Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 840 – 857

doi:10.3906/elk-1210-20

Discovery of hydrometeorological patterns

Mete ÇELİK[1,*], Filiz DADAŞER-ÇELİK[2], Ahmet Şakir DOKUZ[3]
[1]Department of Computer Engineering, Erciyes University, Kayseri, Turkey
[2]Department of Environmental Engineering, Erciyes University, Kayseri, Turkey
[3]Department of Computer Engineering, Niğde University, Niğde, Turkey

Received: 04.10.2012 • **Accepted:** 09.12.2012 • **Published Online:** 17.06.2014 • **Printed:** 16.07.2014

Abstract: Hydrometeorological patterns can be defined as meaningful and nontrivial associations between hydrological and meteorological parameters over a region. Discovering hydrometeorological patterns is important for many applications, including forecasting hydrometeorological hazards (floods and droughts), predicting the hydrological responses of ungauged basins, and filling in missing hydrological or meteorological records. However, discovering these patterns is challenging due to the special characteristics of hydrological and meteorological data, and is computationally complex due to the archival history of the datasets. Moreover, defining monotonic interest measures to quantify these patterns is difficult. In this study, we propose a new monotonic interest measure, called the hydrometeorological prevalence index, and a novel algorithm for mining hydrometeorological patterns (HMP-Miner) out of large hydrological and meteorological datasets. Experimental evaluations using real datasets show that our proposed algorithm outperforms the naïve alternative in discovering hydrometeorological patterns efficiently.

Key words: Data mining, hydrometeorological pattern, association rule mining, hydrological databases, meteorological databases

1. Introduction

Data mining is the process of discovering previously unknown and potentially useful information from large datasets. It offers semiautomatic or automatic techniques to analyze large and multidimensional datasets that are difficult to interpret by analysts. Data mining mainly deals with the problems of clustering, classification, anomaly detection, and association analysis [1,2]. The focus of this study is to develop association analysis techniques to mine large and multidimensional hydrological and meteorological datasets efficiently.

Hydrometeorological patterns can be defined as meaningful and nontrivial associations between hydrological and meteorological parameters over a region. They can also be interpreted as the patterns that represent the cause-effect relationship between hydrological and meteorological parameters and are present at significant number of locations (stations and grids) over a region. In this study, we focus on discovering hydrometeorological patterns that reveal the effect of meteorological parameters on hydrological parameters at a sufficient number of stations over a region. This analysis will help us identify the spatial distribution of the rules (in other words, generalize the rules) generated for individual stations.

In association analysis, the aim is to discover any rules of the form $X \rightarrow Y$ that seem to occur in data with a frequency above a given threshold. Here, X and Y are events of a certain type, connected by the rule

*Correspondence: mcelik@erciyes.edu.tr

'if X occurs, then Y occurs'. The rules can be extended into the form $X_1, X_2, \ldots, X_H \rightarrow Y$, which can be interpreted as 'if $X_1, X_2, \ldots, X_H$ all occur, then Y will occur'. An example hydrometeorological pattern can be formed using precipitation and stream flow. Let X be precipitation of certain magnitude and Y be stream flow of certain magnitude. An association between these 2 variables in the form $X \rightarrow Y$ can be read as 'if precipitation at certain magnitude occurs, then stream flow at certain magnitude occurs'. If the relationship between X and Y is present at a significant number of locations over a region, then $X \rightarrow Y$ can be called a hydrometeorological pattern.

Discovering hydrometeorological patterns is important for several applications that affect our everyday life. For example, once defined, hydrometeorological patterns can be used to forecast natural hazards such as floods and droughts and help develop emergency preparedness or early warning plans. Another use of hydrometeorological patterns could be estimating hydrologic responses of ungauged basins, which is important for determining water availability and developing sustainable water management practices. Filling in missing hydrological or meteorological records could be another application.

However, it is challenging to mine hydrometeorological parameters due to the specific characteristics of hydrological and meteorological datasets and computational issues [3,4]. First, the hydrological and meteorological data are geographical data and include spatial and temporal correlations. Second, they have nonlinear dependencies, a long memory in time, and teleconnections in space. Third, the linkages between the hydrological and meteorological parameters are based on complex physical processes that are difficult to model. Fourth, discovering patterns from large hydrometeorological datasets is computationally expensive due to the archival history of the datasets. Fifth, developing monotonic interest measures to quantify hydrometeorological patterns is challenging. The aim of this study is to develop computationally efficient techniques for discovering the hydrometeorological patterns between hydrological (i.e. stream flow) and meteorological (i.e. precipitation, air temperature, wind speed, and relative humidity) parameters out of large datasets.

This study defines hydrometeorological patterns, proposes a new composite interest measure to quantify these patterns, and a novel and computationally efficient hydrometeorological pattern mining algorithm (HMP-Miner) to mine large hydrological and meteorological datasets.

The rest of this paper is organized as follows. Section 2 presents related works and Section 3 provides the basic concepts related to hydrometeorological pattern mining and presents the problem of hydrometeorological pattern mining. Section 4 discusses the naïve approach and the proposed HMP-Miner algorithm. Section 5 presents the experimental evaluation. Section 6 presents the evaluation of the results and the final section presents conclusions and future works.

2. Related works

Many studies are available in the literature related to data mining methods, including association analysis [2,5,6].

Association rule mining has been used in a broad range of application domains including health [7], medicine [8], and business [9,10]. The use of data mining, particularly association analysis, in environmental research, however, is very limited. Tadesse et al. determined association rules between climatic and oceanic variables to analyze drought in Nebraska [11,12]. Lin et al. used association analysis for discovering the relationships between surface precipitation and sea surface temperatures [13]. Tan et al. used association analysis to find interesting spatiotemporal patterns in earth science data [14]. Dhanya and Nagesh Kumar [15] and Nagesh Kumar et al. [16] analyzed rainfall data to discover association rules for droughts and floods

in India. Shu et al. studied fuzzy association rules in climatological datasets [17]. Dadaser-Celik et al. [18] investigated the associations between stream flow and climatic variables at 3 locations at the Kızılırmak River Basin in Turkey. However, these studies that examined the associations in environmental data focused on the analysis of patterns at individual stations. This prevents understanding of 'spatial relationships of rules' or 'spatial pattern' over certain regions [11].

In contrast, in this study, we focus on determining hydrometeorological patterns over multiple stations by defining a new interest measure and developing a new algorithm to explore spatial relationships between hydrological and meteorological variables computationally.

3. Basic concepts and problem definition

This section introduces basic concepts related to this study. First, definitions related to the association analysis are given. Second, the proposed hydrometeorological prevalence index is discussed. Finally, a formal definition of the 'hydrometeorological pattern mining problem' is presented.

3.1. Basic concepts

For completeness, first we present the definitions related to the classical association rule mining. Next, we define 2 interest measures, the station prevalence index and hydrometeorological prevalence index, for quantifying hydrometeorological patterns that are present at a significant number of locations (stations and grids) over a region.

