ISSN (Print): 2077-7973
ISSN (Online): 2077-8767
DOI: 10.6977/IJoSI.202109_6(5)

International Journal of Systematic Innovation

VOL. 06 NO. 05
September, 2021

Published by the Society of Systematic Innovation

Opportunity Identification
&
Problem Solving

IJoSI

ISSN (Print): 2077-7973
ISSN (Online): 2077-8767
DOI: 10.6977/IJoSI.202109_6(5)

The International Journal of Systematic Innovation

Publisher:

The Society of Systematic Innovation

Editorial Office:

The International Journal of Systematic Innovation

5F, # 350, Sec. 2, Guanfu Rd, Hsinchu,
Taiwan, R.O.C., 30071

e-mail:editor@systematic-innovation.org

web site: http://www.IJoSI.org

IJoSI

INTERNATIONAL JOURNAL OF SYSTEMATIC INNOVATION

CONTENTS

SEPTEMBER 2021 VOLUME 6 ISSUE 5

FULL PAPERS

A Hybrid of Heuristic Orderings and Variable Neighbourhood Descent for a Real Life University Course Timetabling Problem

…..……………………..……Mei Ching Chen, San Nah Sze, Say Leng Goh and Sei Ping Lau **1-10**

Medium-Term Wind Speed Prediction using Bayesian Neural Network (BNN)

……... Sana Mohsin, Sofia Najwa Ramli and Maria Imdad **11-20**

Designing a Real-time Interactive Spatial Augmented Reality Platform

……………………………………… Aye Chan Zay Hta, YunLi Lee1 and Wai Chong Chia **21-30**

Hybrid Black Widow Optimization and Variable Neighborhood Descent Algorithm for Traveling Salesman Problem

……………………………………… Ayad Mohammed Jabbar, Ku Ruhana Ku-Mahamud **32-43**

Finite mixture of Burr type XII for bus travel time in Klang Valley

………………………………Wooi Chen Khoo, Cynthia C. T. Cheok and Hooi Ling Khoo **44-54**

A Text Analytics Approach to Study Python Questions Posted on Stack Overflow

……………………Lee Yong Meng, Soo Yin Yi, Gan Keng Hoon and Nur-Hana Samsudin **55-66**

A Hybrid of Heuristic Orderings and Variable Neighbourhood Descent for a Real Life University Course Timetabling Problem

Mei Ching Chen[1], San Nah Sze[1]*, Say Leng Goh[2] and Sei Ping Lau[1]

[1]Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, Jalan Datuk Mohammad Musa, 94300 Kota Samarahan, Sarawak, Malaysia

[2]Optimisation Research Group, Faculty of Computing & Informatics, Universiti Malaysia Sabah Kampus Antarabangsa Labuan, Jln Sungai Pagar, 87000 Labuan, Malaysia

* Corresponding author E-mail: snsze@unimas.my

(Received 8 July 2021; Final version received 22 July 2021; Accepted 7 September 2021.)

Abstract

Academic institutions face timetabling problem every semester. Addressing timetabling problem at academic institutions is a challenging combinatorial optimisation task both in theory and practice. This is due to the size of the problem instances as well as the number of constraints that must be satisfied. Over the years, timetabling problem has attracted many researchers in proposing ways to find an optimal solution. In this paper, we investigate a hybrid of heuristic orderings and variable neighbourhood descent approach in tackling course timetabling problem at the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS). At FCSIT, some events of 4 lecture hours are not evenly spread over minimum working days and some events are conducted until 9 pm. The objectives of the study are to shorten the daily lecture hours and evenly distribute events' lecture. In stage 1, heuristic orderings are utilised to find a feasible solution. In stage 2, a hybrid of heuristic orderings and variable neighbourhood descent approach are utilised to improve the quality of the solution. The proposed algorithm is tested on real-world data instances (semesters 1 and 2 of 2019/2020) of FCSIT, UNIMAS. Results show that certain heuristic ordering (largest degree or the combination of largest degree and largest enrolment) are better than others in generating a feasible solution. In addition, the number of timeslots required by heuristic ordering are less compared to that required by the existing timetabling software. In stage 2, the proposed algorithm manages to achieve soft constraint violations of 0 and 1 for instances for semesters 1 and 2, respectively. However, all HO manage to achieve 0 violation for both instances when the proposed algorithm is executed 30 times. Each neighbourhood structures defined in this study contributes to lowering the soft constraint violations thus ensuring a high-quality timetable. Results show that the order of neighbourhood structures do impact the number of soft constraint (SC1) violations achieved.

Keywords: combinatorial optimisation, course timetabling problem, heuristic orderings, hybrid, variable neighbourhood descent

1. Introduction

Educational timetabling is defined as a task of allocating events such as exams, subjects and courses to rooms and timeslots by fulfilling certain constraints (Tan et al., 2021; Thepphakorn & Pongcharoen, 2020; Tan et al., 2020; Assi et al., 2018). Timetabling is a challenging combinatorial optimisation problem in theory and practice (Schaerf, 1999). Universiti Malaysia Sarawak (UNIMAS) devotes a significant number of resources in developing a feasible and high-quality course scheduleor each faculty. Efficient allocation of courses may result in more effective use of valuable resources (Burke et al., 2005). Therefore, it is crucial to find an optimal configurations for the variables defined to achieve specific objectives (Habashi et al., 2018).

University course timetabling problem (UCTTP) involves allocating a set of courses to limited resources namely lecturers, venues and timeslots by fulfilling certain constraints (Goh et al., 2020; Goh et al., 2019;

Erdeniz & Felfernig, 2018; Goh et al., 2017). UCTTP can be divided into two different categories based on problem settings and requirements, namely curriculum-based course timetabling problem (CBCTTP) and post-enrolment course timetabling problem (PECTTP). UCTTP in UNIMAS is closely related to CBCTTP. Constraints can be classified into two types namely hard and soft. The fulfilment of hard constraints is mandatory in generating a feasible timetable. Meanwhile, the fulfilment of soft constraints is optional but will determine the quality of the timetable generated.

To date, there are many papers on UCTTP either tackling benchmark or real-world UCTTP. For most real-world search problems, automatically generating high-quality solutions is a difficult challenge (Muklason et al., 2019). The objective is to find a feasible timetable with the lowest possible soft constraint violations. Furthermore, the requirements of UCTTP differ across academic institutions as policies and regulations are unique in each institution. This paper is addressing UCTTP at the Faculty of Computer Science and Information Technology (FCSIT), UNIMAS using real-world dataset. We investigate the performance of the hybrid of heuristic ordering (HO) and variable neighbourhood descent (VND). We also compare its performance against the existing timetable which was constructed using commercial timetabling software.

The structure of this paper is as follows. Next section presents the related work on HO and VND. We describe the UCTTP at UNIMAS in Section 3. The proposed algorithm is presented in Section 4. Section 5 presents the numerical results of the research. Finally, conclusions are presented in section 6.

2. Related work

A variety of approaches have been proposed in solving UCTTP. Babaei et al. (2015) had categorized the approaches into five, namely operational research (OR) based techniques, metaheuristic approaches, multi criteria/ objective approaches, intelligent novel approaches and distributed multi agent systems approaches. Each approach has its own advantages. In order to take advantage of each approach, researchers have proposed hybrid approaches in solving UCTTP. Among the hybrid approaches are Hybrid Genetic Algorithm (Akkan & Gülcü, 2018; Matias et al., 2019) and combination of VNS and Tabu Search (Vianna et al., 2020).

VNS is used to solve combinatorial optimisation problem in two phases, namely descent phase and perturbation phase (Hansen et al., 2018). Descent phase helps to achieve local optimum whereas perturbation phase helps to escape from local optimum. VNS is well known for its ability in avoiding traps (local optimum) by considering different neighbourhood structures (Hansen & Mladenovi´c, 2014). Its success has been proven in a wide range of applications with large instances and challenging number of constraints (hard and soft) (Hansen et al., 2018).

Variable Neighbourhood Descent (VND) method was proposed by Borchani et al. (2017) to solve UCTTP for Faculty of Economics and Management Sciences of Sfax in Tunisia. The authors aimed to minimize the total number of holes and the number of isolated lessons. Neighbourhood structures proposed by authors were implemented using simple move. Six real datasets were used as testbeds. Results showed that the proposed algorithm was able to eliminate 52.47% of holes and isolated lessons.

Heuristic ordering (HO) is derived from graph colouring heuristics such as largest degree (LD), saturation degree (SD), largest weighted degree (LWD) and colour degree (CD) (Burke & Petrovic, 2002). In LD, the event with the largest number of conflicts/clashes with other events are assigned first because it is hard to find a valid timeslot for an event that has many conflicts/clashes with other events. LWD associates the number of students with the conflicted events. Therefore, the event with largest number of students is assigned first. In SD, the event with the least number of valid timeslots will be selected for assignment. The valid timeslots for the remaining events are updated in each iteration. Meanwhile, CD takes into consideration the conflict between events to be scheduled with the scheduled events. Priority is given to events with the largest number of conflicts with the scheduled events. These heuristics play an important role in generating initial solutions which quality would then be improved by other methods (Pillay & Özcan, 2019).

Vianna et al. (2020) proposed a hybrid of Variable Neighbourhood Search (VNS) and Tabu Search (TS) in tackling the UCTTP for Federal Fluminense University. Framework for the Implementation of metaheuristics based on Neighbourhood Structure Search (FINESS) framework was used in developing the proposed algorithm which enabled constraints to be added and removed easily. The datasets used in their work were obtained from two undergraduate courses. Results showed that the hybrid produced better solutions than those produced using VNS and TS separately.

Muklason et al. (2019) proposed a Tabu-Variable Neighborhood Search based Hyper-Heuristic algorithm in addressing the UCTTP for the Department of Information Systems, Institut Teknologi Sepuluh Nopember, Indonesia. This approach does not require parameter tuning as required in metaheuristic approaches such as simulated annealing. The algorithm was tested using two real-world datasets from 2017/2018 session. The solution obtained was better in terms of quality compared to the one created manually.

3. Problem description

UNIMAS is one of the public universities in Malaysia established on 24 December 1992. It has 10 faculties offering more than 90 programmes. The timetabling problem in this study is based on the real-world scenario at the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS).

All this while, each faculty's administrator/timetable planner in UNIMAS constructs course timetable based on curriculum (as information on course pre-registration is not available) manually. They started utilising commercial timetabling software in 2014. In this study, we focus in timetable at FCSIT. Courses offered by FCSIT can be divided into a few categories, namely lecture, lecture with tutorial and lecture with lab. These courses are ranged from 2 to 4 credit hours. The credit hour indicates the number of lecture hours per week for a course. It is recommended to split long lecture hours (4 credit hours course) in 2 days. For example, 2 hours on Monday and another 2 hours on Wednesday, which can be represented as "2+2".

In term of venue, FCSIT conducts lectures at either its own venue (available all the times) or shared venue (only available at certain times). Sharing of venues is a common feature showcased by most academic institutions especially if the venue can accommodate many students. Table 1 shows the capacity of the shared and fixed venues.

Table 1 Teaching venues and its capacity

Feature	Usage	Venue	Quantity	Capacity
Shared	Limited	DK	Vary from semester to semester	500
		BS	Vary from semester to semester	150
Fixed (Faculty)	All the time	TMM	1	120
		MM2	1	100
		ARTLNT	1	80
		ISLAB	1	80
		MM1	1	80
		TL1	1	80
		TL2	1	80
		CSLAB	1	60
		NETLAB1	1	60
		NETLAB2	1	40
		TR	8	40

Table 2 shows the timeslots used in this study. Gray area indicates that the timeslots are blocked. No assignment of faculty courses on these timeslots are allowed. Therefore, only 31 timeslots are allocated for the courses to the latest 5pm for Monday, Tuesday and Thursday, and 12pm for Friday. Table 3 shows the data instances used as testbeds for the algorithm proposed. In this study, all individual courses are referred as events.

Table 2 Timeslots

Day\Time	0800-0900	0900-1000	1000-1100	1100-1200	1200-1300	1300-1400	1400-1500	1500-1600	1600-1700
Monday	01	02	03	04	05	06	07	08	09
Tuesday	10	11	12	13	14	15	16	17	18
Wednesday									
Thursday	19	20	21	22	23	24	25	26	27
Friday	28	29	30	31					

Table 3 FCSIT data instances for academic years 2019/2020

Instance	Events	Rooms	Students	Timeslot requirement	Event enrolment
Semester 1 2019/2020	102	19 (fixed) 4 (shared)	1397	31	5073
Semester 2 2019/2020	77	18 (fixed) 2 (shared)	1040	31	3394

The constraints considered are listed below:

Hard constraints

HC1: Lectures taught by the same lecturer cannot be conducted in the same timeslot.

HC2: Only one lecture can be assigned to a venue at a specific timeslot.

HC3: A room assigned to a lecture must be big enough to accommodate the number of students.
HC4: Lectures for all events must be scheduled.
HC5: Blocked timeslots for lectures must be taken into considerations.
HC6: A student can only attend one lecture at a specific timeslot.

Soft constraints:
SC1: Events with 4 lecture hours are evenly spread over minimum working days.

4. Proposed algorithm

In this study, a two-stage heuristic algorithm is proposed. In stage 1, HO (LD) in descending order is utilised to generate a feasible solution by ensuring all hard constraints are satisfied. In stage 2, a hybrid of HO Largest Enrolment (LE) in descending order and VND is proposed to improve the quality of the solution by satisfying soft constraint as much as possible. This proposed algorithm consolidates the features of both HO and VND, which is not attempted in the existing literature reviews. Fig. 1 shows the general framework for solving UCTTP at the FCSIT, UNIMAS.

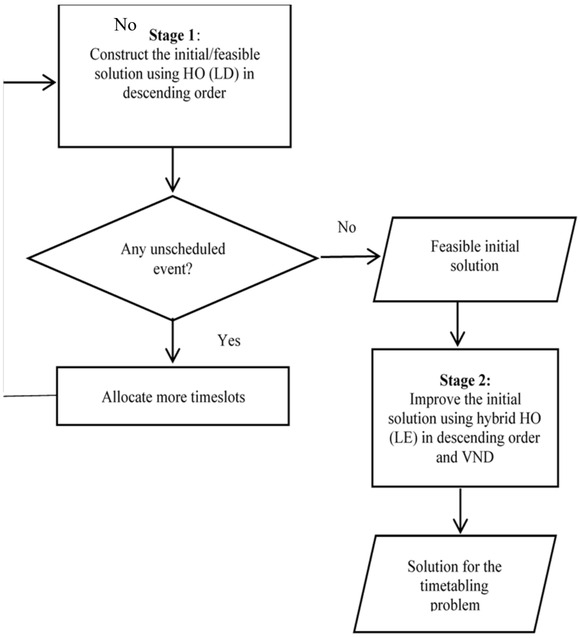

No
Stage 1: Construct the initial/feasible solution using HO (LD) in descending order
Any unscheduled event?
No
Yes
Feasible initial solution
Allocate more timeslots
Stage 2: Improve the initial solution using hybrid HO (LE) in descending order and VND
Solution for the timetabling problem

Fig. 1 General framework for solving UCTTP at FCSIT, UNIMAS

In stage 1, HO (LD) in descending order is used for event selection. In LD, the event with the largest number of conflicts/clashes with other events is assigned first. If there is any unscheduled event, more timeslots will be allocated, and stage 1 is repeated to generate a feasible initial solution.

In stage 2, a hybrid of HO (LE) in descending order and variable neighbourhood descent (VND) is used to minimise soft constraint violations. VND is known as best improvement local search. Sequential VND is used where the algorithm will walk through all the neighbourhood structures (NS) in a sequential order. It will start with the first NS and continue with the next one sequentially. Fig. 2 shows the details of this hybrid algorithm. *k* is initialised to 1. The algorithm starts with a feasible initial solution obtained from stage 1. *orderedEvents* is a list of events ordered based on HO (LE) in descending order. For each event in *orderedEvents*, we search the timeslots and venues sequentially until a feasible *candidateSolution* is found. Once it is found, the values of *f(candidateSolution)* and *f(currentSolutiom)* are compared. If the value of *f(candidatesolution)* is less than the value of *f(currentsolution)*, then the *candidatesolution* will be set as the *currentsolution*. Then, the next event in the *orderedEvents* will be considered. Otherwise, if the value of *f(candidatesolution)* is greater than or equal the value of *f(currentsolution)*, then the search to find the next feasible *candidateSolution* will continue. If no feasible *candidateSolution* can be found, the next event in the *orderedEvents* will be considered.

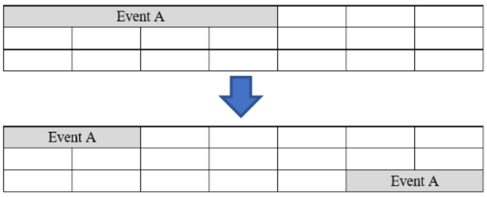
```
PROCEDURE variable neighbourhood descent

Input neighbourhood structures $N_k$, $k$=1,2,3,4,5
$k$ ← 1
currentSolution ← initialSolution //initial solution is obtained from stage 1
orderedEvents ← events ordered based on HO

REPEAT
   FOR each e in orderedEvents
      FOR each timeslot
         FOR each venue
            IF feasible (e, timeslot, venue, $N_k$)
               candidateSolution ← move (e, timeslot, venue, $N_k$)
               IF f(candidateSolution) < f(currentSolution) THEN
                  currentSolution ← candidateSolution
                  moved←true;
               END IF
            END IF
            IF moved=true
               Break;
            END IF
         END FOR

         IF moved=true
            Break;
         END IF
      END FOR
   END FOR

      $k$=$k$+1
   UNTIL $k$=5

END PROCEDURE
```

Fig. 2 Hybrid of HO and VND algorithm

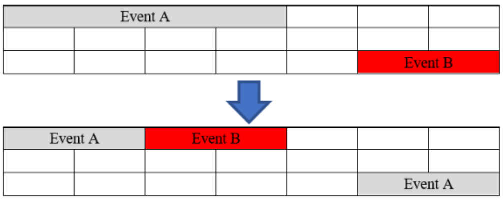
Fig. 3 illustrates the neighbourhood structures (NS) adopted in the proposed algorithm:

- Neighbourhood structure 1 (NS1): attempts to split a course with 4 continuous lecture hours by moving two of its lecture hours to other timeslots.
- Neighbourhood structure 2 (NS2): attempts to split a course with 4 continuous lecture hours by swapping two of its lecture hours with another course with 2 lecture hours.

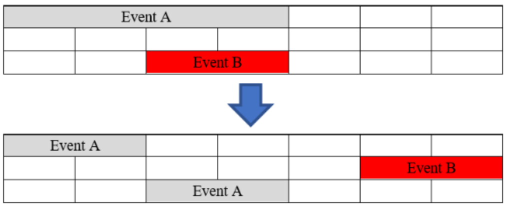
NS1: Split a course with 4 continuous lecture hours by moving two of its lecture hours to other timeslots.



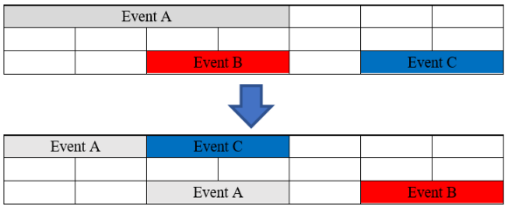
NS2: Split a course with 4 continuous lecture hours by swapping two of its lecture hours with another course with 2 lecture hours.



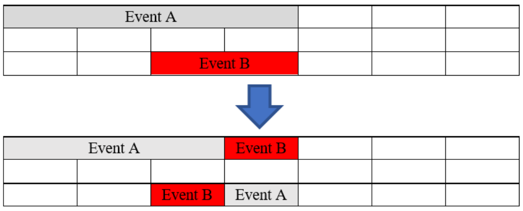
NS3: Split a course with 4 continuous lecture hours by executing 2 moves involving another course.



NS4: Split a course with 4 continuous lecture hours by executing 2 swaps involving 2 other courses.



NS5: Split a course with 4 continuous lecture hours by swapping one of its lecture hours with another course or by moving one lecture hour to other timeslot.



**Note: Column – timeslot, Row - venue*

Fig. 3 Neighbourhood structures: NS1 to NS5

- Neighbourhood structure 3 (NS3): attempts to split a course with 4 continuous lecture hours by executing 2 moves involving another course.
- Neighbourhood structure 4 (NS4): attempts to split a course with 4 continuous lecture hours by executing 2 swaps involving 2 other courses.
- Neighbourhood structure 5 (NS5): attempts to split a course with 4 continuous lecture hours by swapping one of its lecture hours with another course or by moving one lecture hour to other timeslot.

NS3 and NS4 are new neighbourhood structures introduced and included in the proposed algorithm to further improve the quality of the timetable. As shown in Fig. 3, Event A is selected from a list ordered by HO (LE) in descending order. Whereas Event B and Event C are selected when the timeslots and venues are scanned sequentially. The five different NS are used to improve the connectivity of the search space and therefore the quality of the solution. If the resulting solution from applying the NS is feasible (not breaching any hard constraints), it is returned as a *candidateSolution* and evaluated for acceptance.

5. Numerical result

The algorithms are coded using visual basic (VB.Net). We use Microsoft Access as the database management software. Table 4 shows the distance to feasibility (number of unallocated courses) for initial solutions generated using different HO in stage 1. In LD, the event with the largest number of conflicts/clashes with other events is assigned first. Whereas in LE, the event with the largest number of enrolments is assigned first. In (LD+LE), both LD and LE are taken into considerations when allocating events to timetable.

Table 4 Distance to feasibility (number of unallocated courses) for initial solutions generated using different HO in stage 1. *N* = 30 runs.

