Sains Malaysiana 49(10)(2020): 2345-2358
http://dx.doi.org/10.17576/jsm-2020-4910-02

Assessing the Relationship between Pollution Sources and Water Quality Parameters of Sungai Langat Basin using Association Rule Mining

(Menilai Hubungan antara Sumber Pencemaran dan Parameter Kualiti Air di Lembangan Sungai Langat menggunakan Peraturan Persatuan Perlombongan)

Nurizah Abdul Hasib & Zalinda Othman*

ABSTRACT

Due to water pollution, the water quality in Sungai Langat Basin is declining. This study was conducted in this regard to identify the sources of pollution and to analyse the relationship between sources of pollution and water quality parameters. The Malaysian Department of Environment (DOE) has provided data on pollution sources and water quality. Both data are combined and pre-processed through association rule mining for further modelling processes. Apriori algorithm was used to generate rules in finding any relationships between sources of pollution and water quality parameters. Water quality experts had analysed and validated the generated rules. The analysis of expert acceptance showed that 65% of the formulated rules were agreed by the experts while 35% of the rules were disagreed. Furthermore, the findings showed that all experts agreed that there was a relation between NH_3-N values ranging from 0.9 to 2.7 and the sewage as a source of pollution. Besides, as agreed by all experts, either effluent, sewage or livestock is the pollution source of Sungai Langat.

Keywords: Association rule mining; pollution source identification; Sungai Langat Basin; water pollution

ABSTRAK

Kualiti air di Lembangan Sungai Langat semakin merosot ekoran daripada pencemaran air yang berlaku. Sehubungan itu, kajian ini dijalankan untuk mengenal pasti punca pencemaran dan menganalisis hubungan antara punca pencemaran dan parameter kualiti air. Data punca pencemaran dan data kualiti air diperoleh daripada Jabatan Alam Sekitar Malaysia. Kedua-dua data digabungkan dan diproses lebih awal menggunakan kaedah perlombongan petua sekutuan untuk digunakan dalam proses pemodelan seterusnya. Algoritma Apriori yang terdapat dalam kaedah perlombongan petua sekutuan digunakan untuk mendapatkan petua yang mengaitkan hubungan antara punca pencemaran dan parameter kualiti air. Petua yang dijana dianalisis sebelum disahkan oleh pakar. Analisis penerimaan pakar menunjukkan rumusan petua yang dipersetujui oleh pakar adalah sebanyak 65% manakala 35% tidak dipersetujui oleh pakar. Keputusan kajian menunjukkan semua pakar bersetuju bahawa terdapat hubungan antara nilai NH_3-N dalam lingkungan 0.9 hingga 2.7 dengan punca pencemaran kumbahan. Selain itu, punca pencemaran Sungai Langat yang telah dipersetujui oleh semua pakar adalah sama ada efluen, kumbahan atau ternakan.

Kata kunci: Lembangan Sungai Langat; pencemaran air; pengenalan punca pencemaran; perlombongan petua sekutuan

Introduction

Malaysia's water resources that come from surface water, such as rivers, lakes, ponds, and groundwater are used for domestic, commercial, and industrial. In Malaysia, 97% of water supply are supplied by rivers as the main source of water (Chan 2012). Sungai Langat Basin provides Selangor residents with 27% of raw water (STCPD 2012). Water pollution will therefore, affect the river's main function of supplying raw water. Ammonia and manganese pollution, for example, caused the Semenyih Water Treatment Plant to be closed on March 3, 2015 (Mazlinda 2015). This pollution caused water supply disruption to the residents of Petaling, Sepang, Klang, Kuala Langat, and Hulu Langat (Liang 2015). In 2016, Sungai Langat was contaminated eight times due to an increase in ammonia (Anon 2016).

According to the Malaysian Department of Environment's (DOE) Environmental Quality Report, Sungai Langat Basin Water Quality Index (WQI) deteriorated within five years from 2012 to 2016. For example, the water quality in Sungai Langat declined from 73 WQI values in 2012 to 64 WQI values in 2016 (DOE 2016, 2012). WQI deterioration also occurred in Sungai Langat Basin, Sungai Semenyih, Sungai Jijan, Sungai Langat, and Sungai Batang Nilai.

Industrial pollution source affects the water quality parameters in the middle stream of Sungai Langat, namely DO, BOD, and NH_3-N (Abidin et al. 2018). Sand mining activity also has an impact on the parameters of total suspended solids (TSS) in Sungai Langat upstream and downstream (Abidin et al. 2018). In the middle stream and downstream of Sungai Langat Basin, industrial, municipal waste, and agricultural runoff are sources of pollution (Juahir et al. 2011). In the upstream of Sungai Langat Basin, both domestic and agricultural runoffs are pollution sources (Juahir et al. 2011).

Upstream Sungai Langat is moderately polluted by identified pollutants due to agricultural, industrial, domestic waste, sewage, and livestock (Gasim et al. 2010). In the meantime, Sungai Langat has been identified as moderately polluted by agricultural, industrial, and residential areas (Ahmad et al. 2015). This study will, therefore, identify sources of pollution from the location of premises in DOE's Geographic Information System (GIS).

