
17	87	19631108	42	613	19810903	67	6244	19970206*
18	89	19631110	43	638	19810928	68	6719	19980527*
19	97	19631201	44	699	19811128	69	6782	19980729**
20	102	19631209	45	700	19811129	70	6833	19980918**
21	105	19631222	46	750	19820118	71	6997	19990302**
22	122	19640320**	47	889	19820609*	72	8331	20021026*
23	139	19640421**	48	916	19820706*	73	9680	20060707*
24	153	19640526**	49	1050	19821117*			
25	155	19640529**	50	1131	19830206*			

Note: YEARMODA = timestamp; TEMP = mean temperature; DEWP = mean dew point temperature; VISIB = mean visibility; WDSP = mean wind speed; MXSPD = maximum sustained wind speed; MAX = maximum temperature; MIN = minimum temperature;* represents El Nino occurrence and ** represents La Nina occurrence

index 1131 (1983-02-06), at input data index 2317 (1986-05-07), and at input data index 3550 (1989-09-22), respectively. These three weather patterns were also detected as rare events by the algorithm and less input data was classified into these three weather patterns as is evident in Fig. 8. Similar to the previous results obtained in Scenario 1 and Scenario 2, the early weather patterns disappeared after a period of time. In fact, more than one weather pattern disappeared, according to Fig. 7 and Table IV. Compared with the results obtained previously by the models in Scenario 1 and Scenario 2, it is observed that the initial weather patterns obtained with lower resolutions were at this stage divided into four different weather patterns owing to the use of models with higher resolutions. Such increases in details of weather patterns may plausibly be the result of the marked differences between wet and dry days in the same season. Once again it is obvious from Table IV that a considerable number of new patterns coincided with ENSO events [13], [14].

IV. CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

In this study, we present an adaptive Markov chain pattern detection method for disclosing weather pattern in Singapore. The proposed method is capable of simultaneously dealing

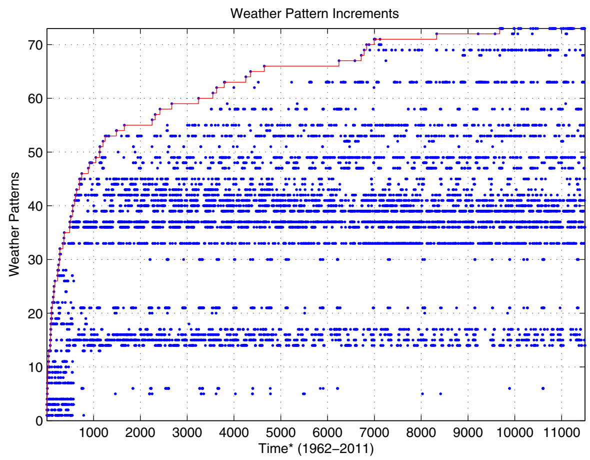

Fig. 7. Weather patterns obtained with a high cluster threshold (0.3) by using the normalized weather variables

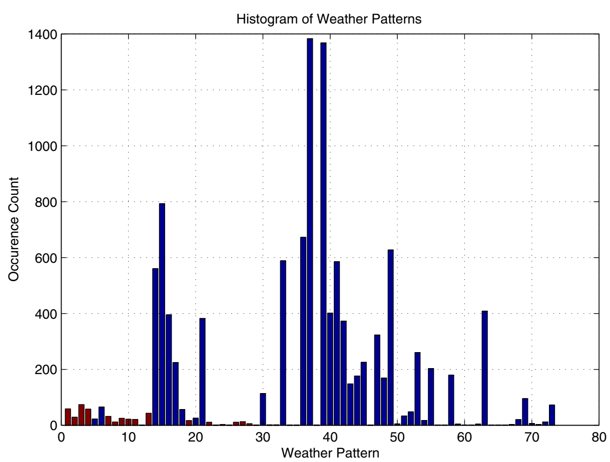

Fig. 8. Histogram of the 73 weather patterns shown in Fig. 7

with all meteorological or climate variables to detect hidden weather patterns. The results reveal the different weather patterns existing in Singapore from 1962 to 2011. The proposed method is enabled to (i) identify weather patterns in the long term while consistent with global climate change on weather patterns, (ii) identify rare/outlying patterns that coincide with ENSO events, and (iii) detect weather patterns in greater detail by adjusting the resolution of the proposed detection model.