Definition 1 *An association rule is defined as follows: Let* $I = \{i_1, i_2, \ldots, i_n\}$ *be a set of n binary attributes called items. Let* $D = \{t_1, t_2, \ldots, t_m\}$ *be a set of records called the database. Each record in D has a unique record ID and contains a subset of the items in I. A rule is defined as an implication of the form* $X \rightarrow Y$*, where* $X, Y \subseteq I$ *and* $X \cap Y = \emptyset$*. The sets of items (for short itemsets) X and Y are called the antecedent and consequent of the rule, respectively. A pattern P is defined as set of* $\{X, Y\}$ *[20].*

Definition 2 *Given a pattern P and a dataset* D_i *of a station* S_i*, the support of pattern P in station* S_i *is the fraction of the number of records containing P to the total number of records of dataset* D_i *of the station* S_i *[1,2,19]. Support of pattern P at station* S_i *can be formulized as follows:*

support (P, S_i*) = (number of records containing P) / (number of all records of station* S_i*).*

Definition 3 *Pattern P is frequent (support prevalent) if its support is equal to or greater than a user-defined support threshold, min_support.*

Definition 4 *Given a pattern P and a dataset* $D = \{D_1, D_2, \ldots, D_n\}$*, the station prevalence of the pattern P is the fraction of stations in which the pattern P exists to the total number of stations S in the dataset (in the region) D.*

station_prev(P, D)= (number of stations S_i *containing P) / (total number of stations S)*

Definition 5 *Given a pattern P and a dataset D, a hydrometeorological pattern can be defined as a subset of itemsets of hydrological and meteorological parameters, in which at least 1 meteorological and 1 hydrological parameter occur in the antecedent and/or consequent part of the rule.*

The focus of this paper is to reveal the effect of meteorological parameters on hydrological parameters. Hence, in this study, the antecedent part of the hydrometeorological rule contains meteorological parameters and the consequent part of the rule contains hydrological parameters (such as 'if precipitation at certain magnitude occurs, then stream flow at certain magnitude occurs').

Definition 6 *Given a pattern P, stations $S = \{S_1, S_2, \ldots, S_n\}$ and a dataset $D = \{D_1, D_2, \ldots, D_n\}$ belonging to the stations, the hydrometeorological prevalence index of pattern P is the composition of support and station prevalence measures formulized as below:*

$station\_prev_{Si \in S}$ (support(P, S_i) $\geq$ min_support).

The hydrometeorological prevalence index only counts the stations in which pattern P satisfies the min_support threshold.

Definition 7 *Given a pattern P, the support and station prevalence threshold values of min_support and min_station, respectively, the pattern P is hydrometeorological prevalent, if it satisfies the user-defined support and station prevalence thresholds.*

$station\_prev_{Si \in S}$ (support(P_i, S_i) $\geq$ min_support) $\geq$ min_station

Definition 8 *Given a rule $X \rightarrow Y$ (where $P = \{X, Y\}$), the confidence of rule $X \rightarrow Y$ determines how frequently the parameters in Y appear in the records that contain X parameters and is formulized as follows [1,2,19]:*

confidence ($X \rightarrow Y$) = support ($X \cup Y$) / support (X)

Confidence is used to find probability $P(Y|X)$. In our case, Y is a hydrological parameter and X is a set of meteorological parameters.

Definition 9 *A rule is called meaningful if its confidence value satisfies the minimum confidence threshold min_conf.*

3.2. Problem definition

Given:

- Hydrometeorological database, D.
- Minimum support threshold, *min_support.*
- Minimum station prevalence threshold, *min_station.*
- Minimum confidence threshold *min_conf.*

Output:

- Hydrometeorological patterns (rules) that satisfy the *min_support*, *min_station*, and *min_conf* thresholds.

Objective:

- Minimization of the computational cost.

Constraints:

- Find correct and complete hydrometeorological rules with given parameters.
- Consequent part of the rule should include hydrological parameter.

4. Hydrometeorological pattern mining

In this study, the aim is to find the associations between hydrological and meteorological parameters. In other words, we would like to discover rules that explain the effect of meteorological parameters on hydrological parameters over a region that includes more than one station. That is, the antecedent part of the rule will contain meteorological parameters and the consequent part will contain a hydrological parameter.

To mine hydrometeorological patterns, we propose 2 algorithms, the naïve approach and HMP-Miner (Figures 1a and 1b).

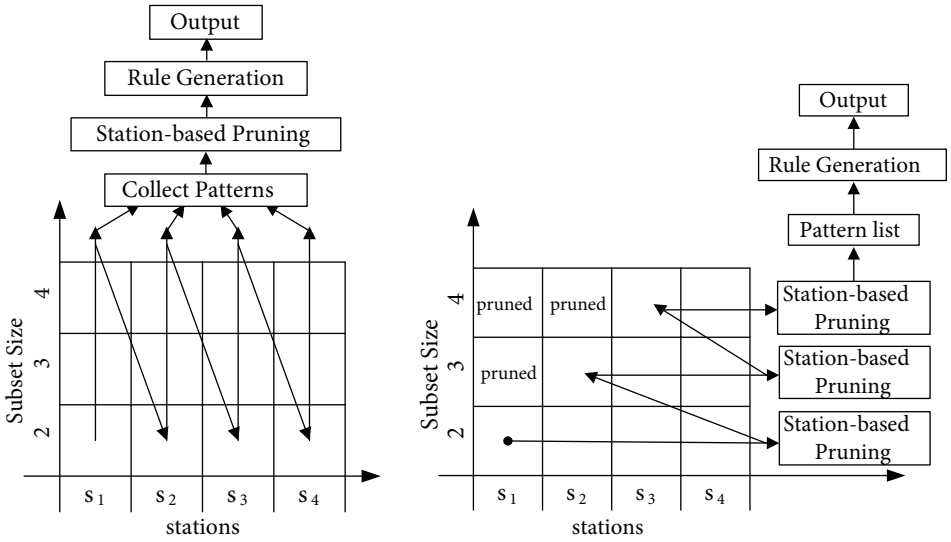

Figure 1. a) Naïve approach. b) HMP-Miner.

4.1. Naïve approach

The naïve approach (Figure 1a) is an extended version of the Apriori algorithm [19] to handle multiple stations to mine hydrometeorological patterns. It first discovers all size-frequent patterns for each station in a region and applies a postprocessing step to prune nonprevalent hydrometeorological patterns that do not occur at a sufficient number of stations. After pruning the nonprevalent hydrometeorological patterns, meaningful ones are discovered using the minimum confidence threshold *min_conf*. The limitation of this algorithm is the generation of the nonprevalent hydrometeorological candidates before the postpruning stage. However, if a size k pattern is not hydrometeorologically prevalent at a sufficient number of stations, that pattern should not be used to generate size $k + 1$ candidates. The unnecessary generation of nonprevalent patterns increases the complexity of the algorithm.

Algorithm 1 gives the pseudocode of the naïve approach.

Algorithm 1. Pseudocode of the naïve approach

Inputs:

min_support: Minimum support threshold

min_station: Minimum station prevalence index threshold

min_conf: Minimum confidence threshold

D: Hydro meteorological dataset

Output: Frequent and meaningful hydrometeorological patterns that satisfy the thresholds of *min_support*, *min_station*, and *min_conf*.