Zubir et al. (2016) used a chemometric method to assess the use of organic pollutants as a chemical indicator in Sungai Langat Basin. On the other hand, Ahmad et al. (2015) applied water sampling and statistical methods to assess water quality in Sungai Langat. Gasim et al. (2015) conducted a study using water sampling and statistics methods specifically on physicochemical and microbial water quality parameters on the upstream of Sungai Langat. Abidin et al. (2018) used water quality modelling, namely QUAL2 K, to determine the diversity of water quality and evaluate the effect of land use on Sungai Langat. Al-Badaii et al. (2016) utilised environmetric methods to evaluate dissolved heavy metals in Sungai Semenyih.

Association rule mining is a data mining technique to identify relationships from frequently identified items in the data (Othman et al. 2018). The aim is to discover knowledge from the frequent itemset about the relationship. This study will, therefore, use the association rules method to determine the relationship between sources of pollution and parameters of water quality. Previous researchers used the association rules method to explore the relationship between item datasets in various fields such as medicine (Fahrudin et al. 2017), meteorology (Harun et al. 2017), agriculture (Gandhi & Armstrong 2016), and the environment (Gour et al. 2016). Fahrudin et al. (2017) employed the association rules method to use data from cancer patients to diagnose breast cancer. In Terengganu, Malaysia, Harun et al. (2017) also used the association rules method to study flood predictions. In this case, from flood data, researchers are able to study the relationship between water levels and flood areas. Gandhi and Armstrong (2016) also applied the association rules method to study the rainy season's effects on paddy yields in Rajasthan State, India. Gour et al. (2016) used association rules in the field of environment to analyse the relationship between water quality parameters in the Narmada River, India. The purpose of this study was, therefore, to identify sources of pollution and the relationship between sources of pollution and parameters of water quality in Sungai Langat Basin using the association rules method.

MATERIALS AND METHODS

STUDY AREA

Selangor has five river basins, namely Sungai Selangor Basin, Sungai Langat Basin, Sungai Klang Basin, Sungai Buloh Basin, and Sungai Tengi Basin. The area of this

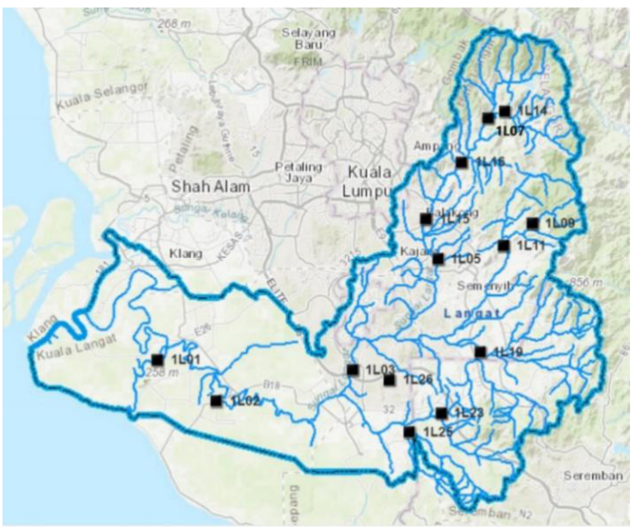

FIGURE 1. Location Map of Sungai Langat Basin and Water Quality Monitoring Stations
Source: GEOSpAS DOE

study is Sungai Langat Basin, consisting of several rivers including Sungai Langat, Sungai Semenyih, Sungai Lui, Sungai Pajam, Sungai Batang Nilai, Sungai Jijan, and Sungai Batang Labu. It covers an area of 2,394.38 square kilometres in an area of 40'U latitude 2 to 3° 20'U and longitude L01° 10'E to 102000'E (Khairul et al. 2000). The basin area includes Selangor, Putrajaya, Kuala Lumpur, and a small part of Nilai, Negeri Sembilan as well as across the districts of Sepang and Kuala Langat as shown in Figure 1.

The Malaysian DOE provided the water quality data in this study for a period of five years from 2012 to 2016 as shown in Table 1.

TABLE 1. Water quality data

No.	Data information	Details		
1	Number of attributes		46 attributes	
2	Attribute types	1) Nominal	2) Numeric	3) Date
3	Total instances		560	
4	Missing data		56 (10% of total data)	
5	Duplicate data		7	

The water quality data was collected from Sungai Langat Basin water quality monitoring stations. Information about Sungai Langat Basin monitoring stations is shown in Table 2.