HO	Instance					
	Semester 1 (31 timeslots)		*Semester 1 (35 timeslots)*		*Semester 2 (31 timeslots)*	
	Best	Average	Best	Average	Best	Average
LD Ascending	11	13.43	6	8.87	5	7.23
LD Descending	1	2.03	0**	0.27	0**	0.43
LE Ascending	12	14.20	7	8.10	6	7.47
LE Descending	5	5.80	1	1.73	1	1.00
(LD + LE) Ascending	12	15.80	8	11.07	6	7.13
(LD + LE) Descending	2	3.93	0	0.60	0	0.27
Random	8	8.00	5	5.00	3	3.00

Note: ** Selected HO ordering in stage 1 (used as feasible initial solution in stage 2)

As shown in Table 4, both LD and (LD + LE) in descending order manage to find feasible solutions for semester 2's instance using 31 timeslots. However, they failed to do so for semester 1's instance using the same number of timeslots. This is because the instance for semester 1 is larger than that of semester 2 in terms of events, students and course enrolment. There are 77 events (282 lecture hours) for semester 2, compared to 102 events (347 lecture hours) for semester 1. Furthermore, FCSIT has limited timeslots, since Wednesday and Friday afternoons are blocked. As the size of the data instance grows larger, this makes allocating lecture hours a challenging task. Nevertheless, the algorithm manages to find feasible solution for semester 1's instance when the number of allocated timeslots is increased to 35 (6 pm). In a comparison, the solution generated by existing timetabling software required 48 timeslots (9 pm) to achieve feasibility.

Table 5 shows the number of timeslots required by the existing UNIMAS timetabling software in obtaining a feasible solution for semester 1's instance. A total of 48 timeslots required. As shown, some of the lectures are conducted until 9 pm. This will consume extra resources such as electricity cost. One the other hand, Table 6 shows the timeslots required by our approach in generating a feasible solution for the same instance. A total of 35 timeslots required, where the latest lectures end at 6pm. Comparatively, there are only 3 timeslots compared to 12 timeslots from existing timetable (Table 5) are scheduled after 5pm.

Table 7 shows the number of soft constraint (SC1) violations of the proposed VND algorithm with different HO for semester 1's instance. From the table, the lowest number of soft constraint (SC1) violations achieved is 0 using LE Ascending, LD Descending, LE Descending, (LD+LE) Descending and random ordering. The number 0 indicates that all the courses can be spread over minimum working days (2 days). Note that the number of allocated timeslots is 35. Each NS defined in this study contributes to lowering the soft constraint violations thus ensuring a higher-quality timetable.

Table 8 shows the number of soft constraint (SC1) violations of the proposed VND algorithm with different HO for semester 2's instance. From the table, the lowest number of soft constraint (SC1) violations achieved is 1 using LE Ascending, LD Descending, LE Descending, (LD+LE) Descending and random ordering. The number 1 indicates that there is one course which cannot be spread over minimum working days.

Table 5 The number of timeslots required by the existing timetabling software (semester 1's instance).

Day\Time	0800 - 0900	0900-1000	1000-1100	1100-1200	1200-1300	1300-1400	1400-1500	1500-1600	1600-1700	1700-1800	1800-1900	1900-2000	2000-2100
Monday	01	02	03	04	05	06	07	08	09	10	11	12	13
Tuesday	14	15	16	17	18	19	20	21	22	23	24	25	26
Wednesday													
Thursday	27	28	29	30	31	32	33	34	35	36	37	38	39
Friday	40	41	42	43	44	45	46	47	48				

Table 6 The number of timeslots required by our approach (semester 1's instance).

Day\Time	0800-0900	0900-1000	1000-1100	1100-1200	1200-1300	1300-1400	1400-1500	1500-1600	1600-1700	1700-1800
Monday	01	02	03	04	05	06	07	08	09	10
Tuesday	11	12	13	14	15	16	17	18	19	20
Wednesday										
Thursday	21	22	23	24	25	26	27	28	29	30
Friday	31	32	33	34	35					

Table 7 The number of soft constraint (SC1) violations of the proposed VND algorithm with different HO (semester 1's instance) with 35 timeslots.

	LD Ascending	LE Ascending	(LD+LE) Ascending	LD Descending	LE Descending	(LD+LE) Descending	Random
Initial solution	32	32	32	32	32	32	32
NS1	6	5	6	5	5	5	6
NS1 + NS2	4	3	4	3	3	3	5
NS1 + NS2 + NS3	3	1	3	3	3	3	3
NS1 + NS2 + NS3 + NS4	3	1	3	2	3	2	3
NS1 + NS2 + NS3 + NS4 + NS5	1	0	1	0	0	0	0

It is hard to spread a course with many students when; 1) the number of venues (high seating capacity) that can fit the students is limited, 2) the timetable is tight (as not many vacant places are available, and it is difficult to satisfy the conflict requirement).

In order to achieve 0 soft constraint (SC1) violations, the number of allocated timeslots needs to be increased to 32. As shown in Table 9, the lowest number of soft constraint (SC1) violations achieved is 0 which is achieved using LE Ascending, LE Descending and (LD+LE) Descending. By increasing the number of allocated timeslots, more venues (high seating capacity) that can accommodate large number of students, are made available. This increases the chances of a course with many students being spread over minimum working days.

Table 10 shows the number of soft constraint (SC1) violations of the proposed VND algorithm with different HO for semester 2's instance with 31 timeslots when NS is applied in different order. From the table, the lowest number of soft constraint (SC1) violations achieved is 0 by using LD Ascending. This shows the order of NS is one of the parameters which will impact the number of soft constraint (SC1) violations achieved in this study.

In further analysis, the proposed algorithm is executed 30 times for both instances (semesters 1 and 2). The aim is to find the best and average of soft constraint (SC1) violations. Each run uses different initial solution generated from LD Descending (stage 1). As shown in Table 11, all HO manage to achieve 0 violation for both instances.

Table 8 The number of soft constraint (SC1) violations of the proposed VND algorithm with different HO (semester 2's instance) with 31 timeslots.

	LD Ascending	LE Ascending	(LD+LE) Ascending	LD Descending	LE Descending	(LD+LE) Descending	Random
Initial solution	38	38	38	38	38	38	38
NS1	11	12	10	8	8	8	10
NS1 + NS2	7	8	9	8	8	8	7
NS1 + NS2 + NS3	6	5	7	5	5	5	7
NS1 + NS2 + NS3 + NS4	5	5	5	5	5	5	6
NS1 + NS2 + NS3 + NS4 + NS5	2	1	2	1	1	1	1

Table 9 The number of soft constraint (SC1) violations of the proposed VND algorithm with different HO (semester 2's instance) after the number of allocated timeslots is increased to 32.

	LD Ascending	LE Ascending	(LD+LE) Ascending	LD Descending	LE Descending	(LD+LE) Descending	Random
Initial solution	39	39	39	39	39	39	39
NS1	9	12	8	8	8	8	8
NS1 + NS2	6	8	7	8	8	8	7
NS1 + NS2 + NS3	5	5	5	5	5	5	7
NS1 + NS2 + NS3 + NS4	4	5	4	5	5	5	6
NS1 + NS2 + NS3 + NS4 + NS5	1	0	1	1	0	0	1

Table 10 The number of soft constraint (SC1) violations of the proposed VND algorithm with different HO (semester 2's instance) with 31 timeslots when NS are applied in different order.

	LD Ascending	LE Ascending	(LD+LE) Ascending	LD Descending	LE Descending	(LD+LE) Descending	Random
Initial solution	38	38	38	38	38	38	38
NS2	6	8	8	5	5	7	6
NS2 + NS4	3	6	6	3	3	3	6
NS2 + NS4 + NS1	3	6	6	3	3	3	6
NS2 + NS4 + NS1 + NS3	3	5	6	3	3	3	4
NS2 + NS4 + NS1 + NS3 + NS5	0	2	2	1	1	1	2

Table 11 The number of soft constraint (SC1) violations of the proposed VND algorithm with different HO. *N*= 30 runs.

HO	Instance			
	Semester 1 *(35 timeslots)*		*Semester 2* *(31 timeslots)*	
	Best	Average	Best	Average
LD Ascending	0	1.13	0	1.53
LD Descending	0	0.63	0	1.07
LE Ascending	0	1.23	0	1.57
LE Descending	0	0.40	0	0.77
(LD + LE) Ascending	0	1.00	0	1.57
(LD + LE) Descending	0	0.80	0	1.03
Random	0	0.47	0	1.17

6. Conclusion

We address the UCTTP at the FCSIT, UNIMAS utilising a 2-stage approach. In stage 1, HO is used to find a feasible solution. In stage 2, a hybrid of HO and VND is used to improve the quality of the solution. The proposed algorithm is tested on real-world data instances for semester 1 and 2 of 2019/2020.

LD Descending and (LD+LE) Descending ordering manage to generate feasible solutions for both the instances when the number of allocated timeslots is increased to 35, which is less compared to the number of allocated timeslots (48) required by the existing timetabling software.

We also compare different HO and NS in VND. VND works best with LE Ascending, LD Descending, LE Descending, (LD+LE) Descending and random ordering for both the instances by executing single iteration. The proposed algorithm manages to achieve soft constraint (split a course with 4 continuous lecture hours over minimum working days) violations of 0 and 1 for instances for semesters 1 and 2, respectively. However, all HO manage to yield 0 violation for both instances after 30 iterations of the proposed algorithm. Results show that the order of NS also will impact the number of soft constraint (SC1) violations achieved in this study. Future research may focus on other soft constraints such as one-hour lunch break and minimising isolated events, which are also the concern of most universities.

Acknowledgements

The work was supported by Hadiah Latihan Persekutuan (HLP) scholarship from Ministry of Higher Education Malaysia.

References

Akkan, C., & Gülcü, A. (2018). A bi-criteria hybrid Genetic Algorithm with robustness objective for the course timetabling problem. Computers and Operations Research, 90, 22–32. https://doi.org/10.1016/j.cor.2017.09.007

Assi, M., Halawi, B., & Haraty, R. A. (2018). Genetic Algorithm Analysis using the Graph Coloring Method for Solving the University Timetable Problem. Procedia Computer Science, 126, 899–906. https://doi.org/10.1016/j.procS.2018.08.024

Babaei, H., Karimpour, J., & Hadidi, A. (2015). A survey of approaches for university course timetabling problem. Computers and Industrial Engineering, 86, 43–59. https://doi.org/10.1016/j.cie.2014.11.010

Borchani, R., Elloumi, A., & Masmoudi, M. (2017). Variable neighborhood descent search based algorithms for course timetabling problem: Application to a Tunisian University. Electronic Notes in Discrete Mathematics, 58, 119–126. https://doi.org/ 10.1016 /j.endm.2017.03.016

Burke, E., Curtois, T., Post, G., Qu, R., Veltman, B., Burke, C. E., Curtois, T., Post, G., Qu, R., & Veltman, B. (2005). University of Nottingham Jubilee Campus Computer Science Technical Report No . NOTTCS-TR-2005-9 A Hybrid Heuristic Ordering and Variable Neighbourhood Search for the Nurse Rostering Problem A Hybrid Heuristic Ordering and Variable Neighbourhood Search for.

Burke, E. K., & Petrovic, S. (2002). Recent research directions in automated timetabling. European Journal of Operational Research, 140(2), 266–280. https://doi.org/10.1016/S0377-2217(02)00069-3

Erdeniz, S. P., & Felfernig, A. (2018). OCSH : Optimized Cluster Specific Heuristics for The University Course Timetabling Problem. 0–5.

Goh, S. L., Graham, G., Sabar, N. R., & Abdullah, S. (2020). An effective hybrid local search approach for the post enrolment course timetabling problem. OPSEARCH. https://doi.org/10.1007/s12597-020-00444-x

Goh, S. L., Kendall, G., & Sabar, N. R. (2017). Improved local search approaches to solve the post enrolment course timetabling problem. European Journal of Operational Research, 261(1), 17–29. https://doi.org/10.1016/j.ejor.2017.01.040

Goh, S. L., Kendall, G., & Sabar, N. R. (2019). Simulated annealing with improved reheating and learning for the post enrolment course timetabling problem. Journal of the Operational Research Society, 70(6), 873–888. https://doi.org/10.1080/01605682.2018.1468862

Habashi, S. S., Yousef, A. H., Salama, C., & Fahmy, H. M. A. (2018). Adaptive Diversifying Hyper-Heuristic Based Approach for Timetabling Problems. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 259–266.

Hansen, P., & Mladenovi´c, N. (2014). Variable neighborhood search. In Edmund K. Burke & G. Kendall (Eds.), Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques (pp. 313–338). Springer New York Heidelberg Dordrecht London. https://doi.org/10.1007/978-1-4614-6940-7_12

Hansen, P., Mladenovi´c, N., Brimberg, J., & Pérez, J. A. M. (2018). Variable Neighborhood Search. In International Series in Operations Research & Management Science (Vol. 272). Springer New York LLC. https://doi.org/10.1007/978-3-319 -91086-4_3

Matias, J. B., Fajardo, A. C., & Medina, R. P. (2019). A hybrid genetic algorithm for course scheduling and teaching workload management. 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2018, 1–6. https://doi.org/10.1109/HNICEM.2018.8666332

Muklason, A., Irianti, R. G., & Marom, A. (2019). Automated course timetabling optimization using tabu-variable neighborhood search based hyper-heuristic algorithm. Procedia Computer Science, 161, 656–664. https://doi.org/10.1016/j.procs. 2019.11.169

Pillay, N., & Özcan, E. (2019). Automated generation of constructive ordering heuristics for educational timetabling. Annals of Operations Research, 275(1), 181–208. https://doi.org/10.1007/s10479-017-2625-x

Schaerf, A. (1999). Survey of automated timetabling. Artificial Intelligence Review, 13(2), 87–127. https://doi.org/10.1023/A:1006576209967

Tan, J. S., Leng, S., Kendall, G., & Sabar, N. R. (2021). A survey of the state-of-the-art of optimisation methodologies in school timetabling problems. Expert Systems With Applications, 165(May 2020), 113943. https://doi.org/10.1016/j.eswa.2020.113943

Tan, J. S., Say, T., Goh, L., Sura, S., Kendall, G., & Sabar, N. R. (2020). Hybrid particle swarm optimization with particle elimination for the high school timetabling problem. Evolutionary Intelligence, 0123456789. https://doi.org/10.1007/s12065-020-00473-x

Thepphakorn, T., & Pongcharoen, P. (2020). Performance Improvement Strategies on Cuckoo Search Algorithms for Solving the University Course Timetabling Problem. Expert Systems With Applications, 113732. https://doi.org/10.1016/j.eswa.2020.113732

Vianna, D. S., Martins, C. B., Lima, T. J., Vianna, M. de F. D., & Meza, E. B. M. (2020). Hybrid VNS-TS heuristics for University Course Timetabling Problem. Brazilian Journal of Operations & Production Management, 17(1), 1–20. https://doi.org/10.14488/bjopm.2020.014

AUTHOR BIOGRAPHIES

Mei Ching Chen is a PhD student in the Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak. She obtained her Master of Education in Technical and Vocational Study and Bachelor of IT in Computational Science. Her current research interest is hybrid algorithms with specific interest in timetabling problems.

Dr. San Nah Sze is presently working as a senior lecturer at the Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak (UNIMAS). She received her Doctor of Philosophy (PhD) in University of Sydney in 2011. She has more than 10 years of experience in research and development (R&D). Her research

interests include educational timetabling, vehicle routing, heuristics and meta-heuristics. She had presented at numerous conferences and published her work in ISI and Scopus publications.

Dr. Say Leng Goh is an academic with the Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia. He graduated with a first class B.I.T. degree from Universiti Malaysia Sabah. He received his M.Sc. degree and Ph.D. degree from Imperial College London and University of Nottingham, respectively. He has published in various top-tier journals such as European Journal of Operational Research, The Operational Research Society, Expert Systems with Applications etc. His current research interests include artificial intelligence, operational research, discrete optimisation, meta-heuristics, timetabling and scheduling.

Dr. Lau Sei Ping is a senior lecturer at Faculty of Computer Science & Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS). He received his Doctor of Philosophy (PhD) from University of Southampton, UK at 2015. His in-depth knowledge in computer science had given him the trust to handle wide varieties of course in UNIMAS including commercial postgraduate program courses. Besides the teaching and learning activities, he also actively participates and had more than 15 years experience in research and development (R&D) activities. The area of interests includes wireless sensor networks, applied computing, and cybersecurity.

Medium-Term Wind Speed Prediction using Bayesian Neural Network (BNN)

Sana Mohsin[1], Sofia Najwa Ramli[2*] and Maria Imdad[2]
[1] Asia Pacific University of Innovation and Technology, Kuala Lumpur, Malaysia
[2] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Parit Raja, Johor, Malaysia
* Corresponding author E-mail: sofianajwa@uthm.edu.my

(Received 11 July 2021; Final version received 22 July 2021; Accepted 7 September 2021.)

Abstract

Renewable energy has become an emerging source of daily energy utilization in recent years. Non-conventional sources are extensively applied in the smart grid due to their environment friendly and relatively easy maintenance. Wind energy unlike other conventional sources has drawn attention in terms of clean energy production. Due to unpredictable nature of wind, it is difficult to trade energy to the smart grid without any power loss. Variations in wind energy affect power scheduling, wind power generation, and energy management. Therefore, wind speed forecasting is an important tool to address such problems. Machine learning approaches have always been considered for accurate wind speed prediction. To evaluate the performance of machine learning algorithms, several models have been tested to achieve precise prediction. Amongst these several models, Neural Networks perform best and optimizes the prediction at its maximum. Apropos, in this paper, Bayesian Neural Network (BNN) is used for predicting medium-term wind speed on different time horizons. The input for training purposes is taken from Numerical Weather Prediction (NWP) model and sifted as per the model's requirement. After successive training, it is evident from the percentage Mean Absolute Percentage Error (MAPE) and Normalized Mean Absolute Error (NMAE) criterion that BNN has achieved good accuracy as compared to Least Absolute Shrinkage and Selection Operator (LASSO). Ultimately, the proposed model has shown that it can bring precision and accuracy for prediction and can be applied for other renewable sources as solar and water as well.

Keywords: Bayesian Neural Network (BNN), Least Absolute Shrinkage and Selection Operator (LASSO). Medium-term wind speed prediction, Numerical Weather Prediction (NWP).

1. Introduction

Among all other renewable energy sources, the share of wind energy is increasing day by day globally. Wind energy supplies 4.7% of electric power Worldwide (Dyatlov, Didenko, Ivanova, Soshneva, & Kulik, 2020). The rise in wind energy usage has drawn attention towards forecasting because large-scale integration of wind energy needs accuracy and precision in forecasting. Accurate and effective power generation forecast systems are needed to combat wind energy sources intermittency and variable nature. Days-ahead forecasting is planned and organized from the perspective of the wind power plant owner. The demand for days-ahead forecasting is highly used, especially for the power grid's energy trade, transmission, distribution, optimization, and security (Nema, Nema, & Rangnekar, 2009; Ogundiran, 2018).

Numerous amounts of work have been established in the domain of wind speed forecasting. Primarily, it has been observed that forecasting is done on the bases of time horizon. The time scale for forecasting purpose, influences, decision-making, as the ranges are discussed briefly (Prakesh, Sherine, & BIST, 2017).

Short-term (from few seconds to hours): It is purposely used for storage control and the electricity market. In the smart grid, short-term forecasting plays an important role.

Medium-term (from 6 to 72 hours ahead): this range is a bit crucial to handle deciding management and planning. It deals with economic dispatch and operational management of the grid.

Long-term (up to one week ahead): It is helpful for maintenance, scheduling, and distribution of utilities, etc.

Moreover, the forecasting also includes the nature and structure of wind power plant, terrain and data (Buhan & Çadırcı, 2015) (Buhan, Özkazanç, & Çadırcı, 2016). Wind speed generation forecasting is categorized into three models, which are physical, statistical, and hybrid models. As the names suggest, physical models which process the physical data are usually obtained from NWP or wind power plant' s landscape. The statistical model depends upon the historical meteorological data, while the hybrid model is the combination of physical and statistical models that have captivated more researchers (Zhou, Wang, & Zhang, 2019).

Depending on the time horizon, a suitable forecasting model is selected (Soman, Zareipour, Malik, & Mandal, 2010). There are typically two kinds of statistical forecasting models that are linear and non-linear. In the previous research carried out in this field, it has been seen that both models are widely adopted for prediction purposes. Linear models generally deal with statistical and historical data. It can be safely compared with the persistence model (Soman et al., 2010). Typically, linear models that have been used extensively are Linear Regression (LR), Autoregressive with exogenous variable (AR), Autoregressive with integrated moving average (ARIMA), and Kalman filtering. While on the other hand, non-linear models are generally operated with NWP models, which predict weather parameters and range from Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Fuzzy logics are the best machine learning algorithms (Y. Liu, Zhang, Chen, & Wang, 2018) (Lydia, Kumar, Selvakumar, & Kumar, 2016).

2. Literature Review

Forecasting is an essential feature in the operations of the grid. It regularizes and manages the performance of an energy management system (EMS). The main challenge behind forecasting the wind speed or is its volatile and variable nature. The unpredictable and intermittent nature of wind speed creates hurdles in the planning and controlling of a grid system (H. Liu, Duan, Wu, Li, & Dong, 2019). Therefore, the system requires an efficient and accurate predictive tool to solve this issue at its best. (Ahadi & Liang, 2018) proposed a neural network model to predict wind speed time series. The proposed neural network is trained with different training models, such as Bayesian Regularization, Levenberg-Marquardt, and Scaled-conjugate gradient. The results were compared with ARMA and showed that the neural network approach demonstrates more accurate output. It is observed very keenly that wind speed forecasting serves to schedule and dynamic control of power management systems. (Ye, Ding, & Wan, 2021) distinguishes few facts about wind speed's randomness, irregularity, and non-linear nature and brought significance in using the Bayesian model. The study added more than the Bayesian model with Gaussian process prior is adopted for high flexibility, probabilistic evaluation, and predictive variance. It was concluded from the research that Bayesian modeling is a good choice as a predictive tool for predicting wind speed.