Variables:

NS: number of stations

C_k: set of size *k* candidates

L_k: set of support-prevalent size *k* patterns

TP: set of candidate hydrometeorological pattern

HM: set of hydrometeorological pattern

HM_patterns: set of meaningful hydrometeorological patterns

Algorithm:

```
1.  initialization; k = 0, L_0 = D
2.  for (s = 1 to NS) {
3.                  while (not empty L_k) {
4.                          If k = 0 then
5.                                  C_{k+1}(s) = get_singletons (D(s))
6.                          else
7.                                  C_{k+1}(s) = generate_candidates (L_k(s))
8.                          C_{k+1}(s) = calculate_supports (C_{k+1}(s), D(s))
9.                          L_{k+1}(s) = prune_support_non_prevalent (C_{k+1}(s), min_support)
10.                         k = k + 1
11.                 }
12. }
13. TP = calculate_station_support (L)
14. HM = prune_station_non_prevalent (TP, min_station)
15. HM_patterns = discover_meaningful_patterns (HM, min_conf)
16. return HM_patterns
```

In between Steps 2 and 12, the naïve approach generates hydrometeorological patterns for each station; in Steps 13 and 14, prevalent hydrometeorological patterns are discovered; and in Step 15, hydrometeorological rules are generated. The functions used in the algorithm are explained below.

Getting singletons (Step 5): This function takes the dataset of a station s as input and outputs size *1* (singletons) candidates of that station. Size *1* candidates are the features that exist in that station s.

Generating candidates (Step 7): In this step, size $k + 1$ candidate patterns are generated using prevalent size k patterns L_k. The *for* loop between Steps 2 and 12 is run for each station to generate support-prevalent hydrometeorological patterns.

Calculating supports of patterns (Step 8): This function calculates the support values of candidate patterns C. The inputs of the function are candidate patterns and the dataset of station s. The output of the function is the support values of the candidate patterns.

Pruning support nonprevalent candidates (Step 9): Patterns that do not satisfy the minimum support threshold, *min_support*, are pruned by these functions.

Calculating station supports of patterns (Step 13): In this step, the station support values of each hydrometeorological pattern are calculated.

Pruning station nonprevalent patterns (Step 14): In this step, the patterns that do not satisfy the *min_station* threshold are pruned.

Discovering meaningful patterns (Step 15): In this step, meaningful hydrometeorological patterns that satisfy the minimum confidence threshold *min_conf* are discovered. This function first generates rules for each time slot. In this study, the rules of interest are those that contain a hydrological parameter in the consequent part of the rule. After the rule generation, the ones that are not satisfying the minimum confidence threshold *min_conf* are pruned. Next, the station prevalences of the rules are rechecked to see if they still satisfy the *min_station* threshold after pruning based on the minimum confidence threshold. If the rules do not satisfy the *min_station* threshold, these hydrometeorological rules are pruned.

Finally, frequent and meaningful patterns are returned by the algorithm.

The limitation of the naïve approach is the unnecessary generation of station nonprevalent hydrometeorological patterns before the postprocessing step.

4.2. HMP-Miner algorithm

To overcome this limitation of the naïve approach, we propose HMP-Miner (Algorithm 2) by applying station nonprevalent pruning after each size k generation in all stations. In contrast to the naïve approach, HMP-Miner does not wait until the postprocessing step to discover patterns.

The strategy of HMP-Miner is to apply station prevalence pruning as early as possible. The algorithm first generates size k frequent patterns for all stations and then eliminates size k station nonprevalent hydrometeorological patterns by applying station prevalence pruning. The size k prevalent hydrometeorological-prevalent patterns are then used to generate candidate size $k + 1$ patterns for all stations. This process is applied until all size-prevalent hydrometeorological patterns are discovered. With this algorithm, we discard the generation of unnecessary station nonprevalent hydrometeorological candidates that increase the cost of the mining hydrometeorological patterns. The pseudocode of HMP-Miner is given in Algorithm 2. The explanations of the functions of the algorithm are as given in Section 4.1.

Algorithm 2. The pseudocode of HMP-Miner

Inputs:

min_support: Minimum support threshold

min_station: Hydrometeorological prevalence index threshold

min_conf: Minimum confidence threshold

D: Hydro meteorological dataset

Output: Frequent and meaningful hydrometeorological patterns that satisfy the thresholds of *min_support*, *min_station*, and *min_conf*

Variables:

NS: number of stations

C_k: set of size k candidates

L_k: set of support-prevalent size k patterns

TP_k: set of candidate size k hydrometeorological pattern

HM_k: set of size k hydrometeorological pattern

HM_patterns: set of meaningful hydrometeorological patterns

Algorithm:

```
1.  initialization; k = 0, TP_0 = D
2.  while (not empty TP_k) {
3.                  for (s = 1 to NS) {
4.                          If k = 0 then
5.                                  C_{k+1}(s) = get_singletons (D(s))
6.                          else
7.                                  C_{k+1}(s) = generate_candidates (TP_k(s))
8.                          C_{k+1}(s) = calculate_supports (C_{k+1}(s), D(s))
9.                          L_{k+1}(s) = prune_support_non_prevalent (C_{k+1}(s), min_support)
10.                 }
11.                 TP_{k+1} = calculate_station_support (L_{k+1})
12.                 HM_{k+1} = prune station_non_prevalent (TP_{k+1}, min_station)
13.                 k = k + 1
14. }
15. HM_patterns = discover_meaningful_patterns (HM, min_conf)
16. return HM_patterns
```

4.3. Execution trace of HMP-Miner

This section presents an execution trace of HMP-Miner. An example dataset is given in Table 1. The dataset includes 6 months of monthly average (monthly total for precipitation) values of hydrological and meteorological parameters for 5 stations. The parameter values are discretized as low (L), medium (M), and high (H). Table 1 presents the real and discretized values of the parameters for 5 stations. Details of the dataset preparation are given in Section 5.2.

Table 1. Original and discretized (labeled) parameter values of the example dataset.

	Months	Station #1203	Station #1221	Station #1222	Station #1224	Station #1226
Wind Speed	1	1.41 – M	1.70 – M	3.73 – M	1.91 – L	1.83 – M
	2	1.47 – M	1.61 – L	5.56 – H	3.08 – H	2.53 – H
	3	1.74 – H	1.65 – M	2.51 – L	2.47 – M	2.01 – M
	4	1.64 – H	1.99 – H	3.24 – M	3.01 – H	1.69 – L
	5	0.88 – L	1.62 – L	2.96 – L	1.95 – L	1.69 – L
	6	1.33 – M	1.73 – M	3.24 – M	2.61 – M	1.85 – M
Humidity	1	86.55 – H	73.11 – H	78.89 – L	83.72 – H	84.17 – H
	2	82.99 – H	77.05 – H	79.29 – L	74.43 – H	78.12 – H
	3	66.58 – L	63.20 – L	78.75 – L	56.59 – L	68.19 – M
	4	66.68 – L	62.35 – L	80.77 – M	59.84 – L	61.94 – L
	5	76.04 – M	71.27 – M	84.22 – H	72.34 – M	72.24 – M
	6	71.84 – M	71.64 – M	84.78 – H	59.68 – L	60.41 – L
Precipitation	1	1.49 – L	3.02 – M	3.60 – H	1.00 – L	3.61 – H
	2	1.88 – L	3.78 – H	3.67 – H	0.77 – L	1.15 – L
	3	3.18 – H	2.40 – M	2.79 – M	0.98 – L	1.56 – L
	4	1.84 – L	0.73 – L	1.31 – L	1.78 – M	2.43 – M
	5	3.54 – H	4.64 – M	1.78 – L	4.94 – H	5.49 – H
	6	2.29 – M	3.74 – H	2.73 – M	3.61 – H	2.38 – M
Temperature	1	–0.99 – L	6.04 – L	6.34 – L	–2.01 – L	–2.58 – L
	2	–0.17 – L	4.81 – L	5.76 – L	0.14 – L	–0.49 – L
	3	7.45 – M	11.74 – M	10.70 – M	7.40 – M	6.18 – M
	4	11.52 – M	15.03 – M	14.02 – M	12.15 – M	11.70 – M
	5	13.78 – H	17.65 – H	17.18 – H	13.75 – H	12.91 – H
	6	17.83 – H	21.54 – H	21.61 – H	18.67 – H	17.90 – H
Stream Flow	1	5.78 – L	118.00 – L	22.00 – L	7.41 – L	4.67 – L
	2	7.25 – L	191.00 – H	35.40 – M	7.71 – L	6.31 – L
	3	12.10 – H	147.00 – M	48.20 – H	12.40 – H	14.50 – M
	4	10.10 – M	134.00 – L	43.30 – H	7.88 – M	16.50 – M
	5	11.90 – H	194.00 – H	38.40 – M	11.70 – H	46.30 – H
	6	8.83 – M	132.00 – L	19.50 – L	6.90 – L	12.20 – M