TABLE 2. Information on water quality monitoring stations

No.	Location	Station	River
1	Kg Air Tawar	1L01	Sg. Langat
2	Telok Datok	1L02	Sg. Langat
3	Kg. Dengkil	1L03	Sg. Langat
4	Bandar Kajang	1L05	Sg. Langat
5	Jambatan Bt 18	1L07	Sg. Langat
6	Semenyih	1L09	Sg. Semenyih
7	Kg. Pasir Semenyih	1L11	Sg. Semenyih
8	Kg. Masjid	1L14	Sg. Lui
9	Hulu intake Point Loji Langat	1L15	Sg. Langat
10	Batu 14, Hulu Langat	1L16	Sg. Langat
11	Kg Kuala Pajam	1L19	Sg. Pajam
12	Bandar Nilai	1L23	Sg. Batang Nilai
13	Kampung Jijan	1L25	Sg. Jijan
14	Batang Labu	1L26	Sg. Batang Labu

Data from pollution sources were retrieved through the Geographical Information System (GIS) web services, namely GEOSpAS, based on the location of premises that discharge waste from a point source or non-point source into the river. The location of the premises together with the location of the monitoring stations were retrieved and mapped into ArcGIS software. The location of pollution sources should be on the upstream while the monitoring station is on the downstream of the river as these are the criteria for determining pollution sources. For example, the shaded area in Figure 2 is a premise that discharges upstream effluent while monitoring downstream stations, namely 1L16. Therefore, 1L16 is an effluent source of pollution.

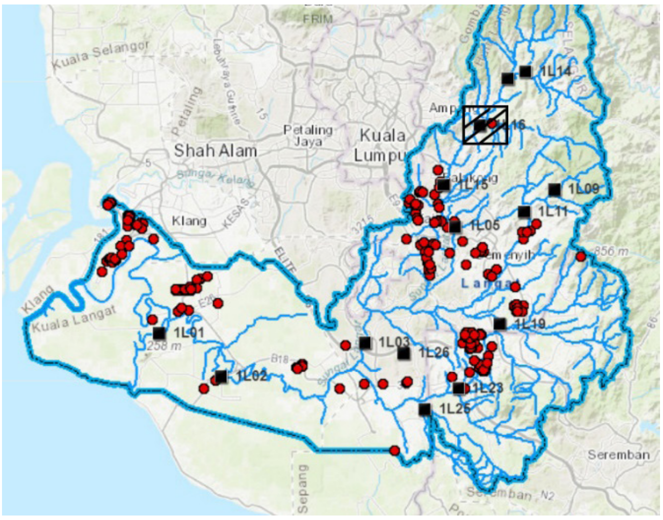

FIGURE 2. Location of premises that discharge effluent
Source: GEOSpAS DOE

Effluent, sewage, livestock, palm oil mill, construction site, and market were pollution sources identified in this study. Scheduled waste was excluded as a source of pollution as it should be disposed of in the prescribed premises (Malaysia 2005). In this respect, the list of pollution sources is as shown in Table 3.

TABLE 3. List of pollution sources

No.	Station	River	Pollution sources					
			Effluent	Sewage	Livestock	Palm oil mill	Construction site	Market
1	1L01	Sungai Langat	Yes	Yes	Yes	Yes	Yes	Yes
2	1L02	Sungai Langat	Yes	Yes	Yes	Yes	Yes	Yes
3	1L03	Sungai Langat	Yes	Yes	Yes	No	Yes	Yes
4	1L05	Sungai Langat	Yes	Yes	Yes	No	Yes	Yes
5	1L07	Sungai Langat	No	No	Yes	No	No	No
6	1L09	Sungai Semenyih	No	No	No	No	No	No
7	1L11	Sungai Semenyih	No	No	No	No	No	No
8	1L14	Sungai Lui	No	No	Yes	No	No	No
9	1L15	Sungai Langat	Yes	Yes	Yes	No	Yes	Yes
10	1L16	Sungai Langat	Yes	Yes	Yes	No	No	No
11	1L19	Sungai Pajam	Yes	Yes	Yes	No	Yes	Yes
12	1L23	Sungai Batang Nilai	Yes	Yes	No	No	Yes	Yes
13	1L25	Sungai Jijan	No	Yes	No	Yes	No	Yes
14	1L26	Sungai Batang Labu	No	Yes	No	No	Yes	No

The label 'Yes' refers to a source of pollution in Table 3, whereas 'No' means not a source of pollution for the particular station. For example, effluent, sewage, livestock, palm oil mill, construction site, and market were identified as pollution sources of the monitoring station 1L01. Meanwhile, monitoring station 1L07's source of pollution was livestock.

RESEARCH WORKFLOW

The research process flow was adapted from Fayyad's Knowledge Discovery in Databases (KDD) process flow (Fayyad et al. 1996). This study's research flow, as shown in Figure 3, consisted of three phases. In the first phase, the research started with collecting water quality data and pollution sources data from the Malaysian DOE. At this phase, the data would be cleaned from missing values, outliers, duplication, inconsistencies and much more. Meanwhile, data from sources of pollution were retrieved via web services with read-only access control.