On the other hand, deep learning seems to be in the limelight in recent studies. Especially when forecasting is involved, deep learning using Long term short memory

(LSTM) has got researcher’s attention. (Bali, Kumar, & Gangwar, 2019) stated that the implementation of extensive data set for prediction purposes is a big challenge to be catered to. But, the usage of LSTM can solve this problem by having a deep analysis of the data set. LSTM is known for its accuracy and pattern remembrance for a more extended period. The LSTM model concludes that wind speed can be best predicted with deep learning models and can bring efficiency to the system. Furthermore, lots of disputes have been seen when dealing with the data set. Pre-processing and sifting of the data set have been taken as an arduous task for further training.

The thumb rule of using any machine learning model is to make the data set able to train. (H. Liu, Mi, & Li, 2018) used different predictive tools for different purposes. This study reveals that the decomposition and organization of the data itself is an important thing to consider. Empirical wavelet transform is used to mortify the data set then LSTM is employed to the data arrays with low frequency. At the same time, Elman Neural Network is used for higher frequency data arrays. It is worth noticing that using a combination of machine learning models for a different purpose can achieve high accuracy in predicting wind speed.

The literature review shows that non-linear machine learning approaches are more effective in terms of accurate prediction. In (Y. Wang, Shen, Mao, Chen, & Zou, 2018) Wang has combined least absolute shrinkage and selection operator (LASSO) with long short-term memory (LSTM) for short term prediction of solar intensity where combined model better effectiveness and accuracy. Similarly, LASSO has been used for solar power generation forecasting in (Tang, Mao, Wang, & Nelms, 2018), whereas no evidence has been found where LASSO has been used for medium term wind speed prediction. In (Blanchard & Samanta, 2020), different ANN models are used to predict wind speed. The results show that non-linear autoregressive (NAR) and non-linear autoregressive with exogenous input (NARX) have achieved better accuracy than the persistence model. Another research conducted in 2019 revealed that any neural network and its respective kind could enhance wind speed forecasting. Besides, feed-forward neural networks are primarily used in prediction, forecasting, or classification (J. Wang, Zhang, & Lu, 2019). In (Ashraf, Raza, & Saleem, 2020) and (Kaur, Kumar, & Segal, 2016), the performance and optimization of different networks are shown. It has been perceived that neural networks with different parameters can attain high accuracy and precision compared to linear models.

In this research work, medium-term wind speed forecasting is originated by employing a feed-forward neural network with a Bayesian regularization training model and LASSO for making a comparison in terms of checking accuracy. Bayesian regularization training model will be called as Bayesian Neural Network (BNN). The study's fundamental approach is to predict 6 to 72 hours ahead wind speed of the wind power plant situated in Sindh, Pakistan. This research’s main objective is to observe the performance of Bayesian regularization with the mentioned time horizon scale. In short-term wind speed forecasting, the persistence model is usually used to compare the precision, but in the medium term, the forecasting persistence model is not useful (Ahmed & Khalid, 2018). In a study by (Kumar & Sahay, 2018) different neural networks with different training models are adopted and are used for wind speed forecasting. This research primarily shows that when BNN is compared with a linear model LASSO, where it attains good accuracy for predicting wind speed. In addition, BNN ignores uncertainties at its best and updates the weights to extend the standard network with maximum likelihood. In this study, mainly BNN is employed for medium term wind-speed forecasting and for comparative analysis LASSO has been used for observing the efficacy.

Bayesian Neural Network (BNN)

Bayesian Neural Networks contain a unique function of regularization with probabilistic approach. This is one of the classic training models used in regression and classification. It is also noted that BNN’s are very much flexible in designing the architecture of the network. Deciding the inputs, hidden layers, and adjustment of distributive weights are easy and effortless in BNN (Niu, Fang, & Niu, 2019). Fig. 1 indicates every connection of input relates to neurons of a hidden layer with a distribution of adjacent weights.

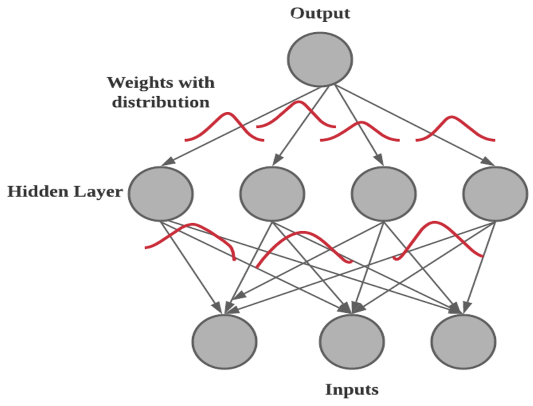


Fig. 1. The Architecture of BNN

The distributed weights are approaching the Probability Density Function (PDF). Using the Bayes theorem, these pdf generated weights are called prior to distributions, which are later converted into the posterior distribution (Maiti, Kumar, Sarkar, Tiwari, & Srinu, 2019). The mathematical expression of the Bayes theorem is shown below in Eq. (1), with X and Y are the events and the rest of them are the probabilities of events chosen.

$$P(Y\,|\,X)=\frac{P(X\,|\,Y)\bullet P(Y)}{P(X)} \qquad (1)$$

This probability is often called evidence of the model. In neural networks, training occurs with a different number of neurons and is assembled with their respective evidence. Eq. (2) and Eq. (3) show the architecture of neural networks concerning their evidence or PDF, whereby y is the desired output, x is the input, w is the distributed weight, and φ is an activation function.

$$y=w_1(P(y\,|\,x))x_1+w_2(P(y\,|\,x))x_2+ \\ w_3(P(y\,|\,x))x_3+.....+w_n(P(y\,|\,x))x_n \qquad (2)$$

$$v=\phi(y) \qquad (3)$$

$$\phi=\frac{1}{1-\exp(-az))} \qquad (4)$$

As illustrated in the equations and the architecture, Bayes theorem has attached a probability to the training model and estimated the forecast. Furthermore, it has also been seen that it can be used as a selection parameter.

3. Model Implementation

The data obtained from the NWP model is pre-processed and assembled according to the proposed predictive model. The data comprises of wind speed in 3-hour resolution up to 72 hours. BNN has always been considered ideal for short term wind speed prediction ranging from 0-6 hours. Here, this has been extended for medium term prediction with a continuous range of 6-72 hours (~3 days ahead). Each hour is trained individually for checking the efficacy and accuracy. BNN model is always used for optimization in prediction models whereas here it has been implemented in two phases, which are training and testing of NN. Seventy percent of the data is allotted for training, while the rest is for testing and validation. The training phase apparently depends on the architecture of the neural network. The parameters chosen for training are indicated below in Table 1. It clearly shows that only wind speed (m/s) is chosen as a variable and has the desired output of the same dimension at the input. After a hit and trial of choosing hidden layers, 30 hidden layers are decided for further continuation.

Table 1. Selected parameters for training the model.

Parameters	Values
Input layer	343×1
Hidden layer	30
Output layer	343×1
Train indices	292×1
Test indices	51×1
Validation indices	51×1
Epochs	982
Best epoch	79
Gradient	847
Mu	0.0050

The simulated architecture of BNN is shown in Fig. 2, with w is the weight and b is biases contained in the neural network.

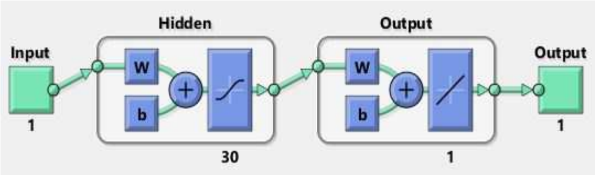


Fig. 2. The simulated architecture of BNN.

Proposed Methodology

As mentioned above, BNN is chosen as a predictive model for forecasting medium-term wind speed along with LASSO as a comparative model. Wind speed is irregular and is not steady always, due to which pre-processing of data is a bit crucial for training and testing purpose. The flowchart of the proposed methodology is demonstrated in Fig. 3. The trend by the NWP data has been depicted in Fig. 4, which can serve as a pattern for showing the raw data for making comparisons with outputs. A step by step demonstration of the procedure from Fig. 3 is given below as:

The physical data is obtained from the NWP model. The chosen variable is wind speed (m/s). The data set is organized in 3-hour resolution from 6 to 72 hours ahead of the 2016 year. The data is sifted according to the predictive model as shown in Table 1.

In every supervised machine learning algorithm, selecting a target or response is a crucial step to follow. The target acts as a catalyst between input and the desired output. Thus, the target is chosen according to the output needed. In this paper, the observed wind speed is taken as the target.

After the successive pre-processing of data, the data set is ready to employ to BNN. Seventy percent of the data samples are utilized for the training and the rest for the testing and validation. The tested sample's accuracy is checked through statistical measures: Mean Absolute Percentage Error (MAPE) and Normalized Mean Absolute Error (NMAE). These quantifying measures are shown in Eq. (5) and Eq. (6), where z is the actual wind speed, while z' is the final predicted wind speed.

$$MAPE = \frac{1}{N}\sum_{i=0}^{n}\left|\frac{z-z'}{z}\right| \times 100 \qquad (5)$$

$$NMAE = \frac{\sum_{i=0}^{n}|z-z'|}{N} \qquad (6)$$

If the MAPE and NMAE criterion is satisfied in the wake of accuracy, then the prediction is successful. The estimation of the MAPE and NMAE depends upon some threshold values which define accuracy level. If the MAPE is achieved below 10 percent, it is considered a higher accuracy or good accuracy. While in the context of NMAE, if the NMAE is achieved less than 1 %, the accuracy goals are achieved (Nespoli et al., 2019) (L. Wang, Lv, & Zeng, 2018).

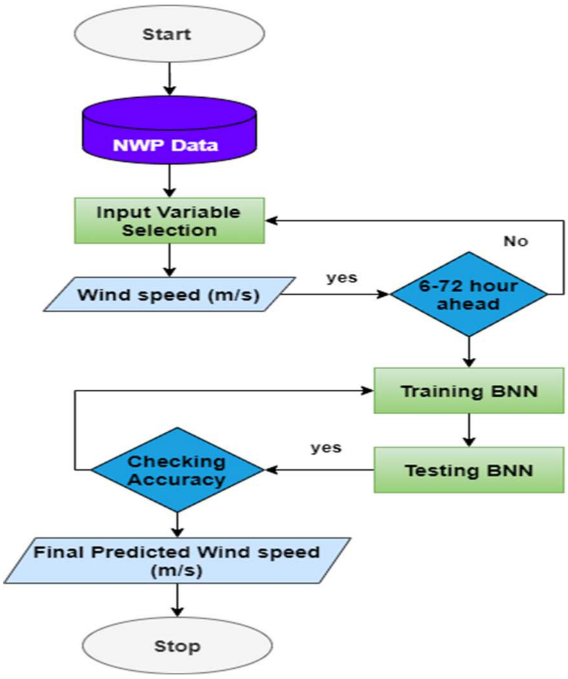


Fig. 3. Flowchart of the proposed methodology.

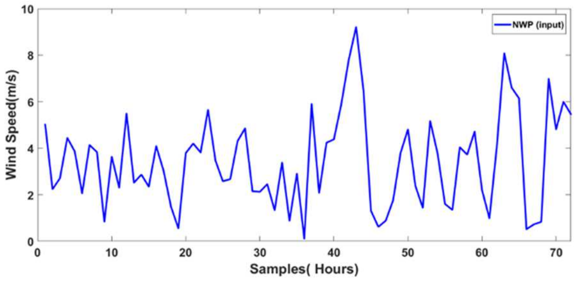


Fig. 4. Trend by NWP data (Original data).

Simulations and discussions

The final predicted wind speed is measured for every 3-hour resolution from 6 to 72 hours ahead for both BNN and LASSO. As the nature of wind speed is always discontinuous, it is hard to decide when the NWP model is forecasting the wind speed. Therefore, for addressing this problem, prediction is made possible to remove the errors between actual and predicted wind speed.

As shown in Table 2, the MAPE and NMAE are used as performance evaluators showing some magnificent accuracy. The BNN model has attained and fulfilled the NMAE criterion of having less than 1% estimation, which is evident from the results. While on the other hand, the estimation from MAPE is dwindling between good to reasonable accuracy. Few of the hours are showing reasonable accuracy above 20 percent. Whereas LASSO has been selected as a comparative model and its quantifying measures are listed in Table 3, where it is evident that the MAPE values are very poor as compare to BNN. Conclusively, it has been observed that BNN has achieved good accuracy in terms of quantifying measures. The forecasting estimation performance for MAPE and NMAE for BNN is shown in Fig. 5 and Fig. 6, respectively.

Table 2. Quantifying measures for BNN.

Hours	MAPE	NMAE	Hours	MAPE	NMAE
9 Hour ahead	19.01	0.003	42 Hour ahead	11.90	0.03
12 Hour ahead	19.39	0.001	45 Hour ahead	17.37	0.04
15 Hour ahead	16.4	-0.021	48 Hour ahead	25.29	0.06
18 Hour ahead	12.71	0.029	51 Hour ahead	30.86	0.083
21 Hour ahead	15.98	0.04	54 Hour ahead	20.79	0.03
24 Hour ahead	33.33	0.03	57 Hour ahead	24.13	0.068
27 Hour ahead	16.81	-0.002	60 Hour ahead	22.58	0.05
30 Hour ahead	18.87	0.01	63 Hour ahead	20.32	-0.02
33 Hour ahead	18.20	-0.03	66 Hour ahead	15.31	0.02
36 Hour ahead	20.84	-0.02	69 Hour ahead	18.16	0.04
39 Hour ahead	18.79	0.002	72 Hour ahead	26.9	0.08

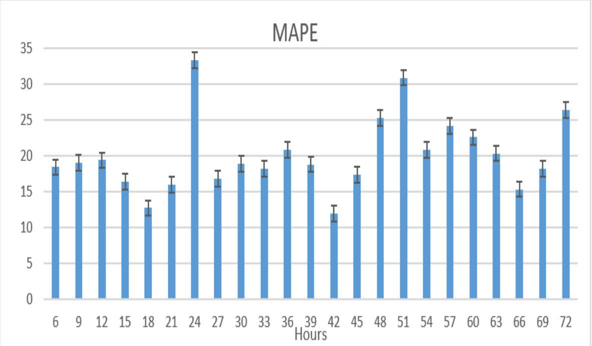


Fig. 5. MAPE criterion for BNN (6 to 72 hours ahead).

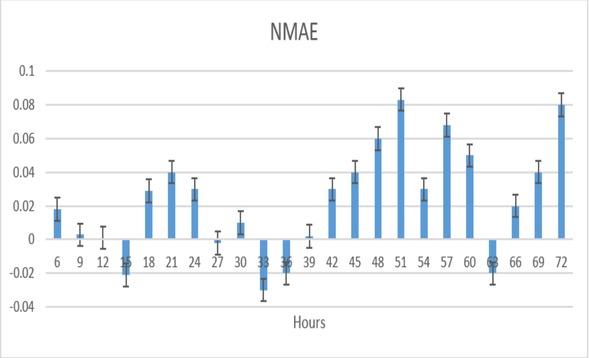


Fig. 6. MAPE criterion for BNN (6 to 72 hours ahead).

Table 3. Quantifying measures for LASSO

Hours	MAPE	Hours	MAPE
9 Hour ahead	249.3	42 Hour ahead	35.84
12 Hour ahead	41.13	45 Hour ahead	36.49
15 Hour ahead	39.99	48 Hour ahead	41.84
18 Hour ahead	36.9	51 Hour ahead	50.89
21 Hour ahead	36.32	54 Hour ahead	42.3
24 Hour ahead	56.8	57 Hour ahead	39668.1
27 Hour ahead	-16670	60 Hour ahead	40.75
30 Hour ahead	40.32	63 Hour ahead	37.45
33 Hour ahead	39.01	66 Hour ahead	35.26
36 Hour ahead	43.41	69 Hour ahead	36.27
39 Hour ahead	38.09	72 Hour ahead	42.6

For checking the precision in machine learning algorithms, the trend between output and the chosen target should be absolute. Wind speed characteristics of several hours such as 9, 36, 48, 57, and 72 are shown in Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11, for observing the trend between target and output. The above mentioned predicted hours showed unique trends, but the other predicted hours were following the same trend which are only exhibited in Table 2. Fig. 7(i) and 7(ii) shows 9 hours ahead prediction for LASSO and BNN respectively. For LASSO, it can be seen that there are no intersecting peaks to observe. The troughs and crests of trends are not showing regularity between output and target which means the there is no accuracy in prediction by LASSO. However, it can be observed that there is a minimal difference between target and output for BNN, which highlights good accuracy by the proposed predictive tool.

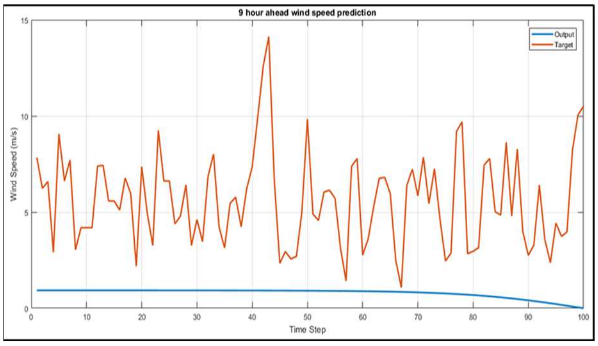


Fig. 7(i). 9 Hour ahead wind speed prediction (LASSO).

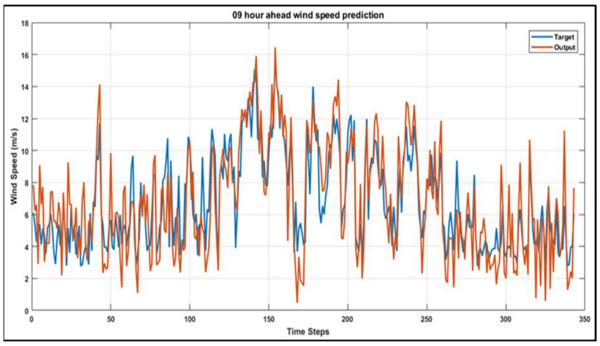


Fig. 7(ii). 9 Hour ahead wind speed prediction (BNN)

After completion of two days, the trend is getting closer and showing precision and efficiency for BNN. These trends have been presented in Fig. 8, Fig. 9, Fig. 10, Fig. 11(i), and Fig. 11(ii) for 36, 48, 57, and 72 hours ahead, respectively. In all the cases, LASSO did not perform well where it was found having no intersecting point between target and output nor was there any convergence observed in Fig. 11(i). Whereas a clear pattern of convergence can be observed in these figures between target and output, which indicates that the BNN model has outperformed well and removed the possible outliers between the target and output.

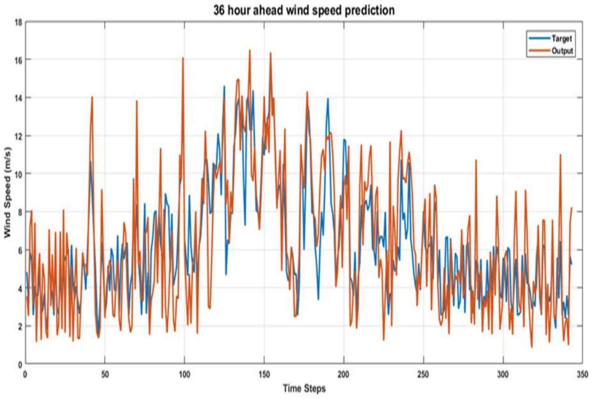


Fig. 8. 36 Hour ahead wind speed prediction (BNN).

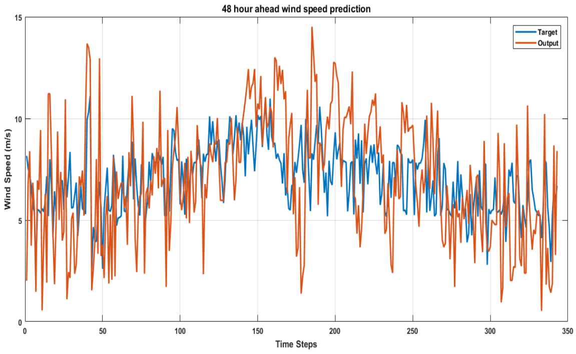


Fig. 9. 48 Hour ahead wind speed prediction (BNN).

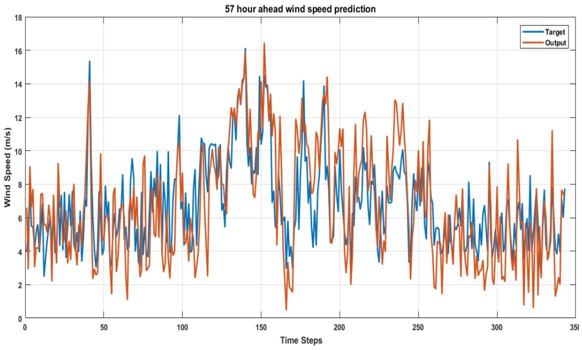


Fig. 10. 57 Hour ahead wind speed prediction (BNN).

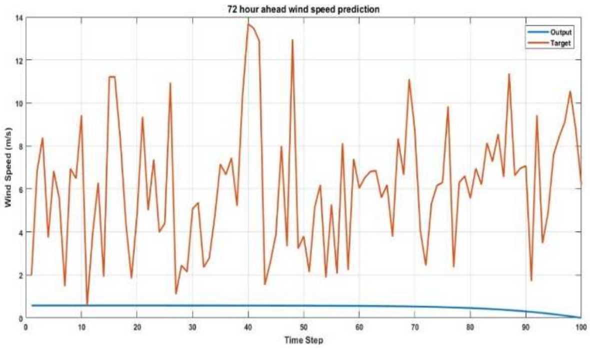


Fig. 11(i). 72 Hour ahead wind speed characteristics (LASSO).