In Tables 2–7, the execution trace of the HMP-Miner algorithm is given. *WSpeed* denotes wind speed, *Hum* denotes humidity, *Pre* denotes precipitation, *Temp* denotes air temperature, and *SFlow* denotes stream flow. Moreover, L, M, and H are used for low, medium, and high, respectively, for characterizing the magnitude of the parameters. If the support prevalence, station prevalence, and minimum confidence thresholds are 0.3, 0.4, and 0.3, respectively, HMP-Miner generates the rules listed in Table 7.

Table 2. Generation of size *1* patterns (for *min_support* = 0.3 and *min_station* = 0.4).

Pattern	Support values of the patterns in each of stations					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
{WSpeed-L}	1/6 – Pruned	2/6	2/6	2/6	2/6	4/5
{WSpeed-M}	3/6	3/6	3/6	2/6	3/6	5/5
{WSpeed-H}	2/6	1/6 – Pruned	1/6 – Pruned	2/6	1/6 – Pruned	2/5
{Hum-L}	2/6	2/6	3/6	3/6	2/6	5/5
{Hum-M}	2/6	2/6	1/6 – Pruned	1/6 – Pruned	2/6	3/5
{Hum-H}	2/6	2/6	2/6	2/6	2/6	5/5
{Pre-L}	3/6	1/6 – Pruned	2/6	3/6	2/6	4/5
{Pre-M}	1/6 - Pruned	3/6	2/6	1/6 – Pruned	2/6	3/5
{Pre-H}	2/6	2/6	2/6	2/6	2/6	5/5
{Temp-L}	2/6	2/6	2/6	2/6	2/6	5/5
{Temp-M}	2/6	2/6	2/6	2/6	2/6	5/5
{Temp-H}	2/6	2/6	2/6	2/6	2/6	5/5
{SFlow-L}	2/6	3/6	2/6	3/6	2/6	5/5
{SFlow-M}	2/6	1/6 – Pruned	2/6	1/6 – Pruned	3/6	3/5
{SFlow-H}	2/6	2/6	2/6	2/6	1/6 – Pruned	4/5

Table 3. Candidate size *2* patterns (for *min_support* = 0.3 and *min_station* = 0.4).

Pattern	Support values of the patterns in each of stations					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
{WSpeed-L, SFlow-L}	-	0 – Pruned	0 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned
{WSpeed-L, SFlow-M}	-	-	1/6 – Pruned	-	1/6 – Pruned	0 – Pruned
{WSpeed-L, SFlow-H}	-	2/6	1/6 – Pruned	1/6 – Pruned	-	1/5 – Pruned
{WSpeed-M, SFlow-L}	2/6	2/6	2/6	1/6 – Pruned	1/6 – Pruned	3/5
{WSpeed-M, SFlow-M}	1/6 – Pruned	-	0 – Pruned	-	2/6	1/5 – Pruned
{WSpeed-M, SFlow-H}	0 – Pruned	0 – Pruned	1/6 – Pruned	1/6 – Pruned	-	0 – Pruned
{WSpeed-H, SFlow-L}	0 – Pruned	-	-	1/6 – Pruned	-	0 – Pruned
{WSpeed-H, SFlow-M}	1/6 – Pruned	-	-	-	-	0 – Pruned
{WSpeed-H, SFlow-H}	1/6 – Pruned	-	-	0 – Pruned	-	0 – Pruned
{Hum-L, SFlow-L}	0 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned	0 – Pruned	0 – Pruned
{Hum-L, SFlow-M}	1/6 – Pruned	-	1/6 – Pruned	-	2/6	1/5 – Pruned
{Hum-L, SFlow-H}	1/6 – Pruned	0 – Pruned	1/6 – Pruned	1/6 – Pruned	-	0 – Pruned
{Hum-M, SFlow-L}	0 – Pruned	1/6 – Pruned	-	-	0 – Pruned	0 – Pruned
{Hum-M, SFlow-M}	1/6 – Pruned	-	-	-	1/6 – Pruned	0 – Pruned
{Hum-M, SFlow-H}	1/6 – Pruned	1/6 – Pruned	-	-	-	0 – Pruned
{Hum-H, SFlow-L}	2/6	1/6 – Pruned	1/6 – Pruned	2/6	2/6	3/5
{Hum-H, SFlow-M}	0 – Pruned	-	1/6 – Pruned	-	0 – Pruned	0 – Pruned
{Hum-H, SFlow-H}	0 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned	-	0 – Pruned
{Pre-L, SFlow-L}	2/6	-	0 – Pruned	2/6	1/6 – Pruned	2/5
{Pre-L, SFlow-M}	1/6 – Pruned	-	1/6 – Pruned	-	1/6 – Pruned	0 – Pruned
{Pre-L, SFlow-H}	0 – Pruned	-	1/6 – Pruned	1/6 – Pruned	-	0 – Pruned
{Pre-M, SFlow-L}	-	1/6	1/6	-	0	0 – Pruned
{Pre-M, SFlow-M}	-	-	0	-	2/6	1/5 – Pruned
{Pre-M, SFlow-H}	-	1/6 – Pruned	1/6 – Pruned	-	-	0 – Pruned
{Pre-H, SFlow-L}	0 – Pruned	1/6 – Pruned	1/6 – Pruned	1/6 – Pruned	1/6 – Pruned	0 – Pruned
{Pre-H, SFlow-M}	0 – Pruned	-	1/6 – Pruned	-	0 – Pruned	0 – Pruned
{Pre-H, SFlow-H}	2/6	1/6 – Pruned	0 – Pruned	1/6 – Pruned	-	1/5 – Pruned
{Temp-L, SFlow-L}	2/6	1/6 – Pruned	1/6 – Pruned	2/6	2/6	3/5
{Temp-L, SFlow-M}	0 – Pruned	-	1/6 – Pruned	-	0 – Pruned	0 – Pruned
{Temp-L, SFlow-H}	0 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned	-	0 – Pruned
{Temp-M, SFlow-L}	0 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned	2/6	1/5 – Pruned
{Temp-M, SFlow-M}	1/6 – Pruned	-	0 – Pruned	-	0 – Pruned	0 – Pruned
{Temp-M, SFlow-H}	1/6 – Pruned	0 – Pruned	2/6	1/6 – Pruned	-	1/5 – Pruned
{Temp-H, SFlow-L}	0 – Pruned	1/6 – Pruned	1/6 – Pruned	1/6 – Pruned	0 – Pruned	0 – Pruned
{Temp-H, SFlow-M}	1/6 – Pruned	-	1/6 – Pruned	-	1/6 – Pruned	0 – Pruned
{Temp-H, SFlow-H}	1/6 – Pruned	1/6 – Pruned	0 – Pruned	1/6 – Pruned	-	0 – Pruned

Table 4. Prevalent size *3* patterns (for *min_support* = 0.3 and *min_station* = 0.4).