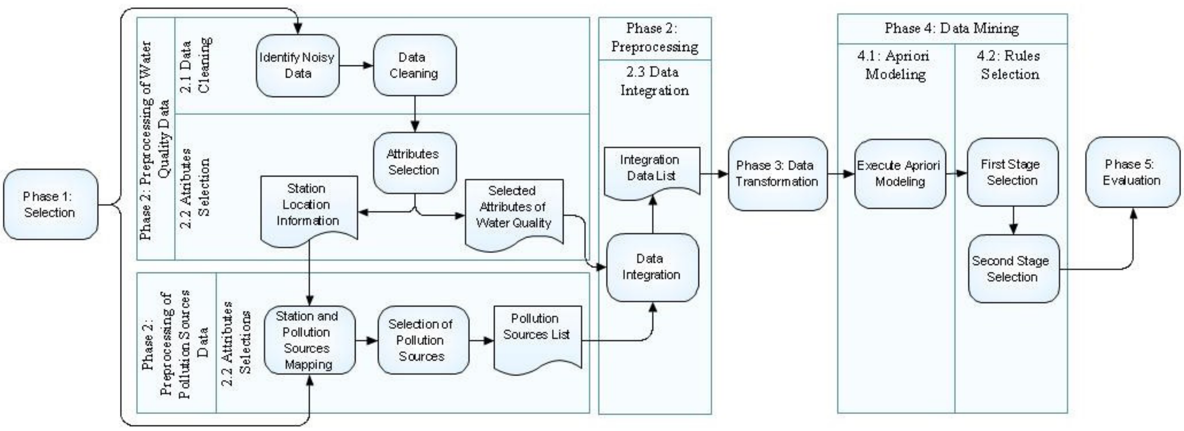

FIGURE 3. Research flow

EXPERIMENT DATASET

The selection phase was to identify the data that would be used to achieve this study's objectives. The attribute selection process was to select attributes for water quality and sources of pollution data. Based on the total of 46 water quality parameters, as shown in Table 1, this study chose six parameters. The water quality parameters used in this study were Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Ammonia Nitrogen (NH3-N), Suspended Solids (SS), and pH. This was based on the water quality index parameters as shown in Table 4.

TABLE 4. Water Quality Index based on parameters

Parameter	Unit	Class				
		I	II	III	IV	V
NH3-N	mgL^{-1}	< 0.1	0.1 - 0.3	0.3 - 0.9	0.9 - 2.7	> 2.7
BOD	mgL^{-1}	< 1	1 - 3	3 - 6	6 - 12	> 12
COD	mgL^{-1}	< 10	10 - 25	25 - 50	50 - 100	> 100
DO	mgL^{-1}	> 7	5 - 7	3 - 5	1 - 3	< 1
pH	-	> 7.0	6.0 - 7.0	5.0 - 6.0	< 5.0	> 5.0
SS	mgL^{-1}	< 25	25 - 50	50 - 150	150 - 300	> 300
WQI	-	> 92.7	76.5 – 92.7	51.9 – 76.5	31.0 – 51.9	< 31.0

Source: DOE (2015)

Rivers and stations were considered in this study because these attributes were related to the pollution sources and water quality index. The list of selected attributes for water quality is shown in Table 5.

TABLE 5. Selected water quality attributes

No.	Attribute	Description
1	DO	Water quality parameter, numeric, chemical characteristic
2	BOD	Water quality parameter, numeric, chemical characteristic
3	COD	Water quality parameter, numeric, chemical characteristic
4	SS	Water quality parameter, numeric, physical characteristic
5	NH3N	Water quality parameter, numeric, chemical characteristic
6	pH	Water quality parameter, numeric, physical characteristic
7	River	River name at Sungai Langat Basin
8	WQI	Water quality classification, numeric
9	Station	monitoring station example IL01

The pollution source attribute selection process was based on location mapping between premises and monitoring stations. Effluent, sewage, livestock, palm oil mill, construction site, and market were identified as sources of pollution from the documented analysis. Criteria for determining data on pollution sources were based on the location of premises whereby the location of pollution source was in the upstream while the monitoring station was in the downstream.

The integration process will combine both selected attributes of water quality and pollution source data. Table 6 shows the list of data used in this study for finding the association between the source of pollution and water quality parameters.

TABLE 6. Data used for experimenting

No.	Attribute	Attribute type
1	DO	Numeric
2	BOD	Numeric
3	COD	Numeric
4	SS	Numeric
5	NH3N	Numeric
6	pH	Numeric
7	River	Nominal
8	WQI	Numeric
9	Station	Nominal
10	Effluent	Nominal
11	Sewage	Nominal
12	Livestock	Nominal
13	Palm oil mill	Nominal
14	Construction site	Nominal
15	Market	Nominal

Then, at data transformation phase, the data would be transformed into the appropriate Apriori algorithm format.

ALGORITHM DEVELOPMENT

Association rule mining (ARM) is a technique used to explore the relationship between items set out in the data based on minimum support and confidence (Agrawal & Srikant 1994; Agrawal et al. 1993). The relationship in the association rules consists of two components, namely antecedent and consequent. For example, consumer buying transactions such as the relationship between bread and milk (bread = > milk) in the supermarket. The bread is the antecedent and the consequent is milk. This relationship shows that milk will be purchased by customers buying bread. Apriori algorithm is one of Agrawal's (1994) association rule algorithms (Agrawal & Srikant 1994; Wu et al. 2008; Yabing 2013). It is a popular algorithm and the foundation of other association rule algorithms (Ali Othman et al. 2018; Nahar et al. 2013; Zhao & Sourav 2003).