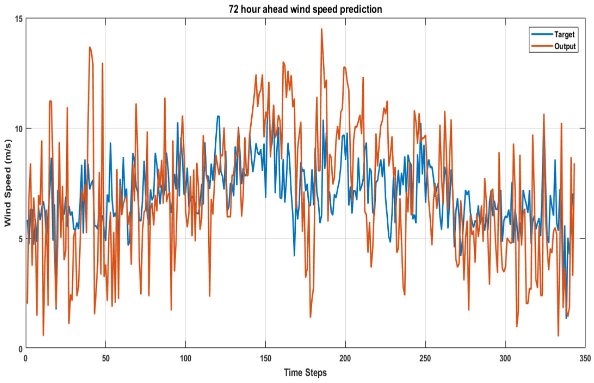


Fig. 11(ii). 72 Hour ahead wind speed characteristics (BNN).

These results mentioned above indicate that, LASSO is not an ideal approach for wind and other power predictions. While on the other hand BNN is an ideal candidate for medium-term wind speed forecasting. As the prediction duration increases, the values are still accurate, and no extreme irregularity has been observed.

4. Conclusions

The surge in need for wind speed forecasting has increased manifolds from the past two decades. Due to variation in weather changes, significant problems came across in the supply and demand of electricity. Therefore, predictive models are utilized and employed to combat this problem. In this paper, medium-term wind speed is predicted using BNN and LASSO. The data is obtained by the NWP model, which usually forecasts the weather parameters. The key objective of predicting the wind speed is to eliminate the errors between predicted and actual wind speed using BNN. The performance indices used for measuring the accuracy are MAPE and NMAE. The result shows that BNN has outperformed exceptionally well and attained good accuracy for 6 to 72 hours ahead of prediction by the estimation forecasting as compare to LASSO which didn’ t provide good accuracy at any hour. Extensive research has been carried out in wind speed forecasting using different machine learning approaches, and still, there is always room for improvement. Additionally, the proposed predictive model seems to be useful for wind speed forecasting for longer time horizons.

In prospects, machine learning techniques have a wide range of algorithms for improving prediction performance. Firstly, BNN can do long-term wind speed forecasting on a broader time horizon. Secondly, the hybridization of different linear and non-linear can also be utilized for wind speed prediction (Babbar & Lau, 2020). In last decade there has been research on combining techniques for prediction purposes, and BNN is observed promising for the ensemble approach. But the small gap has been observed in attaining high accuracy due to the limitation of data set. Lastly, the proposed BNN model can also be applied to other energy forecasting domains, such as solar power forecasting, planning and optimization of the electric grid, and load forecasting.

Acknowledgments

The authors would like to thank all reviewers for their helpful comments. The authors would also like to thank the Ministry of Higher Education Malaysia for supporting the research under Fundamental Research Grant Scheme Vot No. FRGS/1/2019/ICT03/UTHM/03/1 and partially sponsored by Universiti Tun Hussein Onn Malaysia.

References

Ahadi, A., & Liang, X. (2018). Wind Speed Time Series Predicted by Neural Network. Paper presented at the 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE).

Ahmed, A., & Khalid, M. (2018). An intelligent framework for short-term multi-step wind speed forecasting based on Functional Networks. Applied Energy, 225, 902-911.

Ashraf, M. B., Raza, S., & Saleem, U. Z. (2020). A Comparative Analysis on Wind Speed Forecast using Optimized Neural Networks. Pakistan Journal of Engineering and Technology, 3(2), 23-28.

Babbar, S. M., & Lau, C.-Y. (2020). Medium term wind speed forecasting using combination of linear and nonlinear models. Solid State Technology, 63(1s), 874-882.

Bali, V., Kumar, A., & Gangwar, S. (2019). Deep learning based wind speed forecasting-A review. Paper presented at the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

Blanchard, T., & Samanta, B. (2020). Wind speed forecasting using neural networks. Wind Engineering, 44(1), 33-48.

Buhan, S., & Çadırcı, I. (2015). Multistage wind-electric power forecast by using a combination of advanced statistical methods. IEEE Transactions on Industrial Informatics, 11(5), 1231-1242.

Buhan, S., Özkazanç, Y., & Çadırcı, I. (2016). Wind pattern recognition and reference wind mast data correlations with NWP for improved wind-electric power forecasts. IEEE Transactions on Industrial Informatics, 12(3), 991-1004.

Dyatlov, S., Didenko, N., Ivanova, E., Soshneva, E., & Kulik, S. (2020). Prospects for alternative energy sources in global energy sector. Paper presented at the IOP Conference Series: Earth and Environmental Science.

Kaur, T., Kumar, S., & Segal, R. (2016). Application of artificial neural network for short term wind speed forecasting. Paper presented at the 2016 Biennial international conference on power and energy systems: towards sustainable energy (PESTSE).

Kumar, S., & Sahay, K. B. (2018). Wind Speed Forecasting using Different Neural Network Algorithms. Paper presented at the 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech).

Liu, H., Duan, Z., Wu, H., Li, Y., & Dong, S. (2019). Wind speed forecasting models based on data decomposition, feature selection and group method of data handling network. Measurement, 148, 106971.

Liu, H., Mi, X.-w., & Li, Y.-f. (2018). Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. Energy conversion and management, 156, 498-514.

Liu, Y., Zhang, S., Chen, X., & Wang, J. (2018). Artificial combined model based on hybrid nonlinear neural network models and statistics linear models—research and application for wind speed forecasting. Sustainability, 10(12), 4601.

Lydia, M., Kumar, S. S., Selvakumar, A. I., & Kumar, G. E. P. (2016). Linear and non-linear autoregressive models for short-term wind speed forecasting. Energy conversion and management, 112, 115-124.

Maiti, S., Kumar, C. R., Sarkar, P., Tiwari, R., & Srinu, U. (2019). Interface depth modelling of gravity data and altitude variations: a Bayesian neural network approach. Neural Computing and Applications, 1-20.

Nema, P., Nema, R., & Rangnekar, S. (2009). A current and future state of art development of hybrid energy system using wind and PV-solar: A review. Renewable and Sustainable Energy Reviews, 13(8), 2096-2103.

Nespoli, A., Ogliari, E., Leva, S., Massi Pavan, A., Mellit, A., Lughi, V., & Dolara, A. (2019). Day-ahead photovoltaic forecasting: A comparison of the most effective techniques. Energies, 12(9), 1621.

Niu, Z., Fang, J., & Niu, Y. (2019). Comparative study of radial basis function and Bayesian neural network approaches in nuclear mass predictions. Physical Review C, 100(5), 054311.

Ogundiran, P. (2018). Renewable energy as alternative source of power and funding of renewable energy in Nigeria. Asian Bulletin of Energy Economics and Technology, 4(1), 1-9.

Prakesh, S., Sherine, S., & BIST, B. (2017). Forecasting methodologies of solar resource and PV power for smart grid energy management. International Journal of Pure and Applied Mathematics, 116(18), 313-318.

Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. Paper presented at the North American Power Symposium 2010.

Tang, N., Mao, S., Wang, Y., & Nelms, R. (2018). Solar power generation forecasting with a LASSO-based approach. IEEE Internet of Things Journal, 5(2), 1090-1099.

Wang, J., Zhang, N., & Lu, H. (2019). A novel system based on neural networks with linear combination framework for wind speed forecasting. Energy conversion and management, 181, 425-442.

Wang, L., Lv, S.-X., & Zeng, Y.-R. (2018). Effective sparse adaboost method with ESN and FOA for industrial electricity consumption forecasting in China. Energy, 155, 1013-1031.

Wang, Y., Shen, Y., Mao, S., Chen, X., & Zou, H. (2018). LASSO and LSTM integrated temporal model for short-term solar intensity forecasting. IEEE Internet of Things Journal, 6(2), 2933-2944.

Ye, X. W., Ding, Y., & Wan, H. P. (2021). Probabilistic forecast of wind speed based on Bayesian emulator using monitoring data. Structural Control and Health Monitoring, 28(1), e2650.

Zhou, Q., Wang, C., & Zhang, G. (2019). Hybrid forecasting system based on an optimal model selection strategy for different wind speed forecasting problems. Applied Energy, 250, 1559-1580.

AUTHOR BIOGRAPHIES

Sana Mohsin Baabar is currently a Ph.D. scholar at Asia Pacific University of Innovation and Technology, Malaysia. She did her MS in Electrical Engineering from Abasyn University Islamabad, Pakistan in 2019. She did her BS in Electronics from COMSATS Islamabad, Pakistan in 2015. Her research interests are machine learning, Neural Networking and Artificial Intelligence.

Sofia Najwa Ramli received her PhD degree in Information Security from Universiti Teknikal Malaysia Melaka, Malaysia, in 2016. She received her Master’ s degree (M. Eng) in Electrical – Electronics & Telecommunications Engineering and Bachelor’ s degree (B. Eng) in Biomedical Engineering from Universiti Teknologi Malaysia in 2011 and 2009. She is currently a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. Her current research interests include authentication systems, biometrics, biomedical signal processing, cryptography, and information security. She has been actively involved as a program committee of an international conference and a technical editorial committee of the OIC-CERT Journal of Cyber Security. She has delivered articles in various international conferences and journals.

Maria Imdad is a PhD candidate at Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). She received her Master’ s degree in Information Security from Air University Islamabad, Pakistan in 2017. She did her Bachelor’ s in Software Engineering from Riphah International University Islamabad, Pakistan in 2013. Her research interests are, WSN, In-formation Security, Cryptography and Software Engineering.

Designing a Real-time Interactive Spatial Augmented Reality Platform

Aye Chan Zay Hta[1],2*, YunLi Lee[1],2 and Wai Chong Chia[1],2
[1] Department of Computing and Information Systems, School of Engineering and Technology,
[2] Research Centre for Human-Machine Collaboration (HUMAC),
Sunway University, Bandar Sunway, Selangor, Malaysia
* Corresponding author E-mail: ayechan.aczh@gmail.com

(Received 11 July 2021; Final version received 22 July 2021; Accepted 7 September 2021.)

Abstract

The Spatial Augmented Reality (SAR) system is an interactive platform that allows any virtual contents to be projected onto any physical structure (object). However, most of the SAR platforms use pre-rendered models in a static setting that does not allow the physical structure to be modified dynamically. This has limited interactions between the users and the system because users are unable to make rapid changes to the physical structure to express their ideas. This limitation can be a huge challenge when it comes to application such as city planning, where rapid real-time prototyping is able to provide a better visualization to the impacts that the changes could bring to the city environment. Therefore, this research project aims to design and develop a tracker-based SAR system to resolve the aforementioned limitation. The main contributions of this research project include, (i) a SAR system that supports real-time physical structure reconstruction and projection mapping, and (ii) a SAR platform constructed using Lego blocks and easily accessible hardware and software. The hardware involves the design of the physical set-up to support a real time reconstruction, and dynamic projection mapping. The software involves real time object detection, tracking, and projection mapping. Real time object detection is carried out using colour tracking, and recording Lego positions, while dynamic projection mapping is done through marker tracking and coordinate mapping. Based on preliminary evaluations conducted in the laboratory, the experimental results shown that the proposed SAR system is able to (i) successfully project virtual content onto physical structure built using Lego blocks in real-time, and (ii) detect changes made to the physical structure.

Keywords: Dynamic Reconstruction, Lego blocks, Real-time marker assignment, Spatial Augmented Reality (SAR), Tracker-based SAR

1. Introduction

Digital industry predicts that augmented reality (AR)/virtual reality (VR) would continuously grow and the revenue would increase more than twenty-five billion in the next five years (Makarov, 2021). The future applications of these AR/VR technologies are taken place in various sectors such as business, marketing, education, navigation, health, and others. AR constitutes the integration of virtual resources together with real world physical elements, in which computer-generated graphical components are displayed in the user's digital devices along with the elements of real environment. Milgram and Kishino (1994) explained the operational definition of AR by stating the term that describes any case in which the real environment is "augmented' virtually by computer graphics. The mix reality environment is in between the spectrum of extremes of real and virtual worlds, where the user can interact with both real and virtual objects which are presented at the same display as shown in Fig. 1.

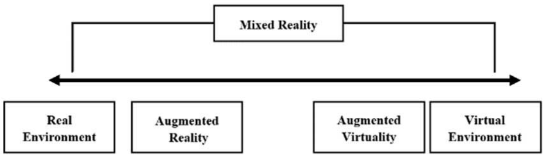


Fig. 1. Virtuality Continuum (Milgram and Kishino, 1994)

Furthermore, with current technologies, limitations in the user's field of view and the ergonomics of wearable AR devices are still challenging. With projection mapping techniques, we can project out these virtual contents onto the real world allowing more seamless blend between the virtual and real world. However, most of the projection mapping systems are static setting without allowing for dynamic reconstruction of physical structure, thus limitations in user interactions (Cortes et al., 2018).

There is a need of interactive tangible user interfaces/platform to allow better data visualization and more innate interactions. With Spatial Augmented Reality (SAR) system, the users can view a larger field in real-time and experience immersive interactions among multiple users. SAR used a projection technology to display the surfaces of a variety of objects with video projection (Ball, 2018; Park et al., 2014). It provides many opportunities for display of events including live shows, museums, exhibitions, conferences, trainings, and designing of products by using audio, video, projectors and software. The audience can appreciate the effect of a combination of audio, video together with 3D modality well beyond the traditional ways (Ball, 2018). SAR allows evaluation of the products by users ahead of physical development of a prototype, thus saving time and cost related to the development of the product (Ball, 2018). Park et al. (2014) mentioned that SAR is useful to show the virtual products which are similar to the real ones without limitation of space issues. From their evaluation of SAR designs, it shows that SAR provides more flexible and intuitive environment with high sense of immersion than using digital display. However, they pointed out that SAR required a more complex set-up of the hardware compared to traditional computer-aided colour design and the immersive experience depends on the projector's performance as the resolution and lumens of the projector are key factors.

1.1 Dynamic Reconstruction SAR Platform

Most SAR platforms were using pre-rendered models and they were used in a static setting without allowing for dynamic manipulation of physical setting. These SAR platforms rely on finger tracking and touch gestures to allow for user interaction. These techniques are viable for 2D projections and touch screen applications. However, in 3D projections the use of movable tangible objects can improve the user's interaction as it allows for a more natural user interaction. A study conducted by Al-Megren and Ruddle (2016) which compares tangible interaction with multi touch interaction showed that the time required to complete tasks were faster as well as less errors occurring in tangible interaction. Those that did allow for dynamic manipulation use markers which limit the virtual projections that are to be mapped and require the use of pre-rendered models, such as in the case of Winder and Larson (2017) which supports 16 types of different markers, all of which are pre-rendered and assigned to the specific marker.

Further literature review was conducted to find SAR platforms that support real-time dynamic reconstruction. Kim et al. (2014) and Guo et al. (2018) support a basic version of dynamic reconstruction by using depth sensors such as the Microsoft Kinect. Based on the depth information, the virtual content responded on to the physical setting. However, they still have limitation in their platforms as they only use the vertical depth information to control the change of virtual content, and therefore their applications cannot detect the actual shape of the object.

There are several issues to be considered such as most SAR platforms do not offer dynamic manipulation of physical setting to the system as it requires object detection and dynamic projection mapping. Additionally, others only used pre-rendered models in a static setting limiting the types of physical structure to pre-determined shapes.

A real-time interactive SAR platform is required to address this problem. In this project, dynamic reconstruction is introduced in SAR system that allows users to manipulate the physical shape of the tangible object that they are interacting with, in real time. Thus, the proposed solution offers an additional layer of user interaction, which overcomes the pre-rendered models and pre-determined shapes in the SAR platform.

1.2 Aims & Objectives

This research aims to develop a general-purpose SAR platform with an additional layer of user interaction through real-time reconstruction, by using a simple, cost effective and easily accessible hardware.

The objectives of the current research project are as follow.

(1) to identify state-of-the-art tracking and calibration techniques that should be used in developing the SAR platform.

(2) to perform real-time projection mapping and 3D reconstruction for objects constructed using Lego blocks.

(3) to track the objects constructed using Lego blocks in real-time using image processing.

(4) to evaluate the usability and generalization of the SAR platform based on feedback /ratings given by end users.

1.3 Project Scope

The proposed real-time interactive SAR platform includes the designing and developing of a tracker-based real-time SAR system and a physical set-up of SAR platform using Lego blocks. Lego blocks are popular and common, therefore supporting the objective of using easily accessible hardware. Furthermore, Lego block can be easily deconstructed and reconstructed by the user allowing for more possibilities in terms of shapes. In this project, only standard 2x2, 2x3 and 2x4 Lego blocks are used. This SAR platform supports the object detection, marker detection and tracking and correct projections mapped from the physical structure to offer a more natural and organic way of user interaction that reduces the barrier between the virtual and the real world.

2. Literature review

A literature review was performed before starting the project in order to identify the existing or similar work on the SAR platforms. This gave a variety of projects with different designs of framework and highlights of issues and challenges that are relevant to SAR platforms. SAR applications require physical set-up design and software architecture design. Most physical designs involve of top, bottom, or even side camera position as well as front projections or rear projections. Common framework design for software architecture were, some form of object detection and tracking which is either marker based (Winder and Larson, 2017; Mousavi et al., 2013; Laviole, 2012) or marker-less (Kim et al., 2014; Guo et al., 2018; Wilson, 2005; Park, 2017), interactive features either direct interaction with the physical content (Winder and Larson, 2017; Kim et al., 2014; Guo et al., 2018; Mousavi et al., 2013; Mousavi et al., 2013; Wilson, 2005) or indirectly (via mobile devices) (Mendes et al., 2019) and projections mapped onto 2D (Mousavi et al., 2013; Laviole, 2012) or 3D objects (Winder and Larson, 2017; Kim et al., 2014; Guo et al., 2018; Laviole, 2012). In addition, the framework of common SAR application designs involves of camera and projector calibration (Fleischmann and Koch, 2016) to support for dynamic projections and to allow accurate projection mapping onto 3D objects. SAR framework design involves of object detection and/or hand/finger detection (Mousavi et al., 2013; Laviole, 2012; Wilson, 2005) to support interactivity. Most SAR platforms are also found to be not portable and some are designing to make it more portable but still these wearable SAR platforms can be bulky, cumbersome and unergonomic to use for extended periods of time.

Furthermore, additional literature review was conducted with a main focus on interactive techniques used as well as the use of dynamic reconstruction in SAR platforms. It was found that most 2D projection-based SAR platforms include virtual buttons and hand gestures as their interactive features. Whereas, most 3D projection-based SAR platforms mainly used tangible objects and some were using real-time reconstruction as their interactive features. Additionally, some 3D projection-based SAR platforms also included an external display which provided more detailed information that supplemented the platform. The common techniques used in existing SAR platforms such as marker-based tracking, depth sensing, free-form tracking, and real-time reconstruction are compared and described in Table 1.

Most SAR platforms prefer to offer direct forms of interaction as it is a main advantage of SAR compared to traditional AR, where users cannot interact with the physical object/world directly and are required to do so through a secondary device such as a touch screen or mobile device. Table 2 shows the comparison of various features of the existing SAR platforms such as whether they support 3D or 2D projections, usage of external display, animation, support of video or audio. In addition, the interactive features like tangible objects, virtual buttons, hand gestures and real-time reconstruction are also compared in Table 2. This table compares and highlights the limitations of each existing work, further confirms the advantages of a system which allows the users to manipulate the physical shape of the tangible object that they are interacting with, in real-time. Furthermore, interaction with tangible objects can provide a more natural form of interaction.

These common techniques and interactive features showed in Table 1 and 2, are useful as a benchmark to consider the required features in the proposed project.

Table 1. Comparison of techniques used in existing SAR platforms.

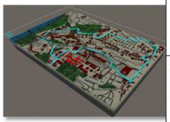
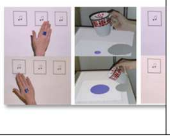
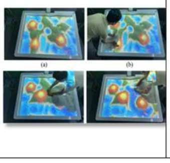
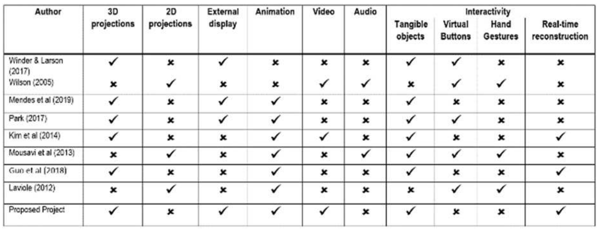
Author	Image	Techniques used			
		Marker based	**Depth sensing**	**Free-form tracking**	**Real-time reconstruction**
Winder & Larson (2017)		✓ (Support for 16 types of different markers)	✗ (No use of depth sensor, uses marker id to determine which content to project)	✗ (No free-form tracking using slot by grid basis)	✗ (Does not support for different shapes that are not predetermined by the marker)
Mendes et al (2019)		✗	✗	✗	✗
		No moving parts, interaction is done through mobile app			
Park (2017)		N/A (Mostly likely static setting with dynamic contents being projected)	✓ (Using an IR camera to gather depth information on pre-defined 3D objects)	✓ (Using OpenCV image processing library to track object)	✗ (No support for building of Lego blocks or changing of new shapes)
Kim et.al (2014)		✗	✓ (Using Kinect to gather 3D depth information	✓	✓ (Allows users to rearrange the square blocks but limited to vertical depth and not actual shape)
Mousavi et al. (2013)		✓ (Uses colour markers on user's fingers for tracking purposes)	✗ (Plans to use Kinect in future works but not yet implemented)	✓ (Supports free-form tracking of hand based on colour markers)	✗
Laviole (2012)		✓ (Uses markers on page sides for tracking purposes)	✓ (Uses Kinect sensor for depth sensing but mainly used for gesture and finger tracking)	✓ (Use of AR markers supports for free-form tracking)	✗
Gou et al. (2018)	 (a) (b)	✗	✓ (Uses Kinect v2 to gather depth information as the sand is build up)	✗ (No tracking needed as projections are changed through depth information)	✓ (Supports real-time reconstruction of the sand box in any shape but limited to vertical depth and not actual shape)

Table 2. Comparison of various interactive features in existing SAR platforms.