Pattern	Support values of the pattern in each of stations					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
{WSpeed-M, SFlow-L}	2/6	2/6	2/6	1/6 – Pruned	1/6 – Pruned	3/5
{Hum-H, SFlow-L}	2/6	1/6 – Pruned	1/6 – Pruned	2/6	2/6	3/5
{Pre-L, SFlow-L}	2/6	-	0 – Pruned	2/6	1/6 – Pruned	2/5
{Temp-L, SFlow-L}	2/6	1/6 – Pruned	1/6 – Pruned	2/6	2/6	3/5

Table 5. Candidate size *3* patterns (for *min_support* = 0.3 and *min_station* = 0.4).

Pattern	Support values of the patterns in each station					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
{Hum-H, Pre-L, SFlow-L}	2/6	-	-	2/6	-	2/5
{Hum-H, Temp-L, SFlow-L}	2/6	-	-	2/6	2/6	3/5
{Pre-L, Temp-L, SFlow-L}	2/6	-	-	2/6	-	2/5

Table 6. Candidate size *5* patterns (for *min_support* = 0.3 and *min_station* = 0.4).

Pattern	Support values of the patterns in each of stations					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
{Hum-H, Pre-L, Temp-L, SFlow-L}	2/6	-	-	2/6	-	2/5

Table 7. Confidence-based pruning with *min_conf* = 0.3 and output of HMP-Miner.

Rule	Confidence values of the rules in each stations					Station Prevalence
	#1203	#1221	#1222	#1224	#1226	
WSpeed-M → SFlow-L	2/3	2/3	2/3	-	-	3/5
Hum-H → SFlow-L	2/2	-	-	2/2	2/2	3/5
Pre-L → SFlow-L	2/3	-	-	2/3	-	2/5
Temp-L → SFlow-L	2/2	-	-	2/2	2/2	3/5
Hum-H, Pre-L → SFlow-L	2/2	-	-	2/2	-	2/5
Hum-H, Temp-L → SFlow-L	2/2	-	-	2/2	2/2	3/5
Pre-L, Temp-L → SFlow-L	2/2	-	-	2/2	-	2/5
Hum-H, Pre-L, Temp-L → SFlow-L	2/2	-	-	2/2	-	2/5

HMP-Miner starts with discovering size *1* patterns. The list of candidate size *1* patterns is given in Table 2. For each pattern, their support values are calculated for each station. The ones that do not satisfy the support prevalence threshold of 0.3 are pruned. Next, the station prevalences of the remaining size *1* patterns are calculated. In Table 2, the last column shows the station prevalence values of the patterns, where it can be seen that none of the patterns are pruned based on the station prevalence, since they satisfy the threshold of 0.4.

In the second step, by joining prevalent size *1* patterns, candidate size *2* patterns are generated. In this study, since we aim to find the effect of meteorological parameters on hydrological parameters, one of the parameters in the size *2* candidates is a hydrological parameter, the stream flow. Table 3 presents size *2* patterns and their support values for each station.

For example, if we inspect the example dataset given in Table 1, the Pre-L and SFlow-L items (that is, the pattern of {Pre-L, SFlow-L}) are together in 2 (1st and 2nd months) out of 6 months at station #1203. Size *2* patterns whose support values are not satisfying the min_support threshold of 0.3 are pruned at this

stage. For example {Pre-L, SFlow-L} is pruned at stations #1222 and #1226 since its support of 1/6 is less than the minimum support threshold 0.3. The pattern {Pre-L, SFlow-L} does not occur at station #1221 and so this pattern is represented with a '-' sign in Table 3. Next, the station prevalence values of the patterns are calculated. The station prevalence of pattern {Pre-L, SFlow-L} is 2/5 since it is support-prevalent at 2 (station #1203 and #1224) out of 5 stations. If the station prevalence is greater than or equal to the threshold of 0.4, the pattern is a hydrometeorologically prevalent pattern and it will be used at the next candidate size k generation of the algorithm. The pattern {Pre-L, SFlow-L} is hydrometeorologically prevalent since its station prevalence value 2/5 satisfies the *min_station* threshold of 0.4. The patterns that do not satisfy the *min_station* threshold are pruned as shown in Table 3. The list of hydrometeorological-prevalent size *2* patterns can be seen in Table 4.

By joining size *2* prevalent patterns, candidate size *3* patterns are generated (Table 5). By joining size *3* prevalent patterns, candidate size *4* patterns are generated (Table 6), and so on. This process continues until no more candidate hydrometeorological patterns are left.

Finally, meaningful hydrometeorological rules that satisfy the minimum confidence threshold *min_conf* of 0.3 are generated as shown in Table 7. In each station, any rule that does not satisfy the *min_conf* threshold is pruned. The rules given in Table 7 all satisfy the *min_conf* threshold of 0.3 and there will be no confidence-based pruning. Next, after these prunings, the station prevalences of the rules are rechecked. The station prevalences of the rules can be seen in the last column of Table 7. None of the rules are pruned at this stage since all station prevalence values satisfy the *min_station* threshold of 0.4. Finally, hydrometeorological-prevalent rules (such as the ones satisfying, *min_support*, *min_station*, and *min_conf* thresholds) are returned by the algorithm as shown in Table 7.

4.4. Experimental evaluation

In this section, we compare the performances of the proposed naïve algorithm and our proposed HMP-Miner algorithm on a real dataset. The experimental evaluation aims to answer the questions of what the effect of the support threshold is, what the effect of the station threshold is, and what the effect of the number of stations is. The experiments are conducted on a computer that has an Intel Core 2-Quad 2.66-GHz CPU and 3-GB RAM.

4.5. Experimental setup

The experimental setup used to test the performance of the algorithms is presented in Figure 2. In the data selection step, the gauging and meteorology stations are selected from a database and a correlation analysis is conducted to ensure that the each gauging station is matched with a meteorology station. In the preprocessing step, the data are prepared for association analysis. First, the daily data are converted to monthly average or total data. Next, the data are discretized. After the preprocessing step, the data are used for the experimental evaluation. Below, these steps are explained in detail.

4.6. Data set

The dataset contains data for precipitation, air temperature, wind speed, relative humidity, and stream flow parameters. Data for all of the parameters were originally daily observed values. They are converted to monthly values by calculating the average for air temperature, wind speed, relative humidity, and stream flow, and total for precipitation. Stream flow data are obtained from 64 gauging stations. These 64 stations were previously selected from more than 300 stations, based on their completeness, homogeneity, and length [20]. The climate

data are obtained from a meteorology station in the same river basin as the stream flow gauging stations (Figure 3). The homogeneity of the meteorological data is tested to ensure data quality [21]. As there is more than one meteorology station in a river basin, we select the meteorology station whose data have the highest correlation with the stream flow data. All of the data are available from 1975 to 2000.