Two measurements involving the generation of meaningful rules are minimum support (1) and minimum confidence values (2). A minimum support value is to obtain all the frequent itemsets, while a minimum confidence value is to get meaningful rules from the frequent itemsets. The formula a relationship between A and B (A=>B) is as follows:

$$Support\ (A => B) = \frac{The\ number\ of\ transactions\ containing\ A\ and\ B}{Total\ transaction} \tag{1}$$

$$Confidence\ (A => B) = \frac{The\ number\ of\ transactions\ containing\ A\ and\ B}{Total\ transaction\ A} \tag{2}$$

The Apriori algorithm processes the data based on minimum support and confidence values in finding any associations in the data. In this experiment, list support and confidence values were set to observe the number of rules generated by the algorithm. Table 7 shows the values of minimum support and minimum confidence.

TABLE 7. Information on minimum support and confidence values

Minimum value	
Support (%)	Confidence (%)
40	100
40	90
30	100
30	90
20	100
20	90

In this study, rules were selected based on a two-stage selection. The first stage of selection would select rules based on minimum support and confidence values. Moreover, priority would be given to the larger minimum support and confidence values, followed by lower minimum support and confidence values. For example, the minimum support value of 40% and the minimum confidence value of 100% were the highest priority in the selection, followed by the minimum support value of 40% and the minimum confidence value of 90%.

Meanwhile, the second stage of selection was to formulate a rule from the selected rules. Based on the preliminary process, it showed that the rules generated had similar antecedents but different consequences. Table 8 displays an example of how a formulated rule is formed based on three generated rules that have similar antecedents with different consequent.

TABLE 8. Example of rule formulation

No.	Generated rules	Antecedent	Consequent	Formulated rules
1	River=Langat 318 ==> Effluent=Yes 294	River= Langat 318	Effluent=Yes 294	River=Langat then pollution sources are effluent, livestock or sewage
2	River=Langat 318 ==> Livestock=Yes 318	River= Langat 318	Livestock=Yes 318	
3	River= Langat 318 ==> Sewage=Yes 294	River= Langat 318	Sewage=Yes 294	

Based on Table 8, the generated rules, numbers 1, 2, and 3, had a similar antecedent to 'River= Langat' with different consequents: 'Effluent= Yes', 'Livestock= Yes', and 'Sewage= Yes'. Therefore, the formulated rule was 'If the river is Langat, then the sources of pollution are effluent, livestock or sewage'. On the other hand, the rule could be rephrased as such that the sources of pollution from Sungai Langat are from effluent, planned waste or sewage.

EVALUATION

The formulated rules have to be evaluated by experts. In this study, there are two stages of rules evaluation. The first stage, the preliminary acceptance test, involved two water quality experts. The experts would verify and accept the list of formulated rules from the selection. Meanwhile, the second stage was a final validation and conducted by five water quality experts. The experts would evaluate a questionnaire that consisted of all the formulated rules. Based on their expertise, they would then validate these rules. The evaluation was made based on five levels of acceptance scales as shown in Table 9.

TABLE 9. Expert acceptance scales

Expert acceptance scales	Descriptions	Acceptance categories
1	Strongly disagree	Disagree
2	Disagree	
3	Slightly agree	
4	Agree	Agree
5	Strongly agree	

Referring to Table 9, the scales of expert acceptance for scales 1, 2, and 3 were classified as disagreeable, while scales of 4 and 5 were classified as agreeable.

RESULTS AND DISCUSSION

RULES ANALYSIS

Table 10 shows some rules generated by the Apriori algorithm. Based on the selected minimum support and minimum confidence values, 159 rules were generated by the algorithm.

TABLE 10. Number of generated rules by Apriori

Minimum value		Number of generated rules
Support (%)	Confidence (%)	
40	100	1
40	90	10
30	100	10
30	90	33
20	100	39
20	90	66
Total		159

Due to the large number of generated rules, the two-stage selection process was applied as mentioned in the previous section. High values of support and confidence would generate general rules; meanwhile, lower values of support and confidence would generate more specific rules. These rules would need to undergo the processes of selection and formulation. After the preliminary expert acceptance test, some formulated rules were not accepted by the experts. The preliminary expert acceptance test results showed that 62.5% of rules were not agreed upon by the experts. This is shown in Table 11.

TABLE 11. Preliminary expert acceptance test results

	Expert acceptance scale					Total
	Disagree			Agree		
	1	2	3	4	5	
Number of formulated rules	0	2	13	8	1	24

They disagreed with some relations between the source of pollution and water quality parameters. As per their comments and suggestions, scheduled waste produced by the industry premises was to be excluded as a source of pollution since it should be disposed of from the premise. Therefore, scheduled waste was excluded from the list of pollution sources. The selected rules were then regenerated and a list of the updated rules is shown in Table 12.