Author	3D projections	2D projections	External display	Animation	Video	Audio	Interactivity			
							Tangible objects	Virtual Buttons	Hand Gestures	Real-time reconstruction
Winder & Larson (2017)	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗
Wilson (2005)	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗
Mendes et al (2019)	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
Park (2017)	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗
Kim et al (2014)	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓
Mousavi et al (2013)	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗
Guo et al (2018)	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓
Laviole (2012)	✗	✓	✗	✓	✗	✗	✗	✓	✓	✗
Proposed Project	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓

3. Methodology

In this section, an overview of the proposed architecture is first described. This includes the set-up of the proposed SAR platform by using low-cost and easily accessible hardware. This is followed by the description of the proposed framework, which highlights the techniques used to achieve real-time 3D reconstruction. However, there are several ways (or algorithms) to implement this module. Therefore, evaluation metrics were set-up to evaluate several potential techniques before choosing the best in slot for the proposed framework. The actual implementation of the proposed framework is discussed in more details in Section 4.

3.1 Architecture of Real-Time Interactive SAR Platform

The proposed real-time interactive SAR platform uses markers for tracking after object detection is completed together with depth sensor to detect the shape of an object. It supports free-form tracking through the marker without restricting the user to a grid and slot basis which was employed in Bits and Bricks (Winder and Larson, 2017). Real-time reconstruction is also supported for the detection of actual shape of an object to allow users to build up Lego blocks as they desire. This makes the project innovative in providing the additional feature of real-time reconstruction for actual shape which was not presented in earlier literatures. Those existing works mainly consisted of pre-assigned shapes and even those that supported real-time reconstruction only allowed limited interactions such as in Kim et al. (2014) and Guo et al. (2018) where the virtual content changes only based on the depth axis and thus do not support for more complex shapes.

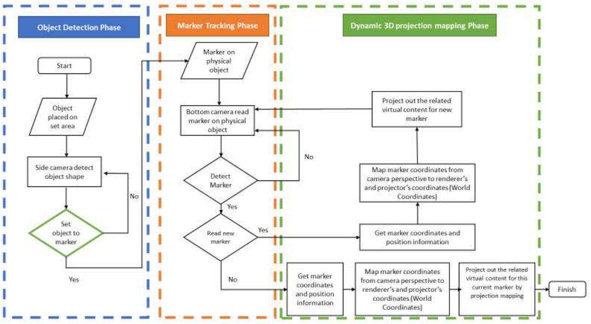

Object Detection Phase
Start
Object placed on set area
Side camera detect object shape
Set object to marker
No
Yes
Marker Tracking Phase
Marker on physical object
Bottom camera read marker on physical object
Detect Marker
No
Yes
Read new marker
Yes
No
Dynamic 3D projection mapping Phase
Project out the related virtual content for new marker
Map marker coordinates from camera perspective to renderer's and projector's coordinates (World Coordinates)
Get marker coordinates and position information
Get marker coordinates and position information
Map marker coordinates from camera perspective to renderer's and projector's coordinates (World Coordinates)
Project out the related virtual content for this current marker by projection mapping
Finish

Fig. 1. Three phases of the proposed real-time interactive SAR platform.

The system design involves three phases as described in Fig. 2.

Phase 1 includes object detection in real-time with the 3D reconstruction that allows the user to create a new object by building Lego blocks and then assign it to a marker in real-time.

Phase 2 consists of detection and tracking the assigned marker in real-time that allows for interaction for user.

Phase 3 consists of dynamic 3D projection mapping which projects out the relevant virtual content and maps it onto the 3D object. This will ensure that the projections will follow the 3D physical object when it is placed in a new position on the platform.

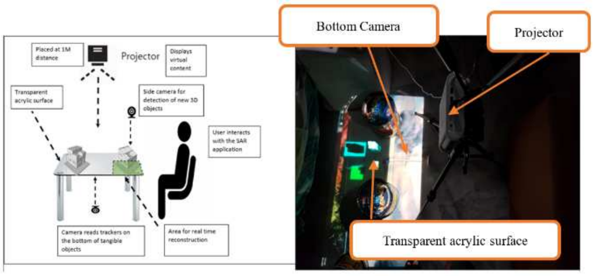

Placed at 1M distance
Projector
Displays virtual content
Transparent acrylic surface
Side camera for detection of new 3D objects
User interacts with the SAR application
Camera reads trackers on the bottom of tangible objects
Area for real time reconstruction
Bottom Camera
Projector
Transparent acrylic surface

Fig. 3. The design of the Physical set-up

To support the SAR platform, physical set-up is designed as shown in Fig. 3. The main hardware resources required are projector (Epson EB-W06), cameras (Acer webcams, Ezviz C1C). These hardware are easier to get and low-cost when compared with more advanced devices. A cheaper projector can also be used at the cost of brightness and resolution of the projection. However, it is recommended to use a projector of 1500 lumens or more to ensure brighter projections. Alternatively, the room can be made darker to make the projections more visible. The projector is used for projection mapping and the camera placed on the bottom of the transparent acrylic surface is used to track markers which are placed on the bottom of the Lego blocks. These markers can be the Lego base's patterns or Fiducial trackers placed onto the bottom of Lego bases. An area is also defined to support for real-time reconstruction where a camera is placed. The users are allowed to change the shape and build the Lego block in this region. By reconstructing the Lego block in this region, it supports coverage by the cameras and reduces the obstructions that may occur. In the next section, the technique adopted by the SAR system to complete the process mentioned in phase 1 to 3 is described.

3.2 Framework for 3D reconstruction

The 3D reconstruction framework involves of the detection of individual Lego block and building it up as described below:

Input: Side camera view image

Output: Reconstruction of Lego 3D model

Algorithm :

1. Detect individual Lego block in the x (defined) region.

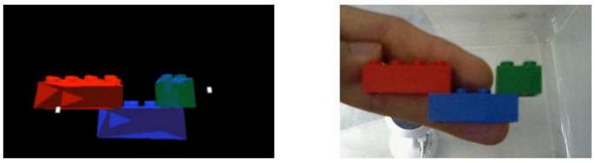
2. Track the Lego block and record the placements when joint with another block.

3. The area for reconstruction is defined as mentioned in the physical set-up to improve the tracking capability and limit camera view obstructions.

4. Assign model to tracker base in real-time.

5. Apply undistort function using intrinsic parameters and map the coordinates to switch to bottom camera.

The individual Lego block is continuously tracked using colour tracking, and their placements are recorded each time they are joined with another block. Trigger areas are assigned on each possible Lego slot (using HitTest VVVV function) to determine where they should join. If the joining Lego block's center touches one of the trigger areas the function with return a Boolean true data type and the index of trigger area will allow it to lock onto that respective Lego slot. This will allow for the support for real-time reconstruction of the 3D model. However, difficulties may arise in the ability to continuously track the Lego blocks. Therefore, to ease the tracking capability the area for reconstruction is defined as mentioned in the physical set-up. Fig. 4 shows the 3D reconstructed virtual model on the left and the real object on the right as seen from the webcam. Once the user has finished building the Lego shape, he/she desires it is then assigned to a marker base so that the tracking can be carried out from the bottom camera. This is done to prevent the projections from interfering with the colour tracker from the top camera.


(a) Digital reconstruction by the system

(b) Image from webcam feed

Fig. 4. Current tests of real-time reconstruction

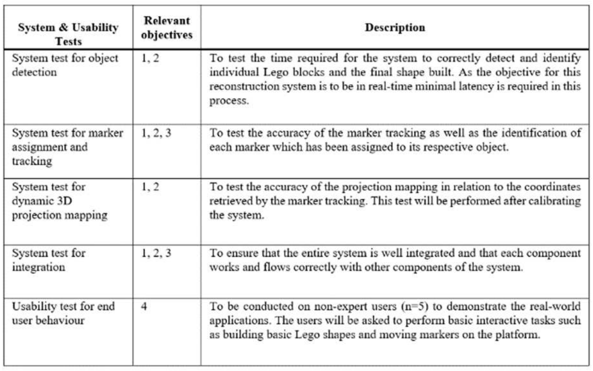
Based on the description given above, several techniques are required to build the proposed framework, and in general they can be divided into three phases (i) trackers, (ii) 3D reconstruction, and (iii) projection mapping. Moreover, there are several ways (or algorithms) to implement the modules mentioned above.

Hence, some preliminary evaluations were carried out to determine which way (or algorithm) works best for each phase. The evaluation metrics as well as the evaluation results are presented in the next section.

3.3 Measurement techniques or evaluation

Different system tests are required to validate the prototype. These include separate tests for object detection, marker assignment and tracking, dynamic 3D projection mapping and integration which are described in the following Table 3. Additionally, user testing is proposed to evaluate the effectiveness of the system from different range of users. It involves the observation techniques to capture the various strategies and approaches that users may take when performing basic interactive tasks with the proposed system.

Table 3. System testing and user evaluation.



System & Usability Tests	Relevant objectives	Description
System test for object detection	1, 2	To test the time required for the system to correctly detect and identify individual Lego blocks and the final shape built. As the objective for this reconstruction system is to be in real-time minimal latency is required in this process.
System test for marker assignment and tracking	1, 2, 3	To test the accuracy of the marker tracking as well as the identification of each marker which has been assigned to its respective object.
System test for dynamic 3D projection mapping	1, 2	To test the accuracy of the projection mapping in relation to the coordinates retrieved by the marker tracking. This test will be performed after calibrating the system.
System test for integration	1, 2, 3	To ensure that the entire system is well integrated and that each component works and flows correctly with other components of the system.
Usability test for end user behaviour	4	To be conducted on non-expert users (n=5) to demonstrate the real-world applications. The users will be asked to perform basic interactive tasks such as building basic Lego shapes and moving markers on the platform.

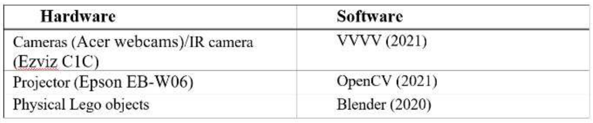
4. Implementation and analysis

The experiments and analysis were carried out to investigate the practicality of the selected techniques. The system is capable of performing colour detection, tracking, real-time real construction of shapes involving 2x2, 2x3, 2x4 bricks. Based on the coordinates retrieved from colour detection and tracking, it is able to perform projection mapping in real-time as well. However, there are still a few issues that need to be resolved which will be further explained in the following section.

4.1 Resources for implementation

The required hardware and software resources were selected in accordance to support, for the objective of using easily accessible hardware and software to develop the proposed SAR platform as described in Table 4. Most hardware resources stated are easily accessible such as cameras, projector and physical Lego blocks. In addition, all of the software used are free and open-source programs.

Table 4. The required resources for development of SAR platform



Hardware	Software
Cameras (Acer webcams)/IR camera (Ezviz C1C)	VVVV (2021)
Projector (Epson EB-W06)	OpenCV (2021)
Physical Lego objects	Blender (2020)

VVVV (platform) (VVVV group, 2021) is used for development as it is a real time interactive live programming environment/toolkit. The real-time aspect of VVVV helps greatly in producing the prototype as changes made in code can be seen in real-time compared to other traditional programming languages where it requires building and compiling. The Open Source Computer Vision (OpenCV) library (2021) is selected for image processing. As the image processing required in this platform involves of object detection and marker tracking. Furthermore, image processing is used in calibration of camera and projector to support for dynamic 3D projection mapping. Lastly, 3D modelling software such as Blender (2020) are chosen to create and edit 3D shapes.

4.2 Phase 1: Implementation of object detection

In this phase, the main aim is to recognise the object constructed by user using a set of Lego blocks. Based on the scope decided, a user can choose from a group of Lego blocks with the size of 2x2, 2x3, and 2x4. To ease the process of recognition, different colours are used for Lego blocks with different sizes. For example, all the 2x2 blocks could be in white colour, 2x3 blocks could be in yellow colour and so forth. With this, colour detection can be adopted to determine the location or where the user has placed the block in the object.

The entire process is continuously monitored. When a user picks a block and adds that to the existing structure, the location is determined and recorded in real-time, so that we can see the 3D reconstruction of the object in real-time on the screen. The trigger area technique is used by setting up trigger areas (using HitTest VVVV function in the shape of a circle) in each of the slots available on a Lego brick. And if the joining Lego block's center touches one of the trigger areas it will lock onto that respective Lego slot as shown in Fig. 5. The translucent green circles represent the trigger areas, and the small white square represents the center point of the Lego object as captured from the webcam. When this center point touches one of the translucent green circles it will turn blue and shift the joining Lego object to its respective slot.

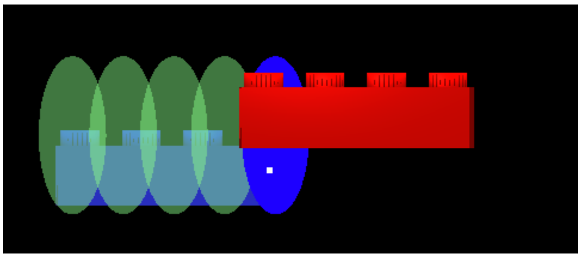


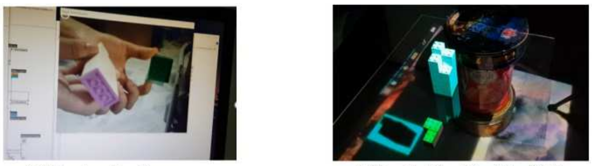
Fig. 5. Implementation of objection detection using trigger areas

4.3 Phase 2: Detection & tracking the assigned marker in real-time

In this phase, the aim is to detect and track a marker continuously, in real-time. Marker-based tracking was implemented by assigning the base pattern and colour of a Lego block to a marker. Both the colour-based tracker and Kernelized Correlation Filter (KCF) tracker were tested in preliminary experiments and results showed satisfactory accuracy in tracking under good lighting conditions as shown in Fig. 6. However, accuracy of tracking suffered under low light conditions, especially in the case of KCF tracker, where it was found to increase the offset in tracking.

Due to these challenges, the system uses col-our-based tracking and place an IR camera under the physical table to better detect the marker features. Additionally, as described in Table 5 in section 4.5 the Aruco marker and fiducial markers were also tested by switching to the bottom camera. This requires remapping of the coordinates as the object shape and size information detected from the top camera must now be tracked by the bottom camera. Undistort functions are also applied to the both the top and bottom cameras to reduce camera distortions using the intrinsic parameters gathered from camera checkerboard calibration.

For the colour tracking function a WithinRange (OpenCV function) is used to filter out the desired HSV (hue, saturation, value) colour range. In addition, a threshold value and some gaussian noise is also added to reduce noise. After this the individual colour masks are then applied separately to the video capture. The contours are then detected using the OpenCV function and the convexhull is calculated in a ForLoop based on the contours. Using the convexhull, it then draws the approximate polygon to display the object shape. This coordinate information must be then mapped to VVVV as most of the OpenCV function reside in VL which uses a different coordinate system.



(a) Colour based tracking

(b) Kernelized Correlation Filter (KCF)

Fig. 6. Implementation of marker-based tracking

4.4 Phase 3: Implementation of dynamic 3D projection mapping

In this phase, it aims to map the virtual projections to the moving physical object in real-time. To support for dynamic 3D projection mapping in real time, the system must first be calibrated using the checkerboard to allow for accurate conversions of virtual coordinates to real world coordinates. The 10 x 7 checkerboard was printed and attached to a solid board. Twenty images were taken to in various positions covering all the x, y and z axis. VVVV (2021) recommends the camera calibration reprojection error to be less than 0.5. Here, the camera calibration reprojection error of the system scored 0.4. The camera calibration reprojection error is calculated by using the OpenCV camera calibration function.

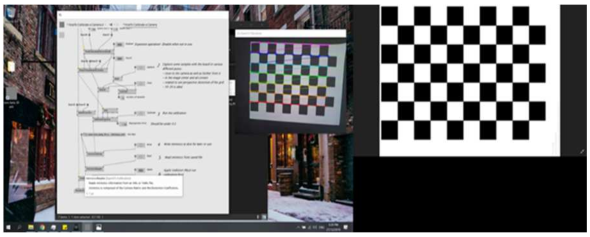


Fig. 7. Implementation of Dynamic 3D Projection Mapping (3D rendering)

The patch shown in Fig. 7 was tested to support for this by using the camera intrinsic and extrinsic parameters. In the preliminary tests, dynamic 3D projection mapping in real time was able to be produced however, the accuracy of the mapping can be improved as sometimes the projections would have some offsets. This is most likely due to the conversion of coordinate systems as shown in Fig. 8.

Additionally, manual calibration technique has also been tested. The manual calibration of the projector involves of marking the physical scene and matching the VVVV renderer to the camera viewpoint using homographic transform applied onto a quad. Homographic transform function is applied here instead of normal transform function as it will allow for individual control and placement of each corner of the renderer, which allows for finer adjustment. Through manual calibration the projector’s renderer view will be the same as the camera view.

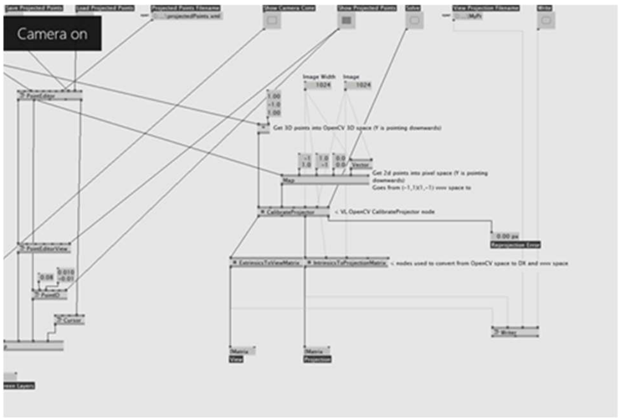


Fig. 8. Implementation of Dynamic 3D Projection Mapping (Mapping Coordinates)

4.5 Evaluation and analysis

This section elaborates details of the test results with comparison of different trackers, techniques for 3D reconstruction and projection mapping.

Table 5 displays the results of the tests conducted on different techniques of the system. The tests for tracking accuracy and consistency involved the subject being tracked to be moved around the scene in different speeds. The tracker must be able to seamlessly track its target and not lose the target when the target is being moved across the scene in three (3) mode of speed. The speed is considered as slow if it is less than 5cm/sec, medium if it is between 5 to 15cm/sec while fast speed was considered to be between 15 to 30cm/sec. The camera and FPS (frames per second) were kept consistent as a control variable.

The tracking precision is calculated based on the concept of intersection over union (IoU) (Khandelwal, 2020) of ground truth bounding box (i.e the actual target) and the predicted bounding box. This allows the assessment of the correct overlaps between the actual target and predicted bounding box. An IoU score higher than or equal to 0.5 is classified as a true positive and an IoU score lesser than 0.5 is classified as a false positive. This test was conducted three (3) times using different objects for each type of tracker and an average IoU score was calculated.

In the detection after loss of tracking test, the object would first be placed on the scene where it is being tracked. After this the object would be removed from the scene and then placed back in the scene. If the tracker is able to detect the object once the object is placed back in the scene, it passes the test.

The colour tracker passes the test for multiple objects of the same colour if it is able to detect two or more objects of the same colour placed part from each other as their own individual objects.

To support real-time reconstruction and record the relative placement of Lego bricks, two techniques were compared. The distance calculation technique involved of comparing the distance and angle from the center of one brick to another. Depending on the this, the Lego brick will lock onto one the adjacent slots available. Whereas the trigger area technique involves of setting up trigger areas (in the shape of a circle) in each of the slots available on a Lego brick. And if the joining Lego block’s center touches one of the trigger areas it will lock onto that respective Lego slot. The evaluation of the accuracy score depends on the numbers of errors to determine the rating scale of 1-5 (from low to high). The high score of 5 would be rated when there is no or one error only; if 2-3 errors it would be rated as 4; if 4-5 errors it would be scored 3; if 6- 8 errors the score would be 2 and when there are more than 8 errors, the score would be rated as low score 1. These errors involve of errors in offsets, wrong slot joining and constant flickering between slots.

For projection mapping two main techniques were tested. In the coordinate lock technique, this involved of locking the camera coordinates retrieved from the colour tracker after the desired object has been created. This allows the projection of the desired shape of the object however, since the colour tracker is still being used the projections can interfere with it and affect the accuracy of the colour tracker. The other technique involves of locking the texture itself once the shape has been constructed. This no longer requires the colour tracker to be active once the shape is created therefore, the projections cannot interfere with the colour tracker.

Table 5. Comparison of various system techniques tested.

Comparison of marker and markerless trackers			
	KCF Tracker	**Aruco Tracker**	**Fiducial Tracker**
Tracking Accuracy & Consistency	Slow, Medium Speed (<10 cm/sec)	Slow Speed (<5 cm/sec)	Slow Speed (<5cm/sec)
Tracking Precision	0.71	0.95	0.74
Detection after loss of tracking	No	Yes	Yes
Requires marker	No	Yes	Yes
Comparison of colour trackers			
	Freeframe Dshow9 colour Tracker	**Proposed Colour Tracker**	
Tracking Accuracy & Consistency	Slow Speed (<5 cm/sec)	Slow, Medium, Fast Speed (<30 cm/sec)	
Tracking Precision	0.82	0.85	
Detection after loss of tracking	Yes	Yes	
Tracking Multiple objects of same colour	No	Yes	
Comparison of techniques for 3D reconstruction			
	Distance calculation	**Trigger area**	
Accuracy	Low (2/5) [Requires fixed distance from the camera]	Medium (3/5) [Requires fixed distance from the camera]	
Comparison of projection mapping techniques			
	Coordinate Lock	**Texture Lock**	
Requires switching to another camera view	Yes	Yes	
Projections interfering with colour tracker	Yes	No	

4.6 Implementation strengths & issues

The strength of the current project is 3D projections allowing the users to move tangible objects. This can improve the user's interaction as it can allow for a more natural interaction. Additionally, the support for real-time reconstruction allows for users to build their own shapes instead of just using the pre-defined shapes. In the tests conducted, the system was able to correctly detect and project the desired texture onto the specific structure constructed of up to 10 Lego bricks placed in random positions.