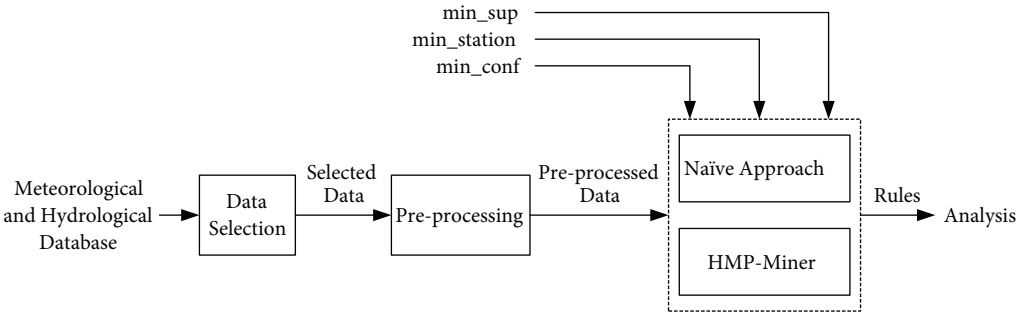

Figure 2. Experimental setup.

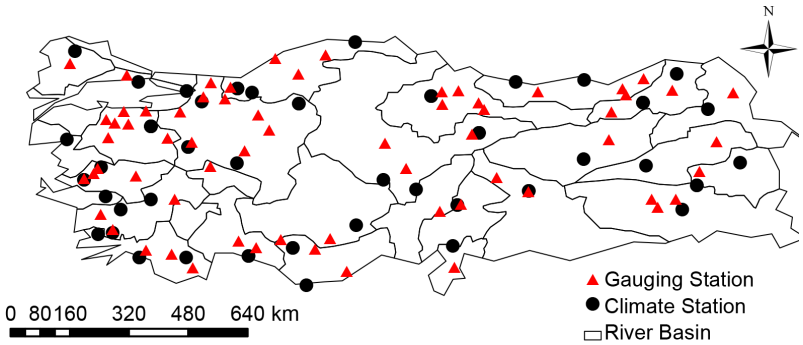

Figure 3. The locations of the stream flow gauging stations and meteorology stations used in this study.

The data are converted to discrete format for the analysis. We discretize the data into 3 groups using their statistical properties (i.e. mean (μ) and standard deviation (σ)). The data are named 'medium (M)' if they are between 'μ - 0.5σ' and 'μ + 0.5 σ'; 'low (L)' if they are 'smaller than μ - 0.5σ'; and 'high (H)' if they are 'higher than μ + 0.5σ'. An example discretization for the stream flow of a station can be seen in Figure 4.

4.7. Experimental results

In this section, we present our experimental evaluations of several design decisions and workload parameters of the HMP-Miner algorithm.

4.7.1. Effect of the support threshold

In this experiment, we evaluate the effect of the support threshold on the execution time for both algorithms (Figure 5). The station and confidence thresholds are set to 0.8 and 0.1, respectively. The results show that the execution times of the algorithms decrease as the support threshold increases. It can also be seen that HMP-Miner takes less time than the naïve approach, because it prunes station nonprevalent patterns as early as possible and the candidate patterns are generated using potentially successful patterns.

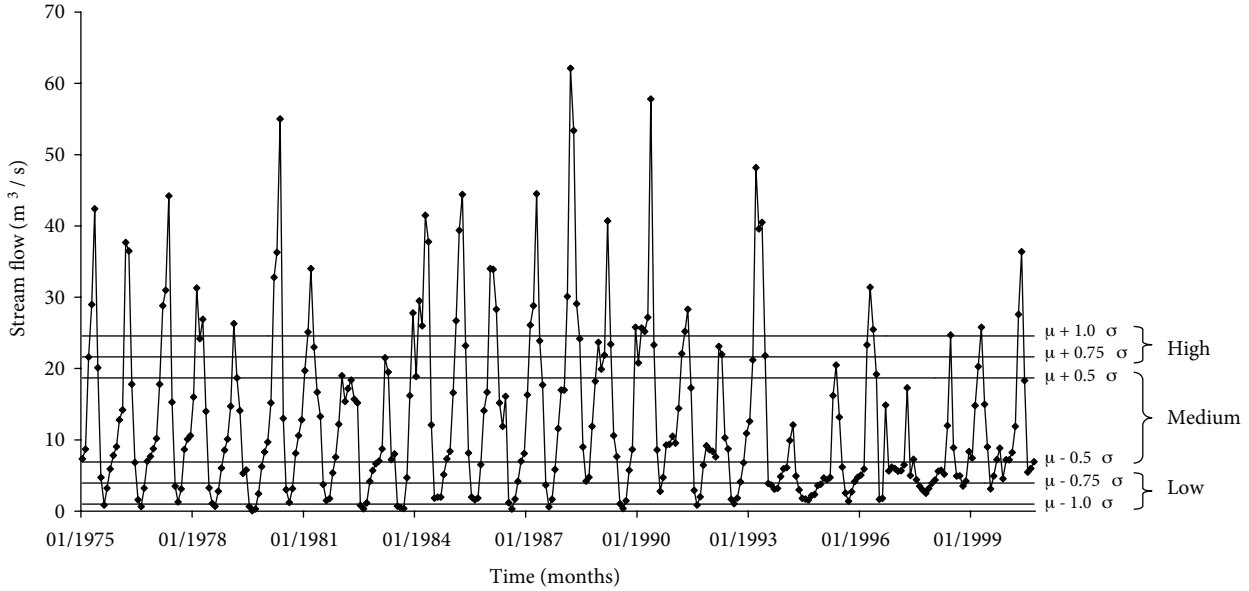

Figure 4. An example of data discretization for stream flow parameter of a station.

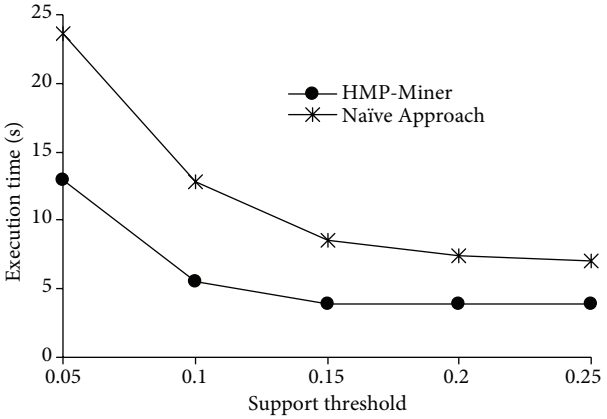

Figure 5. Effect of support threshold.

4.7.2. Effect of the station threshold

In this experiment, we evaluate the effect of the station threshold on the execution time for both algorithms (Figure 6). The support and confidence thresholds are set to 0.1 and 0.1, respectively. The results show that HMP-Miner is sensitive to the value of the station prevalence threshold and the computational cost of the HMP-Miner decreases as the station prevalence threshold value increases. In contrast, the naïve approach is less sensitive to the station prevalence threshold, since it applies station prevalence-based pruning at the postprocessing step, which is computationally cheaper than support-based pruning. As a result, HMP-Miner outperforms the naïve approach, since HMP-Miner generates and deals with fewer patterns than the naïve algorithm.