TABLE 12. List of selected rules

No	Minimum value		Selected rules
	Support (%)	Confidence (%)	
1	40	100	Not Selected
2	40	90	WQI=III 324 ==> Effluent=Yes 293
			WQI=III 324 ==> Sewage=Yes 308
			WQI=III 324 ==> Construction Site=Yes 304
			WQI=III 324 ==> Market=Yes 297
			River=LANGAT 318 ==> Effluent=Yes 294
			River=LANGAT 318 ==> Sewage=Yes 294
			Sungai=LANGAT 318 ==> Livestock=Yes 318
3	30	100	NH3N=0.9-2.7 176 ==> Sewage=Yes 176
			River=LANGAT pH=>7.0 182 ==> Livestock=Yes 182
4	30	90	COD=25-50 WQI=III 178 ==> Sewage=Yes 169
			COD=25-50 WQI=III 178 ==> Construction Site=Yes 166
			pH=>7.0 WQI=III 212 ==> Effluent=Yes 193
			pH=>7.0 WQI=III 212 ==> Sewage=Yes 201
			pH=>7.0 WQI=III 212 ==> Construction Site=Yes 198
			pH=>7.0 WQI=III 212 ==> Market=Yes 196
			BOD=6-12 WQI=III 219 ==> Sewage=Yes 205
			BOD=6-12 WQI=III 219 ==> Construction Site=Yes 201
			DO=5-7 231 ==> Sewage=Yes 215

5	20	100	River=LANGAT BOD=6-12 WQI=III 133 ==> Effluent=Yes 133
			River=LANGAT BOD=6-12 WQI=III 133 ==> Sewage=Yes 133
			River=LANGAT BOD=6-12 164 ==> Livestock=Yes 164
			River=LANGAT BOD=6-12 WQI=III 133 ==> Construction Site=Yes 133
			River=LANGAT BOD=6-12 WQI=III 133 ==>Market=Yes 133
			River=LANGAT DO=3-5 126 ==> Effluent=Yes 126
			River=LANGAT DO=3-5 126 ==> Sewage=Yes 126
			River=LANGAT DO=3-5 126 ==> Livestock=Yes 126
			River=LANGAT DO=3-5 126 ==> Construction Site=Yes 126
			River=LANGAT DO=3-5 126 ==> Market=Yes 126
			River=LANGAT NH3N=0.9-2.7 122 ==> Effluent=Yes 122
			River=LANGAT NH3N=0.9-2.7 122 ==> Sewage=Yes 122
			River=LANGAT NH3N=0.9-2.7 122 ==> Livestock=Yes 122
			River=LANGAT NH3N=0.9-2.7 122 ==> Construction Site=Yes 122
			River=LANGAT NH3N=0.9-2.7 122 ==> Market=Yes 122
6	20	90	River=LANGAT COD=25-50 120 ==> Effluent=Yes 117
			River=LANGAT COD=25-50 120 ==> Livestock=Yes 120
			River=LANGAT COD=25-50 120 ==> Construction Site=Yes 114
			River=LANGAT COD=25-50 120 ==> Market=Yes 114

The selected rules showed that some of the rules had different consequents for the same antecedent. For example, the following rules with minimum support 40% and confidence 90% had an antecedent of 'River = Langat' with different consequents, i.e. effluent, sewage, and livestock. Therefore, as discussed earlier, the second stage of selection was employed to formulate the rules. Table 13 shows the list of rules formulated from the selected rules.

TABLE 13. List of formulated rules

No	Minimum value		Formulated rules
	Support (%)	Confidence (%)	
1	40	100	None
2	40	90	River with WQI = III THEN source of pollution is effluent, sewage, construction site or market.
			River=Langat THEN source of pollution is effluent, sewage or livestock
3	30	100	NH3N = 0.9 – 2.7 THEN source of pollution is sewage
			River=Langat AND pH => 7 THEN source of pollution is livestock

4	30	90	COD = 25 – 50 AND WQI=III THEN source of pollution is sewage or construction site pH=>7 AND WQI=III THEN source of pollution is effluent, sewage, construction site or market BOD = 6 – 12 AND WQI=III THEN source of pollution is sewage or construction site DO = 5-7 THEN source of pollution is sewage
5	20	100	River=Langat AND BOD = 6 – 12 THEN source of pollution is effluent, sewage, livestock, construction site or market River=Langat AND DO = 3 - 5 THEN source of pollution is effluent, sewage, livestock, construction site or market River = Langat AND NH3N = 0.9 – 2.7 THEN source of pollution is effluent, sewage, livestock, construction site or market
6	20	90	River=Langat AND COD=25-50 THEN source of pollution is effluent, livestock, construction site or market

Table 14 shows the total number of rules after the preliminary expert acceptance test. A total of 127 rules were generated, and out of these rules, 37 were selected. From these selected rules, 12 rules were formulated. These 12 rules were then evaluated by the water quality experts in the final validation.