Despite its promising results, there is still space for improvement. This paper focused on the seven meteorological variables (mean temperature, mean dew point temperature, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature and minimum temperature), but other climate variables, such as precipitation and sea level pressure should affect the classification of weather patterns too. Therefore, further studies to incorporate more variables are planned in the future. Second, in our study, all meteorological variables are given equal levels of importance. Different significance of different meteorological variables should be considered accordingly in the future works. More meteorological data from different weather stations can be analyzed in future studies for further investigation on climate patterns over a spatial dimension.

In addition to the analysis of meteorological or weather data, extension of the proposed strategy for other applications, such as identifying influence patterns and sentiment patterns in social networks, is also planned in our future work.

ACKNOWLEDGMENT

The authors would like to thank Dr. Yang Yinping from Institute of High Performance Computing (IHPC) for her advice and essential help in conducting this study. The authors would also like to acknowledge the support from the Strategic Social System (SSS) program of IHPC and Joint Council Office (JCO) project at the Agency for Science Technology and Research (A*STAR).

REFERENCES

[1] UNEP, "United Nations Environment Programme, Climate Change," *Available from:http://www.unep.org/climatechange; [cited 8 March 2013]*.

[2] Y. Meng, M. Dunham, F. Marchetti, and J. Huang, "Rare event detection in a spatiotemporal environment," *Proceedings of the Second IEEE International Conference on Granular Computing (GrC'06)*, pp. 10–12, 2006.

[3] Z. Yang, N. Meratnia, and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[4] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification," *The Computer Journal*, vol. 54, no. 4, pp. 570–588, Mar. 2011.

[5] Z. Wang, C. S. Chang, and Y. Zhang, "A feature based frequency domain analysis algorithm for fault detection of induction motors," *Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on*, pp. 27–32, 2011.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[7] R. M. Li, *Statistical Analysis of the Spatio-Temporal Variability of the Urban Heat Island in Singapore, Honours Thesis.* Department of Geography, National University of Singapore., 2009.

[8] W. Chow and M. Roth, "Temporal dynamics of the urban heat island of Singapore," *International Journal of Climatology*, vol. 26, no. 15, pp. 2243–2260, 2006.

[9] K. L. Ebi, N. D. Lewis, and C. Corvalan, "Climate Variability and Change and their Potential Health Effects in Small Island States: Information for Adaptation Planning in the Health Sector," *Environmental Health Perspectives*, vol. 114, no. 12, pp. 1957–1963, 2006.

[10] NEA, "National Environmental Agency,Weather Wise Singapore," *Available from: http://www.app2.NEA.gov.sg/data/cmsresource/2; [cited 29 July 2011]*.

[11] IPCC, "Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change," *Cambridge University Press*, 2007.

[12] J. Harger, "Air-temperature variations and ENSO effects in Indonesia, the Philippines and El Salvador. ENSO patterns and changes from 18661993," *Atmospheric Environment*, vol. 29, no. 16, pp. 1919–1942, Aug. 1995.

[13] M. Cui, J. Mo, and F. Qiao, "Ei nino phenomenon and extended associate pattern analysis," *Journal of Hydrodynamics Series B-English Edition*, vol. 16, no. 1, pp. 90–100, 2004.

[14] NOAA, "El Nino and La Nina, Climate Prediction Center," *Available from: http://www.cpc.ncep.noaa.gov/products; [cited 5 July 2011]*.