4.7.3. Effect of number of stations

In this experiment, we evaluate the effect of the station number on the execution time for both algorithms (Figure 7). The support, station, and confidence thresholds are set to 0.1, 0.8, and 0.1, respectively. The results

show that naïve approach has a linear-like growth scheme. However, the HMP-Miner algorithm has more flows from the linearity. This shows that HMP-Miner is more sensitive to the station number than the naïve approach.

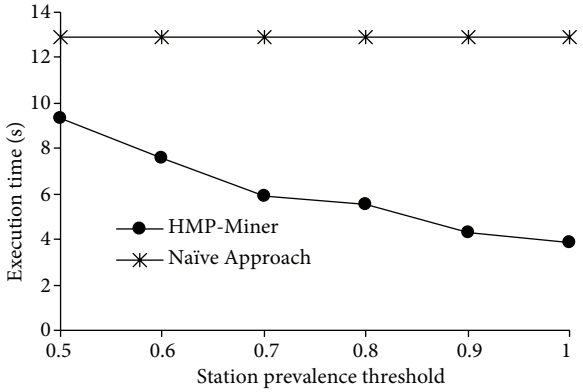

Figure 6. Effect of station prevalence threshold.

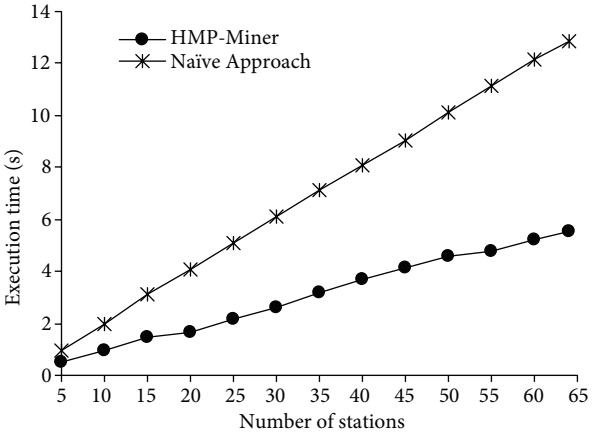

Figure 7. Effect of number of stations.

5. Evaluation of results

HMP-Miner is used to determine the hydrometeorological patterns over Turkey (Table 8; Figure 8). In this analysis, we use stream flow data from 64 stream flow gauging stations and associated meteorology stations. The data are in a monthly timescale and cover the period from 1975 to 2000. A support threshold of 0.15, a station prevalence threshold of 0.5, and a confidence threshold of 0.5 are used for the discovery of frequent rules. The targets were the low (L), medium (M), and high (H) stream flow.

Table 8. Rules discovered by the HMP-Miner (support prevalence threshold = 0.15; station prevalence threshold = 0.5, confidence threshold = 0.5) (L: Low, M: Medium, H: High).

Rules	Support Prevalence in all 64 Stations	Station Prevalence over 64 Stations
Pre-L → SFlow-L	≥ 0.20	0.9
Temp-H → SFlow-L	≥ 0.20	0.8
Temp-L → SFlow-M	≥ 0.15	0.6
Pre-L, Temp-H → SFlow-L	≥ 0.15	0.6
Hum-L → SFlow-L	≥ 0.15	0.6
Hum-H → SFlow-M	≥ 0.15	0.5
WSpeed-M → SFlow-L	≥ 0.15	0.5

With HMP-Miner, 7 rules (Table 8) that satisfy the given thresholds are discovered. The rules discovered clearly show that there is a strong relationship between the stream flow and precipitation and air temperature, particularly for low stream flows. The stream flow appears low when the precipitation is low and temperature is high at a majority of the stations. The rules Pre-L→ SFlow-L and Temp-H → SFlow-L are spatially prevalent all over Turkey, except for a small region located in northeastern Turkey (Figure 8). The station prevalences for these 2 rules are 0.9 and 0.8, respectively. This means that these rules are present at at least 80% of the stations. Relationships with relative humidity and wind speed are also present. The stream flow is low when the relative humidity is low and wind speed is high. Medium ranges of stream flow are found to be associated with low temperatures and high relative humidity.

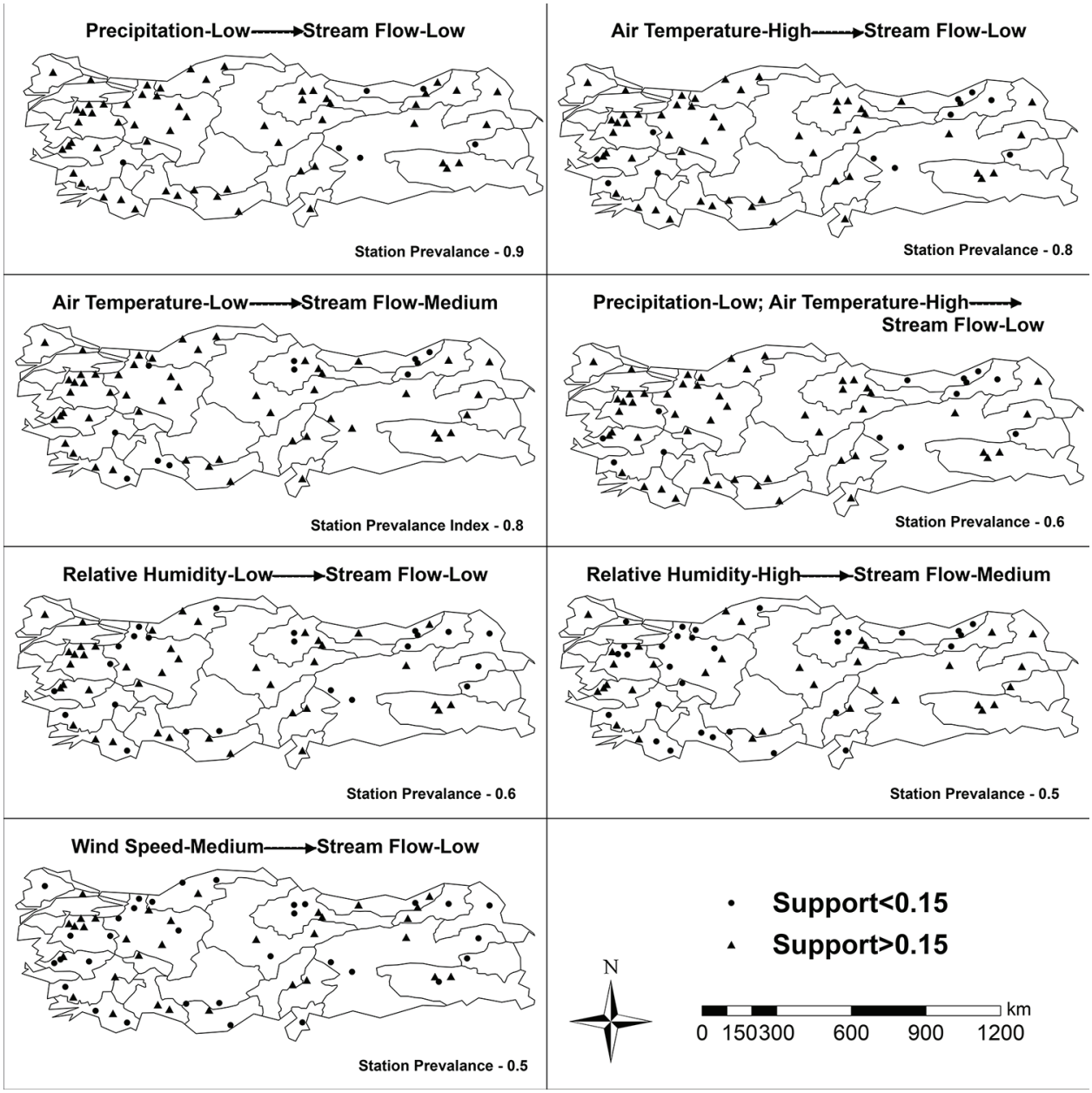

Figure 8. Rules discovered by the HMP-Miner (support prevalence threshold = 0.15; station prevalence threshold = 0.5, confidence threshold = 0.5).