However, there were some issues faced in conversion of 3D virtual coordinates onto real world coordinates where the projections are mapped. This is due to the dynamic nature of the 3D projection mappings and therefore the accuracy in projections is a limitation. Furthermore, some issues faced in tracking of individual Lego blocks to be able to constantly detect the final shape due to camera occlusions. To improve the visibility of the projections most projection mapping applications are carried out in a low lighting environment however, this can be challenging for the camera to perform tracking of marker features.

5. Conclusions

The tracker-based real-time interactive SAR system was successfully developed after testing and analysis. The system projects the SAR virtual contents correctly mapped onto the real-world 3D objects with support for real-time 3D reconstruction. This has introduced a new interaction method that allows to create object detection and tracking in a real-time in the physical SAR platform. Moreover, it would have designed and developed the physical SAR platform suitable for dynamic projection mapping. The benefits from the proposed SAR platforms are stated below.

1. Static Vs Dynamic: Most projection mappings are static and therefore with the addition of dynamic content it can involve users in a more effective manner

2. 2D Vs 3D: It is easier to visualize complex data and ideas/plans through 3D projection mapping compared to 2D

3. AR Vs SAR: SAR provides a more seamless blend between the virtual and the real world compared to AR as the virtual content is projected to the real-world

4. Touch Vs tangible interaction: Tangible interaction provides natural ways of user interaction compared to touch interaction.

6. References

Al-Megren, S. and R. A. Ruddle, 2016. Comparing Tangible and Multi-touch Interaction for Interactive Data Visualization Tasks, TEI '16, Eindhoven, Netherlands.

Ball, C., 2018. Augmented/Mixed Reality for Events: The Pros and Cons for Each Medium and the Likelihood of Adoption, The meeting technology professionals. Available online at: https://www.corbinball.com/article/36-mobile-and-wireless-technology/243-vrvsar (accessed on July 10, 2021)

Blender. 2.83, 2020. Available online at: https://www.blender.org/ (accessed on July 10, 2021)

Cortes, G., E. Marchand, G. Brincin, and A. Lecuyer, 2018. MoSART: Mobile Spatial Augmented Reality for 3D Interaction with Tangible Objects, Front. Robot. AI, 5(93). Available online at: https://doi.org/10.3389/frobt.2018.00093 (accessed on July 10, 2021)

Fleischmann, O. and R. Koch, 2016. Fast projector-camera calibration for interactive projection mapping, IEEE.

Guo, Y., S. C. Chu, Z. Liu, C. Qiu, H. Luo, and J. Tan, 2018. A real-time interactive system of surface reconstruction and dynamic projection mapping with RGB-depth sensor and projector, International Journal of Distributed Sensor Networks, 14(7), Available online at: https://doi.org/10.1177%2F1550147718790853 (accessed on July 10, 2021)

Khandelwal, R., 2020 January 6. Evaluating performance of an object detection model, what is mAP? How to evaluate the performance of an object detection model? Towards data science, Available online at: https://towardsdatascience.com/evaluating-performance-of-an-object-detection-model-137a349c517b (accessed on July 10, 2021)

Kim, H., I. Takahashi, H. Yamamoto, S. Maekawa, and T. Naemura, 2014. MARIO: Mid-air Augmented Reality Interaction with Objects, Entertainment Computing, 5(4), 233-241, Available online at: https://doi.org/10.1016/j.entcom.2014.10.008 (accessed on July 10, 2021)

Laviole, J., 2012. Spatial augmented reality for physical drawing, UIST 12: The 25th Annual ACM Symposium on User Interface Software and Technology Cambridge Massachusetts USA, October 7

- 10, Available online at: https://dl.acm.org/doi/10.1145/2380296.2380302 (accessed on July 10, 2021)

Makarov, A., 2021, February 1. 10 augmented reality trends to watch in 2021: The future is here. Available online at: https://mobidev.biz/blog/augmented-reality-future-trends-2018-2020 (accessed on July 10, 2021)

Mendes, M., J. Almeida, J., H. Mohamed, and R. Giot, 2019. Projected Augmented Reality Intelligent Model of a City Area with Path Optimization, Algorithms, 12(7), 140, Available online at: https://www.mdpi.com/1999-4893/12/7/140 (accessed on July 10, 2021)

Milgram, P., and F. Kishino, 1994. Taxonomy of Mixed Reality Visual Displays, IEICE, Transactions on Information and Systems, vol. E77, 1321-1329.

Mousavi, H. H., M. Khademi, L. Dodakian, S. C. Cramer, and C. V. Lopes, 2013. A Spatial Augmented Reality rehab system for post-stroke hand rehabilitation. Stud Health Technol Inform. 184:279-85. PMID: 23400171

Open. CV., 2021. Available online at: https://opencv.org/ (accessed on July 10, 2021)

Park, M. K., K. J. Lim, M. K. Seo, S. J. Jung, and K. H. Lee, 2015. Spatial augmented reality for product appearance design evaluation, Journal of Computational Design and Engineering, 2, 38–46.

Park H. B., 2017. Available online at: https://www.youtube.com/watch?v=r4DTMedEH4U (accessed on July 10, 2021)

Wilson, A. D., 2005. PlayAnywhere, ACM Press. Available online at: https://dx.doi.org/10.1145/1095034.1095047 (accessed on July 10, 2021)

Winder, J., and K. Larson, 2017. Bits and Bricks Tangible Interactive Matrix for Real-time Computation and 3 D Projection Mapping, Paper presented at Future Technologies Conference (FTC).

VVVV. group., 2021, January 24. VVVV a multipurpose toolkit. Available online at: https://vvvv.org/ (accessed on July 10, 2021)

AUTHOR BIOGRAPHIES

Aye Chan Zay Hta is a master's student at the Department of Computing and Information Systems, School of Engineering and Technology, Sunway University, Malaysia. He graduated with a BSc (Hons) in computer science (2019), validated by Lancaster University and a Diploma in information technology (2017) from the same university.

Yunli Lee is an associate professor in the Department of Computing and Information Systems, School of Engineering and Technology at Sunway University. She was awarded a BIT (Hons) in software engineering in 2002 from Multimedia University, Malaysia. In 2004, she received hermaster's degree in software from Dongseo University, South Korea and obtained her PhD in engineering (Digital Media) in 2009 from Soongsil University, South Korea. She is currently a senior member of the IEEE society, Professional Technologist of MBOT and Malaysia Director of the International Association for Convergence Science & Technology (IACST). Her current research interests include ultrasound imaging, the time series of FOREX data, and augmented reality technology.

Wai Chong Chia is a senior lecturer in the department of computing and information systems, school of engineering and technology at Sunway University. He received his BEng(Hons) in Electrical and Electronics Engineering from University of Nottingham Trent in 2006. He then continues to pursue his studies and receive his MSc in Electronics Engineering and Doctor of Philosophy from The University of Nottingham in 2008 and 2013 respectively. His main research is focused on visual processing for wireless sensor networks, embedded systems and mobile applications.

Hybrid Black Widow Optimization and Variable Neighborhood Descent Algorithm for Traveling Salesman Problem

Ayad Mohammed Jabbar[1]*, Ku Ruhana Ku-Mahamud[2]
[1] Computer Science Department, Shatt Al-Arab University College, Iraq
[2] School of Computing, Universiti Utara Malaysia, Malaysia
* Corresponding author E-mail: ayadjb@gmail.com

(Received 11 July 2021; Final version received 22 July 2021; Accepted 7 September 2021)

Abstract

Local search algorithms in general are better than population-based algorithms in the terms of exploitation capability in finding more local regions in the search space which provide more ability to explore search space in finding global regions. Black widow optimization (BWO) algorithm is one of the best population-based algorithms which was proposed in 2020 to solve engineering optimization problems. However, this algorithm has a limitation in the exploitation of search space and reactivate a search when stagnation occurs during the algorithm run. Thus, deep search and effectively exploring the search space are not possible during the algorithm run. To overcome these drawbacks, this study proposes two modifications to the BWO algorithm. The first modification is the integration of variable neighborhood descent used to enhance the exploitation process in finding more local regions in the neighborhood during the algorithm run. The second modification focuses on the reactive search process by integrating a new convergence indicator for the algorithm during the algorithm run and online reactive search process. Two benchmark datasets were used to evaluate the proposed modification. The minimum tour distance provided by each algorithm has been used as the performance metric in determining the credibility of the hybrid BWO algorithm and results have been compared with best-known algorithms include African buffalo optimization (ABO), ant colony optimization (ACO), artificial bee colony (ABC), particle swarm optimization (PSO) and a hybrid algorithm consisting of harmony search, particle swarm and ACO (HPSACO). The hybrid BWO algorithm has produced better minimum tour distance compared to ABO, ACO, ABC, PSO and HPSACO algorithms which demonstrate that the hybrid BWO can be applied to solve several optimization problems including vehicle routing problem, classification and clustering.

Keywords: Exploration, Exploitation, Local search, Neighborhood search, Swarm algorithms, Traveling salesman problem.

1. Introduction

The optimal solution in the field of artificial intelligence refers to the best solution that can be obtained from the search space compared with other several solutions provided from the same search space (Desale et al., 2015; Jabbar, Ku-Mahamud, & Sagban 2019a, 2019c). This type of solution can be found when solving any NP-hard problems that have complex search space have several landscapes. Examples of NP-hard problems are classification (Stegherr, Heider, & Hähner, 2020), clustering (Hossain et al., 2019), and feature selection (Venkatesh & Anuradha, 2019). The search space of the problem is difficult to solve within the estimated time and requires special algorithm to provide a stochastic search guided by the objective function and randomness covering a wide area of search space (Dao, Abhary, & Marian, 2015). This kind of search is the main foundation of algorithms and known as metaheuristics, which combines heuristic methods in high-level metaphors to find optimal or near-optimal solution in a reasonable time. Examples of metaphors in real life include foraging behavior, memory, annealing, evolution, reproduction style and cannibalism, these metaphors can be represented by several algorithms including ACO (Al-Behadili, Sagban, & Ku-Mahamud, 2020a), tabu search (TS) (Ghany et al., 2020), simulated annealing (SA) (Moriguchi, Ueki, & Saito, 2015), genetic algorithm (GA) (Das & Pratihar, 2018), and black widow optimization algorithm (BWO) (Hayyolalam & Pourhaji Kazem, 2020; Abuhamdah, 2020). The BWO is one of the swarm intelligence algorithms inspired by black widow spiders, which simulates the spider mating process in nature (Houssein et al., 2020). The algorithm has three important stages which simulates the real behavior of black widow. These stages include mating, which starts when the male enters the web of female; reproduction and cannibalism, which start by hatching the egg and offspring engagement; and increasing the density of the population by keeping only strong spiders which is known as sibling cannibalism. However, the main issue is how to achieve better exploitation and exploration of the search space and avoiding local optima problem in the process to obtain the optimal result rapidly and with high accuracy. The BWO the algorithm has limitations in terms of exploitation of the search space and reactivating a search when stagnation occurs during the algorithm run. The algorithm should improve the population iteratively until it converges to local optima and then the reactive the search process by updating its population with new fresh population located far from the region that has that local optimum. Considering this limitation, scholars explore other algorithms to solve the limitation which resulted in a hybrid algorithm. This process can be achieved using metaheuristics algorithms, which use the neighborhood change. An example of these algorithms is the variable neighborhood descent (VND), which descent to the regions that have local optima and can escape from these regions according to the designed VND framework (Duarte et al., 2016). In the optimization problem, the region that has good solution certainly contains neighborhood regions that have better solutions. Thus, several close neighborhoods should be explored to find global solutions by generating several landscapes using the VND algorithm. The algorithm has a befit of generating several landscapes during the algorithm run whereby it can increase the diversity of solutions with different solutions. However, an important issue that should be considered is moving from one landscape to another. Despite of its importance, it is not always sufficient when no knowledge about the exploration state is provided. Hence, more time is required to perform exploitation than exploration. Based on this consideration, two modifications are proposed. VND is used to enhance the performance of the BWO algorithm in terms of exploitation capability to find more solutions around the best regions in the search space. VND enhances the BWO algorithm to find more local regions by improving the neighborhood search during the algorithm run. Reactive search process is proposed as the second modification by integrating a new indicator of the convergence of the algorithm during the algorithm run. The reactive search will enhance the search process by moving the search into new promising region and keeping the history of the search to use them as a guide for future search in advanced iterations. Both modifications will enhance the balance between the exploration and exaptation. Finally, the performance of the proposed algorithm will increase. These two modifications are recruited in BWO to avoid convergence because of its limitation in exploiting the search in finding more local regions at the best so far region and reactivating the search space during the algorithm run. Reactive search is integrated into the algorithm, which will automatically reactivate the search process when algorithm falls into local optima or converges to the same solution.

This article is organized as follows. Section 2 shows the related works of swarm based-algorithms. Section 3 elaborates the proposed hybrid BWO algorithm with its formulation for the traveling salesman's problem. Section 4 describes the benchmark datasets used in the experimental results. Section 5 states the conclusion and future works.

2. Related works

Solving NP hard problem such as TSP where finding the minimum tour distance between all cities is not an easy task, especially for a large number of cities and the search space is complex. Finding an optimal solution where the objective function reaches its minimum value at an acceptable time has several landscapes. This process of reaching the optimal solution is sometimes useless, as the time required to solve these problems may exceed the usefulness of the solution. Here the need arose to use algorithms that have the ability to produce good solutions in an ideal time. Those algorithms are categorized as optimization approaches targeted at finding near or optimal solutions in an ideal time. The optimization approach shown in Fig. 1, classified into several approaches which include estimation approach, exact approach, and approximate approach. The exact approach requires an exponential time to solve the hard problem due to its process to find all solutions. The estimation approach uses a previously defined range of inputs such as used in parameter problem to solve the problem according to the defined inputs. The last approach is the approximate approach can be classified into single-based solution such as local search and population-based techniques (Al-Behadili, Ku-Mahamud, & Sagban, 2020b; Duarte et al., 2016). Local search methods, such as TS and SA, perform neighborhood search to modify single solution by exchanging segments of its components to produce better solutions. Meanwhile, in the population-based techniques such as ACO and BWO, more than one solution are used iteratively during the algorithm run (Al-Behadili, Ku-Mahamud, & Sagban, 2019c; Sicilia et al., 2016). The search process is guided in different processes according to metaphor characteristics employed in the algorithm. Examples of these characteristics' areas are the acceptance criterion and cooling schedule in SA algorithm (Yang & Yang, 2014), neighbor choice in TS algorithm (Zhou et al., 2013), recombination, mutation, and selection in GA algorithm (AL-Behadili, Ku-Mahamud, & Sagban, 2020a; Stegherr et al., 2020), mating, reproduction and cannibalism (Rasekh & Osawa, 2020), siblings in BWO, and pheromone update and probabilistic construction in ACO algorithm (Jabbar, Ku-Mahamud, & Sagban, 2019b). However, there is a big deference between both approaches regarding the stigmergy which is a key role in the nurturing of society that does not exist in the evolutionary-based approach. The swarm-based approach includes different algorithm such as ACO, artificial bee colony (ABC), particle swarm optimization (PSO) and BWO. These algorithms have the stigmergy principle that represents the medium for information transformation. A typical example is pheromone trails that leads to organization of ants in ACO, while in BWO represents the attraction between the male and female. However, the main differences between each algorithm are the algorithm memory and how to represent the search space of the problem. An example of that PSO its memory represents the population of particles, ACO is represented by pheromone matrix that retain the information of ants and ABC represented by its population of bees. The BWO algorithm followed the same procedure of the swarm algorithms, where its memory is represented by the population of the black widows.

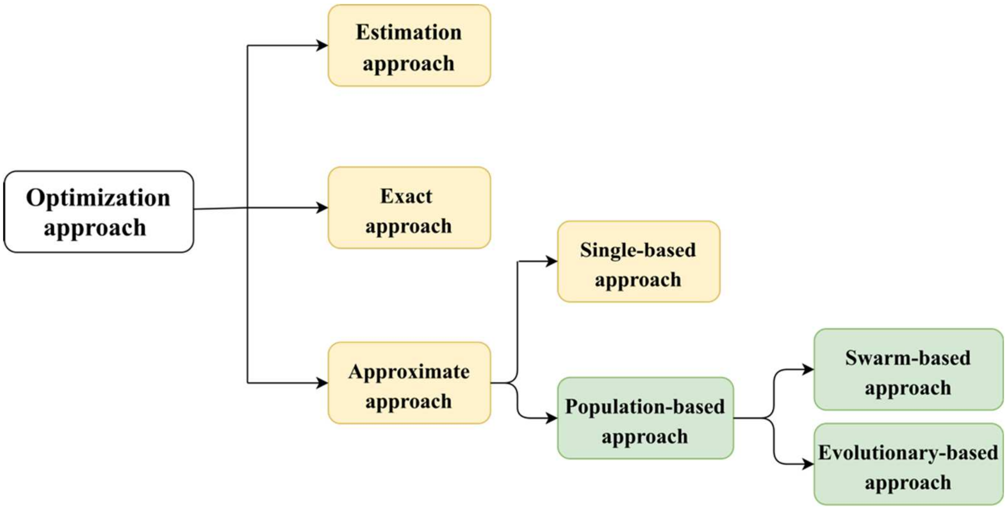


Fig. 1. Taxonomy of optimization approach

Several related works have been proposed in the literature such as Tan et al. (2020) who proposed a hybrid POS algorithm and hill climbing (HC) for high school timetabling problem. The proposed algorithm has two modifications in PSO and HC algorithms. The first modification is the solution transformation of the mutation and crossover operations while the second modification is to increase the efficiency of the exploration and exploitation in search space using HC algorithm. However, it is found that the HC algorithm only accept the candidate solutions that have better fitness. Thus, it limits the exploration capability of PSO algorithm in the terms of looking for global candidate solutions. Other similar research was proposed by Goh et al. (2020) as a hybrid local search algorithm to address the post enrolment course timetabling. There are two phases in the proposed hybrid local search algorithm. The first phase is to find a feasible solution, while the second phase focuses on minimizing the soft constraint of the generated feasible solution from the first phase. In finding a feasible solution, the tabu search (TS) with sampling hybrid algorithm and perturbation with iterated local search (ILS) hybrid algorithm were employed. Simulated annealing with reheating (SAR) algorithm and two preliminary runs (SAR-2P) algorithm are proposed to minimize the soft constraint of the feasible solution. The proposed algorithm has a drawback in terms of exploration capability, but in other aspects it shows promising results. Thus, it requires other algorithm such as ACO algorithm to overcome the exploration problem. Another related research has proposed a re-randomization method coupled with variable neighborhood search (VNS) to solve the optimal allocation of a fixed set of experimental units (Hore, Dewanji, & Chatterjee, 2016). The re-randomization method increases the probability to find more reasonable initial allocations by incorporating the randomization during the search process. However, it would be better to include various criteria to control the randomization of the search process and to avoid losing the exploitation part. An algorithm by coupling variable neighborhood search (VNS) with stochastic search to improve the exploration of VNS has been proposed by Hore, Chatterjee, and Dewanji (2018). The purpose of this coupling is to avoid the problem of local optima solution provided by VNS. Three modifications have been added to increase the stochastic search of the algorithm. These are the initial tour, construction of neighborhood and new stopping criteria. Although the proposed algorithm showed promising result, its time complex was long. Other related research to improve VNS by keeping the characteristics of the best solution during the algorithm run which often be kept and used to obtain promising neighboring solutions (Hore, Dewanji, & Chatterjee, 2014). This kind of VNS algorithm has recently been successfully applied in the field of design of experiments by adding optimum allocation of experimental units with known predictors into two treatment groups. However, adding local search to swarm algorithms is required because of the exploitation capability in finding more local regions. From the literature, swarm algorithms such as BWO remains poor as its exploitation strategy is incapable to intensify the search of local regions. Thus, integration of a local search algorithm such as VND is required with an indicator to check the convergence state of the algorithm. This indicator reactivates the search process during the algorithm run and provide more optimal solutions.

3. Proposed hybrid algorithm

Black widow optimization algorithm, which was proposed by Hayyolalam and Pourhaji Kazem (2020) is one of the best population-based swarm algorithms to solve NP-hard combinatorial problems (Sathish & Ananthapadmanabha, 2021; Sheriba & Rajesh, 2021). The algorithm is inspired by the black widow spider's nature behavior representing the mating process in spiders. The algorithm has three important stages which simulates the real behavior of the black widow. These stages are classified as follows the matting process of black widow spider, the reproduction style and cannibalism and the sibling cannibalism.