[15] G. Stern, R. W. Macdonald, P. M. Outridge, S. Wilson, J. Chételat, A. Cole, H. Hintelmann, L. L. Loseto, A. Steffen, F. Wang, and C. Zdanowicz, "How does climate change influence Arctic mercury?" *Science of the total environment*, vol. 414, pp. 22–42, Jan. 2012.

[16] S. Kovats, K. Ebi, B. Menne, D. Campbell-Lendrum, and U. N. E. Programme, *Methods of assessing human health vulnerability and public health adaptation to climate change*. Regional Office for Europe, World Health Organization, 2003, vol. 1, no. 1.

[17] F. Giorgi and E. Coppola, "Does the model regional bias affect the projected regional climate change? An analysis of global model projections," *Climatic Change*, vol. 100, no. 3-4, pp. 787–795, May 2010.

[18] K. S. Jun, E.-S. Chung, J.-Y. Sung, and K. S. Lee, "Development of spatial water resources vulnerability index considering climate change impacts." *Science of the total environment*, vol. 409, no. 24, pp. 5228–42, Nov. 2011.

[19] A. Augustsson, M. Filipsson, T. Oberg, and B. Bergbäck, "Climate change - An uncertainty factor in risk analysis of contaminated land." *Science of the total environment*, vol. 409, no. 22, pp. 4693–700, Oct. 2011.

[20] S. K. Oni, M. N. Futter, L. a. Molot, and P. J. Dillon, "Modelling the long term impact of climate change on the carbon budget of Lake Simcoe, Ontario using INCA-C." *Science of the total environment*, vol. 414, pp. 387–403, Jan. 2012.

[21] G. Meehl, T. Stocker, W. Collins, A. Friedlingstein, A. Gaye, J. Gregory, A. Kitoh, R. Knutti, J. Murphy, A. Noda, and Others, "Global climate projections," *Climate change*, pp. 747–845, 2007.

[22] M. E. Stuart, D. C. Gooddy, J. P. Bloomfield, and a. T. Williams, "A review of the impact of climate change on future nitrate concentrations in groundwater of the UK." *Science of the total environment*, vol. 409, no. 15, pp. 2859–73, Jul. 2011.

[23] C. Tisseuil, M. Vrac, G. Grenouillet, a. J. Wade, M. Gevrey, T. Oberdorff, J.-B. Grodwohl, and S. Lek, "Strengthening the link between climate, hydrological and species distribution modeling to assess the impacts of climate change on freshwater biodiversity." *Science of the total environment*, vol. 424, pp. 193–201, Mar. 2012.

[24] N. Ye, S. Member, Y. Zhang, and C. M. Borror, "Robustness of the Markov-Chain Model for Cyber-Attack Detection US Air Force Office of Scientific Research," *Engineering*, vol. 53, no. 1, pp. 116–123, 2004.

[25] D. Goldberg, "Maximizing Reward in a Non-Stationary Mobile Robot Environment," *Autonomous Agents and Multi-Agent Systems*, pp. 287–316, 2003.

[26] C. Caragea, D. Caragea, A. Silvescu, and V. Honavar, "Semi-supervised prediction of protein subcellular localization using abstraction augmented Markov models." *BMC bioinformatics*, vol. 11, no. Suppl 8, p. S6, Jan. 2010.

[27] R. M. Kotamarti, M. Hahsler, D. Raiford, M. McGee, and M. H. Dunham, "Analyzing taxonomic classification using extensible Markov models." *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. 2235–41, Sep. 2010.

[28] WMO, "NNDC Climate Data Online, National Climatic Data Center, NESDIS, NOAA ," *Available from: http://www7.ncdc.noaa.gov/CDO/dataproduct; [cited 5 July 2011].*

[29] K. J. Swanepoel, "Unit Distances and Diameters in Euclidean Spaces," *Discrete & Computational Geometry*, vol. 41, no. 1, pp. 1–27, May 2008.

[30] J. Tykesson, "On the visibility in well-behaved random sets in Euclidean space," *Statistics & Probability Letters*, vol. 81, no. 8, pp. 1256–1259, Aug. 2011.