TABLE 14. Summary of number of rules after the Preliminary Expert Acceptance Test

Minimum value		Total number of generated rules	Total number of selected rules	Total number of formulated rules
Support (%)	Confidence (%)			
40	100	1	0	0
40	90	7	7	2
30	100	8	2	2
30	90	25	9	4
20	100	33	15	3
20	90	53	4	1
Total		127	37	12

FINAL EVALUATION

In the final evaluation, 65% of the rules were agreed upon by the experts. Meanwhile, 35% were not agreed on by them. This is shown in Table 15.

TABLE 15. Findings from experts' evaluation

Expert acceptance category	Expert acceptance scales	Total	%
Disagree	1	0	35
	2	2	
	3	19	
Agree	4	28	65
	5	11	
Total		60	100

Tables 16, 17, and 18 exhibit the details of the formulation of rules agreed by the water quality experts.

TABLE 16. Agreed by all five water quality experts

No.	Rules formulation
1	Pollution source of Sungai Langat is effluent, sewage or livestock
2	The NH_3-N is between 0.9 and 2.7 (NH_3-N = 0.9-2.7) due to pollution source from sewage

Based on Table 16, all five water quality experts agreed that either effluent, sewage or livestock was the source of pollution for Sungai Langat. To avoid declining water quality, Sungai Langat's effluent, sewage or livestock pollution source should be addressed comprehensively. In addition, all experts agreed that the NH_3-N parameter in Sungai Langat Basin was between 0.9 and 2.7 due to the source of pollution from sewage. In Class IV, the NH_3-N parameter is based on the classification of water quality that needs to be concerned. Sungai Langat's water quality is generally Class III, but specifically from this finding, the NH_3-N parameter was Class IV. Thus, if no precautionary action is taken, there is a probability that the water quality in Sungai Langat will decline to Class IV in the future. Furthermore, the findings from the experts' assessment showed that the association rules method was capable of discovering water quality domain knowledge.

TABLE 17. Agreed by four out of five water quality experts

No.	Rules formulation
1	The river has WQI is class III then the pollution source is effluent, sewage, construction site or market
2	Sungai Langat has COD is between 25 and 50 due to pollution source is effluent, livestock, construction site or market

Referring to Table 17, four out of five water quality experts agreed that for the river with WQI Class III, the pollution source was either effluent, sewage, construction site or market. This study also showed that Sungai Langat with COD between 25 and 50 had a source of pollution of either effluent, livestock, construction site or market. Therefore, the identified sources of pollution need to be addressed to avoid a future decline in water quality.

TABLE 18. Agreed by three out of five water quality experts

No.	Rules formulation
1	COD = 25 - 50 and WQI = III then the pollution source is sewage or construction site
2	DO = 5-7 then pollution source is sewage
3	River = Langat and BOD = 6 - 12 then pollution source is effluent, sewage, livestock, construction sites or markets
4	River = Langat and DO = 3 - 5 then pollution source is effluent, sewage, livestock, construction site or market
5	River = Langat and NH_3-N = 0.9 - 2.7 then pollution source is effluent, sewage, livestock, construction site or market

Referring to Table 18, three out of five water quality experts agreed that the river had WQI Class III and a COD value of 25 to 50 as a result of either sewage or construction site as the pollution source. Besides, the DO value between 5 and 7 was due to sewage as the source of pollution.

Conclusion

The association rules method has been used in this research to analyse Sungai Langat Basin water quality data. The method shows that either effluent, sewage or livestock is the main source of pollution of Sungai Langat. Furthermore, in Sungai Langat Basin, the river was classified as Class III due to the source of pollution from effluent, sewage, construction site or market. In the future, a particularly new development project needs to be addressed comprehensively to the identified pollution sources. This is to ensure that the water quality in Sungai Langat Basin is not reduced by the pollution sources.

Finding from this research shows that the NH_3-N parameter in Sungai Langat Basin is between 0.9 and 2.7 due to the sewage pollution source. In Class IV, the NH_3-N parameter is based on the classification of water quality that needs to be concerned to avoid future declines in water quality. Furthermore, Sungai Langat has a COD value of between 25 and 50 from effluent, livestock, construction site or market as sources of pollution. Furthermore, sources of pollution will affect water quality parameters such as NH_3-N and COD.

Illegal dumping by premises around Sungai Langat Basin has also resulted in decreasing water quality. In this regard, to tackle illegal dumping, cooperation between DOE and local authorities needs to be streamlined. In Sungai Langat Basin, the proposed future study is to predict illegal dumping. This will alert DOE of illegal dumping by implementing early inspection and enforcement action in the suspected area.

Acknowledgements

The authors would like to thank Universiti Kebangsaan Malaysia for funding this work under the GGP-2017-025 grant, and the Malaysian Department of Environment for providing water quality and pollution source data on Sungai Langat Basin.

References

Abidin, M.Z., Kutty, A.A., Lihan, T. & Zakaria, N.A. 2018. Hydrological change effects on Sungai Langat water quality. *Sains Malaysiana* 47(7): 1401-1411.

Agrawal, R. & Srikant, R. 1994. Fast algorithms for mining association. *Proceedings of the 20th International Conference on Very Large Data Bases.* pp. 487-499.