The mating process starts when the male enters the web of the female. This process occurs when a female black widow desire to mate with one of the males by attracting them using her pheromone. This process is followed by laying her eggs on the sock and wait for hatching. The following stage is reproduction style and cannibalism, which start by egg hatching and offspring engagement. Immediately during or after the matting, the male is consumed by the female black widow, which is a natural behavior that can be seen in different invertebrate societies. Researchers believe that the behavior of male might confer the chance to increase the number of eggs, resulting in greater chance of continuity of offspring. The next stage is where eggs are hatched and spend the time on the web feeding on the yolk and molt. During this time, an important natural behavior known as sibling cannibalism can be observed. This stage increases the density of the population by keeping only strong spiders. The strong spider eats the weak siblings, female black widow eats her husband, and other case such as other spiders consume their mothers. All kind of sibling cannibalism changes the diversity of the population, thus new generation can be better than the older generation. The BWO algorithm employs sibling cannibalism to improve the search process during the run by achieving only high-quality solutions. However, not all cases are desirable in optimization problems, such as improving the quality of solution guided by the objective function will lose its diversification capability (Stützle & Hoos, 2000). In the same manner, high diversification forces the algorithm to lose its exploitation capability. This research proposes a modification which includes a reactive search space during the algorithm run and simultaneously finds the optimal local regions. The mechanism is established by copying the best individuals from the original population and keep them in a temporary memory. The algorithm starts to improve existing individuals in temporary memory until no further improvement can be obtained. This step is considered as the exploration step when reactive search is performed, moving the search process to another promising global region. In each step of the improvement, the best individuals move from the original population to the temporary memory. Through this step, the second modification is added, which apply VND to increase the process of finding more promising local regions in the search space. Variable neighborhood descent enhances the BWO algorithm to find more local regions by improving the neighborhood search during the algorithm run through generation of more landscapes. It moves the best solutions generated during the run and checks when stagnations occur to restart the history of search process as shown in Fig. 2.

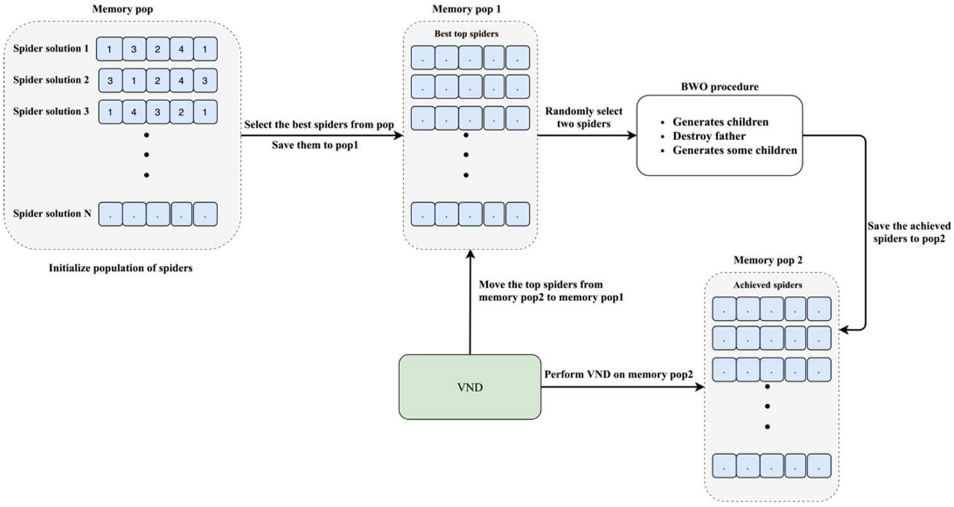


Fig. 2. The mechanism of the modifications

The algorithm starts by initializing the population of spiders in the memory pop (single spider represents single solution). The memory pop1 aims to keep the best solutions during the algorithm run and use them to refresh the search space. This is established after copying the best solutions (best spiders) from the population pop and save them to pop1. The process of the algorithm starts by selecting two solutions, then generate the children and destroy the fathers and some children. The rest of the solutions after this process is kept in the memory called pop2. The next process performs VND to find more optimal solutions and the best produced solution is select from pop2 and saved in pop1. This process ensures deep searching in the local regions of the best solutions, thereby finding the best neighborhood of all in pop2. Once the neighborhood is determined, the best top solutions in pop2 are moved to pop1. However, if convergence occurs in the algorithm, then it will deselect the best solution from pop and save them in pop1, thereby reactivating the search process during the algorithm run. Variable neighborhood descent is used in this research due to its capability of convergence to the optimal solution during long time, giving more time to explore the search space. Different landscapes refer to different neighborhood structures, thus different solutions can be used to avoid the algorithm being trapped at local optima problem. In this research, two operations are performed to provide several landscapes including the pair-swap and inversion. The pair-swap will swap two cities in randomly manner, while the inversion operation inverts a subsequence of cities between two swapped cities which are randomly selected from the solution. This algorithm has the benefit of generating several landscapes during the algorithm run, and it can increase the probability of diversity of solutions and avoid the case of losing the diversity in the population and other memories. VND generates several landscapes iteratively. Thus, the algorithm can use different neighborhood during the run to avoid the local optima as shown in Fig. 3 (El-Ghazali Talb, 2009). The VND algorithm starts by generating a set of different neighborhood structures Nl (l=1,…,lmax). Let N1 be the first neighborhood located in the local regain of the initial solution x. Thus, if no improvement in the current local region based on the fitness function of N1 compared with x occurs, then the neighborhood structure will be changed from N1 to N1+1 until the best neighborhood structure is found. Otherwise, the fitness function of x is kept. The pseudo-code of the variable neighborhood decent algorithm is illustrated in Fig. 4 (El-Ghazali Talb, 2009).

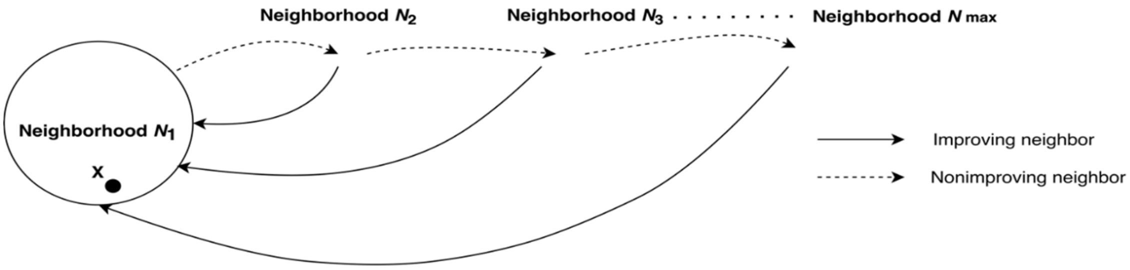


Fig. 3. Principle of variable neighborhood descent algorithm

```
VND algorithm
Input     A set of neighborhood structures $N_l(l = 1, ..., l_{max})$.
Output    The best found solution $(best)$
Step :1   $x = x_0$  %% generate the initial solution
Step :2   repeat
Step :3       $l = 1$
Step :4       while $l \le l_{max}$ do
Step :5           Find the best neighbor $x^`$ of $x$ in $N_{l(x)}$
Step :6           if $f(x^`) < f(x)$ then
Step :7               $x = x^`$
Step :8               $l = 1$
Step :9               else
Step :10              $l = l + 1$
Step :11          end-if
Step :12      end-while
Step :13  until no improvement exists in any $l_{max}$ neighborhoods
Step :14  return $x$;
```

Fig. 4. Pseudo-code of VND algorithm

VND selects a single solution x from the memory in sequence for further improvement by generating a set of neighborhoods and select the first obtained neighborhood that increases the quality of solution (better minimum tour distance). The output of the VND algorithm is a solution has better quality replaced with the original solution x. This process is iteratively performed, as shown in step 6 of Fig. 4 until the best-found solution is obtained in step 14. The complete process of the proposed hybrid BWO algorithm is illustrated in Fig.

```
Proposed hybrid BWO algorithm
Input     TSP dataset
Output    A minimum tour distance $s$
Step :1   Initialize all parameters ($pr$, $nr, nm$, $q_0$, $p$‘ $limit$, $N_l$)
Step :2   Initialize population of spiders $pop$ and evaluate each one
Step :3   Select best $nr$ solutions in $pop$ and save them in $pop1$ based on procreating rate ($pr$)
Step :4   while ($iteration$ $<$ $N_l$ ) do
Step :5           repeat
Step :6               Randomly select two solutions as parents from $pop1$
Step :7               Generates D children
Step :8               Destroy father
Step :9               Based on the cannibalism rate ($CR$), destroy some children
Step :10              Save the remain solutions into $pop2$
Step :11          until stopping $nv$ is met
Step :12      Perform VND on $pop2$ and copy the best $nr$ solutions in $pop2$ and save them in $pop1$
Step :13      if ($av$(solutions in old $pop2$) == $av$(solutions in current $pop2$)) then
Step :14              $limit++$
Step :15      end-if
Step :16      if ($limit >$ $limit\ max$) then
Step :17          Select best $nr$ solutions in $pop$ and save them in $pop1$
Step :18      end-if
Step :19      repeat
Step :20          Select a solution from $pop1$
Step :21          Mutate randomly one chromosome of the solution and generate a new solution
Step :22          Save the new one into $pop3$
Step :23      until stopping $nm$ is met
Step :24      Update $pop$=$pop2$+$pop3$
Step :25  end-while
Step :26  return $s$
```

Fig. 5. Pseudo-code of proposed hybrid BWO algorithm

The hybrid BWO starts by pop1 initializing all parameters and the population of spiders, pop, in steps 1 and 2. In step 3, the algorithm selects the best nr solutions from pop and save them to, which represents the best solutions in the population to be matted latter.

In step 4, the algorithm starts its iterations compared with the maximum number of iterations N_l. In step 5, the algorithm starts its cycling compared with its procreation rating to the maximum allowed procreation which is equals to 50 in this research. The next step is to randomly select two solutions from pop1. In the current time, the best nr solutions exist in pop1. Steps 7 and 8 include matting using the crossover operator between two individuals and destroying the father after the matting process. To introduce more diversity, several children and mothers are destroyed in step 9 and save in pop2 as shown in step 10. The next step highlights the proposed modification where VND is used on all solutions in pop2. However, in general, all local search algorithms provide good solution but is easily stacked at the local optima. Thus, in this research, another modification is added to check if the algorithm has sufficient diversity solution or not and to decide reactivating the search space by adding new solutions into pop1 from pop, as shown in steps 12-18. In the following steps, the algorithm employs mutation to improve the solution. In step 24, the algorithm will update pop by using all solutions that exist in pop2 and pop3. The output of the algorithm is the best solution found during the run as shown in in step 26.

4. Experiment and evaluation

Prepare The experiments have been conducted in two scenarios to evaluate the performance of the hybrid BWO algorithm. In each scenario, different set of TSP datasets (Jia, 2015; Odili & Mohmad Kahar, 2016) were used. The first set includes eight datasets (berlin52, st70, eil76, pr76, kroa100, eil101, ch150, and tsp225). These datasets have been used in a study conducted by Odili and Mohmad Kahar (2016) and the results also available in (TSPLIB, 1995). Results from the experiments conducted on the first set of the datasets have been compared to African buffalo optimization (ABO), ACO, and artificial bee colony (ABC). The second set consists of ten (10) benchmark datasets including the att48, st70, eil76, pr152, gil262, rd400, pr1002, d1291, fnl4461, and brd14051 (Jia, 2015) and results are available in TSPLIB (1995). The algorithms that have been used in the second part of the experiment are ABO, ACO, PSO, and harmony search, particle swarm and ACO (HPSACO) algorithms. Table 1 shows 16 TSP datasets that differ in the number of cities and the optimal distance for each dataset. Note that both st70 and eil70 are mentioned in both benchmarks (Jia, 2015; Odili & Mohmad Kahar, 2016).

Table 1. Benchmark characteristic and optimal solution of each dataset

Dataset	Number of cities	Optimal
att48	48	33522
st70	70	675
Eil76	76	538
pr152	152	73682
gil262	262	2378
rd400	400	15281
pr1002	1002	259045
d1291	1291	50801
fnl4461	4461	182566
brd14051	14051	469385
berlin52	52	7542
pr76	76	108159
kroa100	100	21282
eil101	101	629
ch150	150	6528
tsp225	225	3916

The performance of the proposed algorithm is evaluated based on two criteria including average distance produced by each algorithm calculated as the average Euclidean distance through all cites starting from the start city to all cites and returning to the start city; and the best solution provided by the algorithm in all number of runs to indicate which algorithm surpasses the produced minimum tour distance. For comparison, all parameter settings are fixed for all algorithms as shown in Table 2 in accordance with a previous study (Odili & Mohmad Kahar, 2016).

Table 2. Parameters of all algorithms

ABO	**ACO**	**ABC**	**Hybrid BWO**
Population =40	Ants =D*	Population= D*	Population=40
$\boldsymbol{m.k}$=1.0	β =0.5	ϕij= rand (−1, 1)	$N_l = 1000$
bgmax/bpmax= 0.6	ρ =0.65	ωij= rand (0, 1.5)	$pr = 0.6$
$lp1/lp2$=0.5	α =1.0	SN= NP/2	Limit max=50
$w.k$ =0.1	$Q = 200$	Limit= D^* SN	$nr = 20$
N/A	$q0 = 0.9$	Max cycle number =500	$CR = 0.44$
N/A	N/A	Colony =50	$nm = 0.4$
Total number of runs =50			

Results of the first scenario experiments are shown in Table 3 where best performances are highlighted. For each dataset, the minimum distance (best) and mean distances of the algorithms for 50 runs were recorded.

Table 3. First scenario results

Problem	Tour Length	Algorithm			
		ABO	ACO	ABC	Hybrid BWO
berlin52	Best	7542	7548.99	9479.11	7542
	Mean	7616	7659.31	10,390.26	**7609**
st70	Best	676	696.05	1162.12	676
	Mean	678.33	709.16	1230.49	**677.12**
eil76	Best	538	554.46	877.28	538
	Mean	563.04	561.98	931.44	**541.23**
pr76	Best	108167	115,166.66	195,198.90	108171
	Mean	**108,396**	116,321.22	205,119.61	111,126
kroa100	Best	21311	22,455.89	49,519.51	21298
	Mean	22163.8	22,880.12	53,840.03	**21832.91**
eil101	Best	640	678.04	1237.31	634
	Mean	640	693.42	1315.95	**638.73**
ch150	Best	6532	6648.51	20,908.89	6531
	Mean	**6601**	6702.87	21,617.48	6923.13
tsp225	Best	3917	4112.35	16,998.41	3920
	Mean	3982	4176.08	17,955.12	**3961.23**

Minimum tour distance has been obtained by the hybrid BWO algorithm in six datasets (75% of the datasets) followed by the ABO algorithm which managed to secure the best minimum performance in two datasets. This shows that the hybrid BWO algorithm can find better solution in the local region of the best solution by using VND. Results of the experiments in the second scenario are shown in Table 4. The performance of the proposed hybrid BWO algorithm surpasses all algorithms in eight (8) datasets while in second place is the ABO algorithm, which produces the best results in two datasets. This is again attributed to the use of VND in the BWO algorithm to solve the local optima problem

Table 4. Second scenario results

Problem	Tour Length	Algorithm				
		ABO	PSO	ACO	HPSACO	Hybrid BWO
att48	Best	33524	33734	33649	33524	33523
	Mean	33579	33982	33731	33667	**33553**
st70	Best	676	691.2	685.7	680.3	676
	Mean	**678.33**	702.6	694.7	698.6	677.12
eil76	Best	538	572.3	550.7	546.2	538
	Mean	563.04	589.1	560.4	558.1	**541.23**
pr152	Best	73730	75361	74689	74165	73722
	Mean	73990	75405	74936	74654	**73934**
gil262	Best	2378	2513	2463	2413	2378
	Mean	2386	2486	2495	2468	**2379**

rd400	Best	15301	16964	16581	16067	15298
	Mean	15304	17024	16834	16513	**15300**
pr1002	Best	259132	278923	269758	267998	259740
	Mean	**261608**	279755	271043	269789	269796
d1291	Best	50839	53912	52942	52868	50811
	Mean	50839	54104	53249	52951	**50823**
fnl4461	Best	182745	199314	192964	191352	182712
	Mean	183174	199492	194015	192585	**183123**
brd14051	Best	469835	518631	505734	498471	469745
	Mean	479085	519305	511638	503594	**478257**

The proposed hybrid BWO algorithm can find better tour distance compared to other algorithms. The algorithm iteratively looks for the local regions around the best solution found during the run. This is used as the indicator to reactivate the search process and update the algorithm with new populations. The purpose of intensifying the search process is to force the algorithm to perform deep searching. Once it converges to a local optimum, it jumps to other region in the search space, thereby increasing the probability to explore more regions of the search space. Fig. 6 displays the final comparison is performed between both the proposed hybrid BWO algorithm and VND algorithm to show the benefit of the hybridization. Results show that the proposed algorithm outperforms VND in all datasets. Note that the results of both algorithms have been normalized to produce better presentation

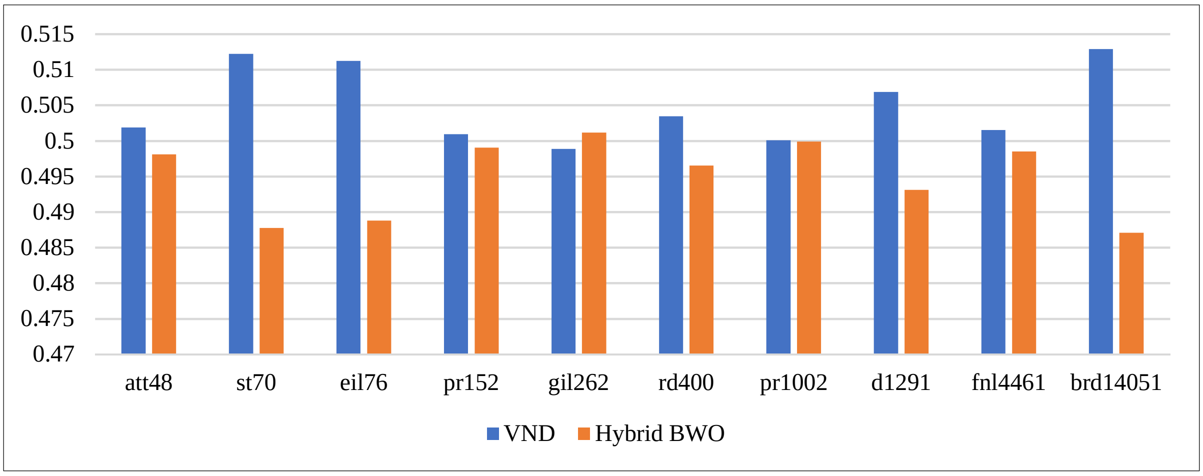


Fig. 6. VND vs. hybrid BWO algorithm

5. Conclusions

This research aims to improve the performance of the BWO algorithm in terms of finding better tour which is the minimum tour distance. The proposed hybrid BWO algorithm has solved the problem by using two modifications and this can be considered as the contribution of this study. First, the use of VND increases the probability to find more solutions with better tour distance and search deeply around the best regions that have the best solution so far. Secondly, an indicator is used to check the convergence of the results. This indicator will reactivate the search process of the algorithm by updating its memory with new populations located in different regions of the search space. The results of the experiments showed that the proposed hybrid BWO algorithm has obtained the minimum tour distances compared to ACO, ABO, ABC, PSO, and HPSACO algorithms. Both modifications have enhanced the BWO algorithm by improving the exploration and exploitation performances during the algorithm run. Despite the superior performance obtained by the BWO algorithm with the two modifications, there are two limitations that can be used as a guide for researchers who are interested in this BWO algorithm. The first one is the time-consuming process as a result of using VND, and the second is the number of parameters used in the BWO algorithm. The proposed limit parameter requires more attention in terms of optimization using adaptive and self-adaptive strategies to find the best value. Future research can also be focused on the application of the hybrid BWO algorithm in several optimization problems, such as vehicle routing, clustering, and classification with other neighborhood search algorithms.

Acknowledgements

The Ministry of Higher Education Malaysia has funded this study under the Transdisciplinary Research Grant Scheme, TRGS/1/2018/UUM/02/3/3 (S/O code 14163).

References

Abuhamdah, A. (2020). Adaptive black widow optimisation algorithm for data clustering. Journal Mathematics in Operational Research, 2002, 1–25.

Al-behadili, H. N. K., Ku-mahamud, K. R., & Sagban, R. (2019). Annealing strategy for an enhance rule pruning technique in ACO-based rule classification. Indonesian Journal of Electrical Engineering and Computer Science, 16(3), 1499–1507

Al-Behadili, H. N. K., Ku-Mahamud, K. R., & Sagban, R. (2020a). Hybrid ant colony optimization and iterated local search for rules-based classification. Journal of Theoretical and Applied Information Technology, 98(04), 657–671.

AL-Behadili, H. N. K., Ku-Mahamud, K. R., & Sagban, R. (2020b). Hybrid ant colony optimization and genetic algorithm for rule induction. Journal of Computer Science, 16(7), 1019–1028.

Al-Behadili, H. N. K., Sagban, R., & Ku-Mahamud, K. R. (2020c). Adaptive parameter control strategy for ant-miner classification algorithm. Indonesian Journal of Electrical Engineering and Informatics, 8(1), 149–162.

Dao, S. D., Abhary, K., & Marian, R. (2015). An adaptive restarting genetic algorithm for global optimization. Lecture Notes in Engineering and Computer Science, 2219, 455–459.

Das, A. K., & Pratihar, D. K. (2018). Performance improvement of a genetic algorithm using a novel restart strategy with elitism principle. International Journal of Hybrid Intelligent Systems, 15(1), 1–15.

Desale, S., Rasool, A., Andhale, S., & Rane, P. (2015). Heuristic and Meta-Heuristic Algorithms and Their Relevance to the Real World: A Survey. International Journal of Computer Engineering in Research Trends, 351(5), 2349–7084.

Duarte, A., Mladenović, N., Sánchez-Oro, J., & Todosijević, R. (2016). Variable Neighborhood Descent. Handbook of Heuristics, 1–27.

El-Ghazali Talb. (2009). METAHEURISTICS from design to implementation.