The results found in this study are evaluated to determine if they are correct and interesting. For this purpose, we compare our results with previous studies conducted on the same topic. Our evaluation shows that the results found in this study are consistent with many other studies that examined the relationships between the stream flow and meteorological parameters [e.g., 22–25]. Similar to our results, previous studies also indicated that precipitation and air temperature are the 2 most important parameters affecting the stream flow. More recently, Dadaser-Celik and Cengiz [21] examined the correlations of the stream flow with various meteorological parameters in Turkey and showed that the correlations between stream flow and air temperatures are strong over Turkey.

The techniques used in this study also provide some advantages over the other techniques used to examine the relationships between the stream flow and meteorological parameters. In many previous studies, techniques such as regression or correlation analysis were used. Regression analysis provides a model of the data in an expected error range. The correlation analysis examines the degree and direction of relationships between 2 variables. The HMP-Miner algorithm developed in this study outperforms the other techniques because with HMP-Miner, we not only examine the relationships between the stream flow and several meteorological parameters, but also produce rules that show the cause–effect relationships between various combinations of variables. HMP-Miner is also efficient for analyzing large datasets that contain several variables.

6. Conclusions and future work

We defined the hydrometeorological patterns and the problem of mining these patterns. We proposed a novel computationally efficient HMP-Miner. We developed an interest measure (hydrometeorological prevalence index) for finding hydrometeorological patterns. The proposed HMP-Miner was compared with the naïve approach. We also evaluated the proposed algorithms experimentally. The results found by the HMP-Miner were evaluated. The experimental evaluations showed that the proposed algorithm outperforms the naïve alternative. The evaluation of results showed that HMP-Miner can successfully find the relationships between hydrological and meteorological parameters and provide many advantages over the classical methods.

As future work, we plan to extend our proposed algorithm for mining hydrometeorological patterns in different spatial and temporal scales.

Acknowledgments

This study was partially supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK), Project Number: CAYDAG 110Y110 and the Research Fund of Erciyes University, Project Number: FBA-09-866. We would like to thank Eda Cengiz for her help at the data preparation and quality assessment steps.

References

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Waltham, MA, USA, Morgan Kaufmann, 2001.

[2] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Boston, MA, USA, Addison-Wesley, 2006.

[3] S. Shekhar, R.R. Vatsavai, M. Celik, Spatial and Spatiotemporal Data Mining: Recent Advances, in Next Generation of Data Mining, In: H. Kargupta, J. Han, P.S. Yu, R. Motwani, V. Kumar, Eds., Boca Raton, FL, USA, CRC Press, 2009.

[4] A.R. Ganguly, K. Steinhaeuser, "Data mining for climate change and impacts", International Workshop on Spatial and Spatiotemporal Data Mining, IEEE International Conference on Data Mining, 2008.

[5] S. Kotsiantis, D. Kanellopoulos, "Association rules mining: a recent overview", GESTS International Transactions on Computer Science and Engineering, Vol. 32, pp. 71–82, 2006.

[6] J. Hipp, U. Güntzer, G. Nakhaeizadeh, "Algorithms for association rule mining - a general survey and comparison", SIGKDD Explorations, 2000.

[7] H.C. Kob, G. Tan, "Data mining application in healthcare", Journal of Healthcare Information Management, Vol. 19, pp. 64–71, 2005.

[8] H. Chen, S.S. Fuller, C. Friedman, W. Hersh, Medical Informatics: Knowledge Management and Data Mining in Biomedicine, New York, NY, USA, Springer, 2005.

[9] P. Giudici, S. Figini, Applied Data Mining for Business and Industry, 2nd Edition, New York, NY, USA, Wiley, 2009.

[10] D. Olson, Y. Shi, Introduction to Business Data Mining, New York, NY, USA, McGraw-Hill, 2005.

[11] T. Tadesse, D.A. Wilhite, M.J. Hayes, S.K. Harms, S. Goddard, "Drought monitoring using data mining techniques: a case study for Nebraska, USA", Natural Hazards, Vol. 33, pp. 137–159, 2004.

[12] T. Tadesse, D.A. Wilhite, M.J. Hayes, S.K. Harms, S. Goddard, "Discovering associations between climatic and oceanic parameters to monitor drought in Nebraska using data-mining techniques", Journal of Climate, Vol. 18, pp. 1541–1550, 2005.

[13] F. Lin, "Discovery of teleconnections using data mining technologies in global climate datasets", Data Science Journal, Vol. 6, pp. 749–755, 2007.

[14] P.N. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, A. Torregrosa, "Finding spatio-temporal patterns in earth science data", KDD Workshop on Temporal Data Mining, 2001.

[15] C.T. Dhanya, D.N. Kumar, "Data mining for evolution of association rules for droughts and floods in India using climate inputs", Journal of Geophysical Research, Vol. 114, pp. D02102, 2009.

[16] D.N. Kumar, M. Ish, C.T. Dhanya, "Data mining and it's applications for modelling rainfall extremes", ISH Journal of Hydraulic Engineering, Vol. 15, pp. 25–51, 2009.

[17] H. Shu, X. Zhu, S. Dai, "Mining association rules in geographical spatio-temporal data", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 37, pp. 225–228, 2008.

[18] F. Dadaser-Celik, M. Celik, A.S. Dokuz, "Associations between stream flow and climatic variables at Kızılırmak River Basin in Turkey", Global Nest Journal, Vol. 14, pp. A-367–A-374, 2012.

[19] R. Agrawal, R. Srikant, "Fast algorithms for mining association rule", 20th International Conference on Very Large Data Bases, 1994.

[20] E. Kahya, M.C. Karabörk, "The analysis of El Nino and La Nina signals in streamflows of Turkey", International Journal of Climatology, Vol. 21, pp. 1231–1250, 2001.

[21] F. Dadaser-Celik, E. Cengiz, "Correlations of stream flow and climatic variables in Turkey", BALWOIS, 2012.

[22] S.A. Changnon, K.E. Kunkel, "Climate-related fluctuations in Midwestern floods during 1921–1985", Journal of Water Resources Planning and Management, Vol. 121, pp. 326–334, 1995.

[23] P.Y. Groisman, R.W. Knight, T.R. Karl, "Heavy precipitation and high streamflow in the contiquous United States: trends in the twentieth century", Bulletin of the American Meteorological Society, Vol. 82, pp. 219–246, 2001.

[24] D.R. Cayan, L.G. Riddle, E. Aguado, "The influence of precipitation and temperature on seasonal streamflow in California", Water Resources Research, Vol. 29, pp. 1127–1140, 1993.

[25] L.L. Kletti, H.G. Stefan, "Correlations of climate and streamflow in three Minnesota streams", Climatic Change, Vol. 37, pp. 575–600, 1997.