Agrawal, R., Road, H. & Jose, S. 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD Conference.* pp. 1-11.

Ahmad, S., Kutty, A.A., Raji, F. & Saimy, I.S. 2015. Water quality classification based on water quality index in Sungai Langat, Selangor, Malaysia. *Jurnal Teknologi* 77(30): 139-144.

Al-Badaii, F., Abdul Halim, A. & Shuhaimi-Othman, M. 2016. Evaluation of dissolved heavy metals in water of the Sungai Semenyih (Peninsular Malaysia) using environmetric methods. *Sains Malaysiana* 45(6): 841-852.

Anon. 2016. Factories along Sungai Langat and Sungai Semantan Caused Water Treatment Plants to Shut Down. *The Star Online*, 8 November.

Chan, N.W. 2012. Managing urban rivers and water quality in Malaysia for sustainable water resources. *International Journal of Water Resources Development* 28(2): 343-354.

Department of Environment (DOE). 2016. *Environmental Quality Report 2016*.

Department of Environment (DOE). 2015. *Environmental Quality Report 2015*.

Department of Environment (DOE). 2012. *Environmental Quality Report 2012*.

Fahrudin, T.M., Syarif, I. & Barakbah, A.R. 2017. Discovering patterns of NED-breast cancer based on association rules using apriori and FP-growth. *International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC).* pp. 132-139.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence* 9078(3): 637-648.

Gandhi, N. & Armstrong, L.J. 2016. Assessing impact of seasonal rainfall on rice crop yield of Rajasthan, India using Association Rule Mining. *International Conference on Advances in Computing, Communications and Informatics (ICACCI).* pp. 1021-1024.

Gasim, M.B., Zakaria, N., Umar, R. & Mustafa, A.D. 2015. Analisis kualiti air fiziko-kimia dan kandungan mikrob di Hulu Sungai Langat, Selangor. *Malaysian Journal of Analytical Sciences* 19(5): 1072-1083.

Gasim, M.B., Jamil, M., Rahim, S.A. & Toriman, M.E. 2009. Water-quality assessment of the Langat River at Kilometre 7, Jalan Kajang-Bangi, Selangor, Malaysia. *The Arab World Geographer* 12(3-4): 188-198.

Gour, S., Jaloree, S. & Gour, M. 2016. Water Quality assessment using association rule mining for River Narmada. *Indian Journal of Science and Technology* 9(10): 1-5.

Harun, N.A., Makhtar, M., Abd Aziz, A., Zakaria, Z.A., Abdullah, F.S. & Jusoh, J.A. 2017. The application of apriori algorithm in predicting flood areas. *International Journal on Advanced Science, Engineering and Information Technology* 7(3): 763-769.

Juahir, H., M. Zain, S., Yusoff, M.K., Tengku Hanidza, T.I., Armi, A.S.M., Toriman, M.E. & Mokhtar, M. 2011. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment* 173(1-4): 625-641.

Khairul, A.M., Juhari, M.A. & Ibrahim, A. 2000. GIS aided groundwater potential mapping of the Langat Basin. *Geological Society of Malaysia Annual Geological Conference 2000.* pp. 405-410.

Liang, T.Y. 2015. Sungai Semenyih Plant Resumes Operations, Water Supply to Return in Stages. *The Star Online*, 3 March.

Malaysia. 2005. *Environmental Quality (Scheduled Wastes) Regulations 2005*.

Mazlinda, M. 2015. High Level of manganese, ammonia forces closure of Sg Semenyih water treatment plant. *New Straits Times*, 3 March.

Nahar, J., Imam, T., S.Tickle, K. & Chen, Y.P.P. 2013. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications* 40(4): 1086-1093.

Othman, Z.A., Ismail, N., Bakar, A.A., Latif, M.T. & Abdullah, S.M.S. 2018. Rules discovery of high ozone in Klang areas using data mining approach. *International Journal on Advanced Science, Engineering and Information Technology* 8(6): 2683-2689.

Selangor Town and Country Planning Department (STCPD). 2012. *Laporan Tinjauan Kajian Rancangan Struktur Negeri Selangor 2035: Bekalan Air*.

Wu, B., Zhang, D., Lan, Q. & Zheng, J. 2008. An efficient frequent patterns mining algorithm based on Apriori algorithm and the FP-tree structure. *Third 2008 International Conference on Convergence and Hybrid Information Technology*, South Korea.

Yabing, J. 2013. Research of an improved apriori algorithm in data mining association rules. *International Journal of Computer and Communication Engineering* 2(1): 25-27.

Zhao, Q. & Bhowmick, S.S. 2003. *Association Rule Mining: A Survey.* Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116.

Zubir, M.R.M., Osman, R. & Saim, N. 2016. Chemometric analysis of selected organic contaminants in surface water of Langat River basin. *Malaysian Journal of Analytical Sciences* 20(2): 278-287.

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: zalinda@ukm.edu.my

Received: 18 September 2019
Accepted: 6 May 2020