Ghany, K. K. A., AbdelAziz, A. M., Soliman, T. H. A., & Sewisy, A. A. E. M. (2020). A hybrid modified step Whale Optimization Algorithm with Tabu Search for data clustering. Journal of King Saud University - Computer and Information Sciences, 5(1), 1–8.

Goh, S. L., Kendall, G., Sabar, N. R., & Abdullah, S. (2020). An effective hybrid local search approach for the post enrolment course timetabling problem. Opsearch, 57(4), 1131–1163.

Hayyolalam, V., & Pourhaji Kazem, A. A. (2020). Black Widow Optimization Algorithm: A novel meta-heuristic approach for solving engineering optimization problems. Engineering Applications of Artificial Intelligence, 87(September 2019).

Hore, S., Chatterjee, A., & Dewanji, A. (2018). Improving variable neighborhood search to solve the traveling salesman problem. Applied Soft Computing Journal, 68, 83–91.

Hore, S., Dewanji, A., & Chatterjee, A. (2014). Design issues related to allocation of experimental units with known covariates into two treatment groups. Journal of Statistical Planning and Inference, 155, 117–126.

Hore, S., Dewanji, A., & Chatterjee, A. (2016). On Optimal Allocation of Experimental Units with Known Covariates into Multiple Treatment Groups. Calcutta Statistical Association Bulletin, 68(1–2), 69–81.

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. Indonesian Journal of Electrical Engineering and Computer Science, 13(2), 521–526.

Houssein, E. H., Helmy, B. E. din, Oliva, D., Elngar, A. A., & Shaban, H. (2020). A novel Black Widow Optimization algorithm for multilevel thresholding image segmentation. Expert Systems with Applications, (12), 114-159.

Jabbar, A. M., Ku-Mahamud, K. R., & Sagban, R. (2019a). An improved ACS algorithm for data clustering. Indonesian Journal of Electrical Engineering and Computer Science, 17(3), 1506–1515.

Jabbar, A. M., Ku-Mahamud, K. R., & Sagban, R. (2019b). Balancing exploration and exploitation in acs algorithms for data clustering. 97(16), 4320–4333.

Jabbar, A. M., Ku-Mahamud, K. R., & Sagban, R. (2019c). Modified ACS Centroid Memory for Data Clustering. Journal of Computer Science, 15(10), 1439–1449.

Jia, H. (2015). A Novel Hybrid Optimization Algorithm and its Application in Solving Complex Problem. International Journal of Hybrid Information Technology, 8(2), 1–10.

Moriguchi, K., Ueki, T., & Saito, M. (2015). An Evaluation of the Use of Simulated Annealing to Optimize Thinning Rates for Single Even-Aged Stands. International Journal of Forestry Research, 2015, 1–15.

Odili, J. B., & Mohmad Kahar, M. N. (2016). Solving the Traveling Salesman's Problem Using the African Buffalo Optimization. Computational Intelligence and Neuroscience. 3(1), 1–13.

Rasekh, A., & Osawa, N. (2020). Direct and indirect effect of cannibalism and intraguild predation in the two sibling Harmonia ladybird beetles. Ecology and Evolution, 10(12), 5899–5912.

Sathish, & Ananthapadmanabha. (2021). Improved black widow-bear smell search algorithm for optimal planning and operation of distributed generators in distribution system. Journal of Engineering, Design and Technology. 7(1), 1–21.

Sheriba, S. T., & Rajesh, D. H. (2021). Energy-efficient clustering protocol for WSN based on improved black widow optimization and fuzzy logic. Telecommunication Systems. 15(10), pages213–230.

Sicilia, J. A., Quemada, C., Royo, B., & Escuín, D. (2016). An optimization algorithm for solving the rich vehicle routing problem based on Variable Neighborhood Search and Tabu Search metaheuristics. Journal of Computational and Applied Mathematics, 291, 468–477.

Stegherr, H., Heider, M., & Hähner, J. (2020). Classifying Metaheuristics: Towards a unified multi-level classification system. 7(3), 1–17.

Stützle, T., & Hoos, H. H. (2000). MAX –MIN Ant System. Future Generation Computer Systems, 16, 889–914.

Tan, J. S., Goh, S. L., Sura, S., Kendall, G., & Sabar, N. R. (2020). Hybrid particle swarm optimization with particle elimination for the high school timetabling problem. Evolutionary Intelligence. (3). 1-16

TSPLIB. (1995). Symmetric traveling salesman problem (TSP).

Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. Cybernetics and Information Technologies, 19(1), 3–26.

Yang, X.-S., & Yang, X.-S. (2014). Simulated Annealing. In Nature-Inspired Optimization Algorithms (pp. 67–75).

Zhou, K., Wan, W., Chen, X., Shao, Z., & Biegler, L. T. (2013). A parallel method with hybrid algorithms for mixed integer nonlinear programming. In Computer Aided Chemical Engineering, 19(1). 271-276.

AUTHOR BIOGRAPHIES

Ayad Mohammed Jabbar was born in Iraq, Basra, in 1985. He holds a Bachelor Certificate in Computer Science in 2008, and his Master degree (IT) in 2011 from Universiti Utara Malaysia. His PhD was obtained from Universiti Utara Malaysia in 2021 in Computer Science. His interests include clustering, ant colony optimization and vehicle routing problems.

Ku Ruhana Ku-Mahamud She holds a Bachelor in Mathematical Sciences and a Master's degree in Computing, both from Bradford University, United Kingdom in 1983 and 1986 respectively. Her PhD in Computer Science was obtained from Universiti Pertanian Malaysia in 1994. As an academic, her research interests include ant colony optimization, pattern classification and vehicle routing problems.

Finite mixture of Burr type XII for bus travel time in Klang Valley

Wooi Chen Khoo1*, Cynthia C. T. Cheok1 and Hooi Ling Khoo2

1 Department of Applied Statistics, School of Mathematical Sciences, Sunway University, Sunway City, 47500 Petaling Jaya, Selangor, Malaysia

2 Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, 43000 Kajang, Selangor, Malaysia

* Corresponding author E-mail: wooichenk@sunway.edu.my

(Received 12 July 2021; Final version received 22 July 2021; Accepted 7 September 2021.)

Abstract

Bus travel time analysis is significant to provide useful information to the users for proper journey planning. This study investigated the bus travel time in Klang Valley, Malaysia, which includes the centered city Kuala Lumpur and the state of Selangor. Two bus routes T786 and 851 which cover suburban and urban region have been studied. The data is collected day-to-day. For each link, it is filtered according to morning peak and evening peak, and the travel time is found dual modes, which is well accommodated by the mixture of Burr distribution. The analysis results show that the two-component finite mixture of Burr distribution is viable solution to explain the link travel time. The aim is two-fold. First, the bus travel time data is fitted by the mixture of Burr distribution, subsequently the computation of reliability metrics is carried out. A skew-width method with median based buffer index is considered to measure the reliability. The buffer index of 851 which covers urban region is found relatively high compared to T786. The links which connected by the signalized intersections and junctions tend to show low reliability. For suburban route of T786, the low reliability may be due to the last station. It is found that an additional of 5 minutes waiting time is necessary.

Keywords: Burr, bus operator, mixture distribution, reliability, travel time

1. Introduction

Bus travel time is always a major concern for road users to plan ahead their travel journey. The travel time estimation is usually done intuitively without a proper study or an investigation, most of the users understand the bus arrival time based on experience. Although the travel time is easily known with real-time operating system nowadays, however, the understanding of the travel time pattern from the historical data is significant, as reliable analysis can be a good indicator for road users to gain better understanding on the traffic condition spatially and temporally, especially for the first-time users.

This paper focused on the study for bus travel time in Klang Valley region, which aims to provide a comprehensive analysis on travel time of two bus routes covering Central Business District (CBD) and suburban region. The data is collected day-to-day, from June to December in 2014 and it contains a total of ten bus routes in Klang Valley. For this paper, we examine the travel time of two particular bus routes of T786 and 851 temporally and spatially. T786 mainly covers residential and institution, while 851 passes through mostly governmental offices and commercial district. For each day, the travel time is filtered according to the peak hour; where the morning peak is defined by 7am-10am and the evening peak is defined by 5pm-8pm. For both routes, the travel time pattern of all links is analyzed and it is observed that some links presents dual modes.

Travel time exists in continuous domain. The spread of the data which contribute to pattern construction can be measured by distributional function. The travel time patterns habitually exhibit heavily right-skewed. However, dual peak is either common observed in the pattern because of the daily peak periods. It is of the interest to study the travel time patterns by fitting to some distribution function. It is aimed that more insight is gained upon the fitting, i.e. the skewness can be measured to determine the reliability. An implication is given based upon the analysis results of the reliability. It is found that the reliability of the bus route which covers urban region is relatively high compared to the bus route which covers suburban region. For some links, the waiting time at the particular bus stop could be up to 5 minutes. This study considers fitting the travel time data with a 2-component finite mixture of Burr distribution (MBD), and the favorable results have been obtained. It is believed that none of the relevant study found in using MBD fitting. This study suggests a plausible solution in bus travel time modelling.

Fitting bus travel time data with probability function is feasible for pattern diagnosis. The variability of the travel time gives uncertainty to the road users that they may not know the exact bus arrival time. Therefore, one can understand the travel time better with the distributional function. Probabilistic models provide plausible solution to measure the reliability. This could at least give certain confidence level to the road user to expect the arrival time. Over the past decades, many of the existing study focused on continuous distributions such as lognormal, gamma, Weibull, exponential and loglogistic to describe the travel time distribution ((Polus, 1979), (Xue et al., 2011), (Mazloumi et al., 2010), (Kieu et al., 2015))). Some applied normal distribution based on large sample size, which is the most common distribution has been used in the modelling, simply because of the elegant properties and readily applicable results. Such application may not be appropriate in terms of accuracy, neither suitable for the interpretation, as the data may not be certainly normally distributed. Burr XII distribution is an appropriate model to explain highly right skewed data because of its flexibility. (Susilawati et al., 2013) studied travel time variability on urban roads with Burr XII distribution. The approach is nonetheless appropriate for description because dual peaks are commonly seen in most of the cases of daily travel time. For instance, it is expected that the peaks are usually shown in the morning and in the evening during weekdays. For such bimodality, mixture model often takes good care of such phenomena and its interpretation. Here, we aim to discuss the fitting of day-to-day travel time data with the MBD distribution. The analysis shows promising results to outperform all unimodal distributions, and for some links it is competency with Gaussian Mixture Models (GMM).

It has been common to apply standard deviation to measure the reliability. In transportation modelling and reliability, standard deviation is used to measure the variation from the estimated mean travel time. Larger standard deviation simply means that the variation is higher, and hence the travel time reliability is low. Such approach is simple and direct, however it is not a good option for the skewed data, which is very likely happened in travel time data. The reliability metrics which uses skew and width based on percentile is considered instead to provide better interpretation (Van Lint et al., 2005), as the measure is based on median which is not sensitive to outliers. The percentile with probabilistic models based is calculated to measure the reliability, and it works well also for the data presenting skewness.

The objectives of this paper are to fit the travel time with MBD, and the analysis of the reliability is subsequently carried out. The secondary day-to-day data is provided by Rapid KL (a public transportation own by Prasarana), and the data is filtered for the peak in the morning 7am to 10am and the peak in the evening 5pm to 8pm. The result indicated that dual modes is presented in the travel time for both routes. For reliability analysis, it is expected that the route which covers urban shows relatively low reliability. See the analysis results in later section. The organization of the paper is as follows. Section 2 briefs some existing study in travel time modelling. Most of them emphasizes the fitting of unimodal distributions and GMM, none of them considered mixture of Burr distribution. Data resources and methodology are discussed in Section 3. Two routes are chosen here; one covers downtown KL which passing by the Central Business District (CBD), and another route covers suburban regions dominant by the residential areas and educational institutions. The details of all links and google map are provided along for better visualization. An explanation of the data collection is given. Section 4 reported significant findings. It is observed that the MBD outperformed all unimodal distributions, and it is competitive with GMM. Higher reliability is observed in the suburban bus route and implication is provided based on the analysis. Section 5 concludes.

2. Literature review

In developing country, travel time modelling is particular of the interest in transportation study as it plays an important role for a proper city transportation planning. Some modelling study is found in private transportation travel time modelling in various environments such as arterial roads, expressway and urban road network (Yu et al., 2020; Chalumuri et al., 2014; Shi et al., 2017) were conducted. This paper suggests a viable solution to explain the travel time modelling in public mode bus operators with mixture distribution. For public transportation, the study of travel time modelling especially in bus operation is necessary. Reliable estimation on bus travel time is not only to secure the confidence of bus passengers, it also beneficial to the bus-service company as it serves as an indicator for route condition. With the analysis results, the arrival time can be clearly tabulated at every bus stop station. The passengers can roughly understand the average travel time from station to station. Therefore, travel time efficiency can be improved if a proper travel time estimation and reliability analysis are carried out.

Time series methods have been widely used in estimating the travel time, as the travel time exist in time context, one can find many studies focus on using seasonal and trend patterns for description. Autoregressive integrated moving average (ARIMA) and Seasonal ARIMA are considered to address such problems. Comi et al. (2017) and Comi et al. (2020) applied time series approach in analyzing the travel time in Ukraine and Rome respectively. Despite of the analysis results provide promising prediction which could be offered to the authority for service enhancement, the approach is however a matured approach in past few decades, and the normality assumption is also a hurdle in the modelling as for such assumption is not always fulfilled by the real situation. Other approaches emerged in conjunction with time series, such as Kalman filtering and Machine Learning models of Artificial Neural Network (ANN). Fan and Gurmu (2015) compared the performance among historical average, Kalman filter and ANN. The results show that the ANN outperformed the counterparts. A mathematical-based model is developed by Wong (2009) to estimate regional bus travel time with ANN. Yuan et al. (2020) designed a mechanism of Recurrent Neural Network (RNN) to capture the dynamic temporal behavior and a Deep Neural Network (DNN) is used for travel time prediction. Noor et al. (2020) applied Support Vector Regression (SVR) to analyze the impact of explanatory variables on travel time of Urban City Bus data in Petaling Jaya, a main business district in Selangor. Result shows the weather has the least influence for travel time. Yu et al. (2017) applied survival models and regression analysis to predict the travel time of campus bus service. Such method resulted in good prediction on the travel time associated with uncertainties. Other relevant studies in travel time modelling involved simulation study to examine and to improve the bus service reliability, see Moosavi et al. (2020). Liyanage et al. (2020) suggested on-demand bus service rather than scheduled bus services, and the analysis results show superior benefits of the on-demand bus service. Both piece of works contributed significantly in sustainability practice.

Travel time exists in an uncertainty context. It is much appropriate to capture the travel time pattern by stochastic models. There have been many investigations in travel time modelling and reliability with probability density function in past three decades. The earlier research was carried out by Taylor (1982) who discussed the section travel times with normal distribution. Since then, the application of unimodal distributions such as normal, log-normal, log-logistic and Weibull has been commonly applied for travel time reliability. See Mazloumi et al. (2010), Ma et al. (2016), Shariat et al. (2019) and Büchel et al. (2020). Taylor and Susilawati (2012) shows that the bus travel time reliability is appropriately fitted by Burr distribution, and the travel time on urban roads often presenting dual modes (Susilawati et al., 2013). A recent study by Low et al. (2021) analyzed the bus travel time in Klang Valley region with Burr distribution. The result favored to the Burr distribution makes this study possible, where we may consider to explain the dual modes scenario of the travel time in Klang Valley region with a mixture of Burr distribution. The usage of applying unimodal distribution to capture travel time has been quite established. Highly right-skewed distribution could be easily explained by the heavy right-skewed distribution such as Weibull and its limiting distribution of Burr distribution. However, one of the important characteristics of the travel time which often presenting dual modes is usually neglected in the literature. Although the mixture of Gaussian models has been considered to deal with dual modes, but for both peaks to appear normal is subsequently a further argument. This study aims to fill up the gap of the study. A mixture of Burr distributions is introduced to handle the dual mode in travel time, which we found it appears in the day-to-day data, for both routes 851 and T786.

Mixture distribution is well-known to cater multimodality data in statistical study. In transportation modelling, it emerged as an important analysis tool for travel time modelling. Sun et al. (2018) classified the traffic flow in real-time with Gaussian mixture models (GMM) for better traffic operation and management. (Guo et al., 2019) analyzed the travel time collected from radio frequency identification technique (RFID) in urban road networks with GMM. Comparison has been done with the unimodal distributions and the results show that GMM defeats the counterparts. Similar study has been carried out by Ma et al. (2016) but focus on modelling the travel time variability for bus operations. Yang and Wu (2016) considered mixture models for fitting freeway travel time data. Three mixture models which have been discussed in the paper, i.e. mixture Gaussian, mixture gamma and mixture lognormal. The results show all mixture models are competence. Similar study has been found by Guessous et al. (2014). The mixture of two gamma and two normal distributions were considered to estimate the travel time distribution under different traffic conditions. To the best of our knowledge, the application of mixture distributions in transportation field is limited, especially in bus operating system. We aim to propose a mixture model, with 2-component of finite mixture of Burr distribution to fit the travel time data. The period of the data is 6 months long, and it is collected via Global Positioning System (GPS) within Klang Valley region. The analysis results show that mixture of Burr distribution is competent to the existing GMM, and it could be considered as a viable solution in further analysis.

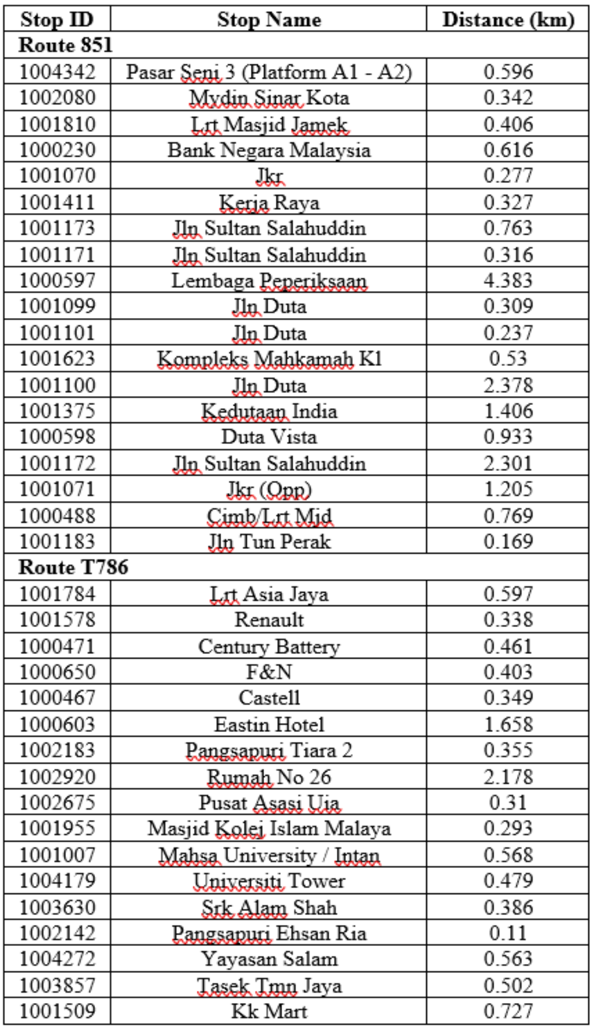
3. Data resources and methodology

3.1 Data resources

The travel time data within Klang Valley is investigated in this study. Klang Valley is located in the centre of Malaysia. It comprises a Federal Territory of Kuala Lumpur and six districts in Selangor, i.e. Petaling, Klang, Gombak, Hulu Langat, Sepang and Kuala Langat. The secondary data is provided by Rapid KL, a public transportation system built by Prasarana Malaysia. We focus on the Rapid KL bus data. Two bus routes are considered in this paper, namely 851 (old route number B115) and T786 (old route number T786). Bus route 851 focuses on the service in Central Business District (CBD), while T786 covers the suburban area. RapidKL buses services covers 6 key areas of the Klang Valley. Route 851 is part of Damansara area coverage while T786 is operated under Lebuhraya Persekutuan area. For both bus routes, the AVL system provides the details such as bus stop ID, street name, route ID, stop ID, stop name and the distance between the bus stops. Some information such as stop ID, stop name and distance are given in Table 1.

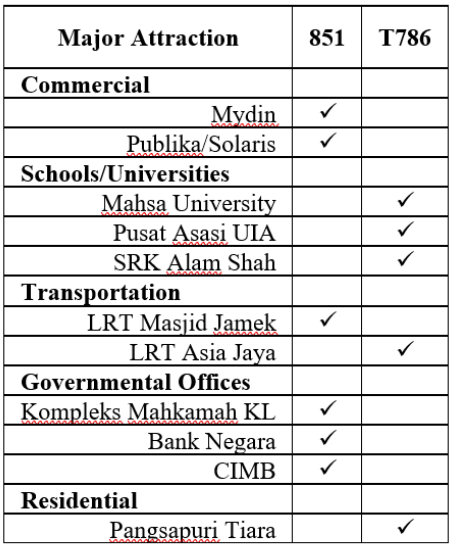
Figure 1 shows the Google map of the bus routes. The total length for bus route 851 is 18.263km and for T786 is 10.277km. Both routes are considered intermediate and short length respectively, and they covered different area in Klang Valley. For bus route 851, it focuses on the service in Central Business District (CBD), while T786 covers the suburban area. RapidKL buses services covers 6 key areas of the Klang Valley. Route 851 is part of Damansara area coverage while T786 is operated under Lebuhraya Persekutuan area. The characteristics of the bus routes are given in Table 2. Both routes cover different area of Klang Valley. For 851, the route focus on Kuala Lumpur (KL) downtown area, which includes the central business district (CBD), such as Masjid Jamek and Pudu. In contrast, T786 operates in Petaling Jaya region, one of the main districts in Selangor. Comparable to 851, it covers suburban, which is dominated by the residential area and the institutions.

Table 1. Route and Stop ID.



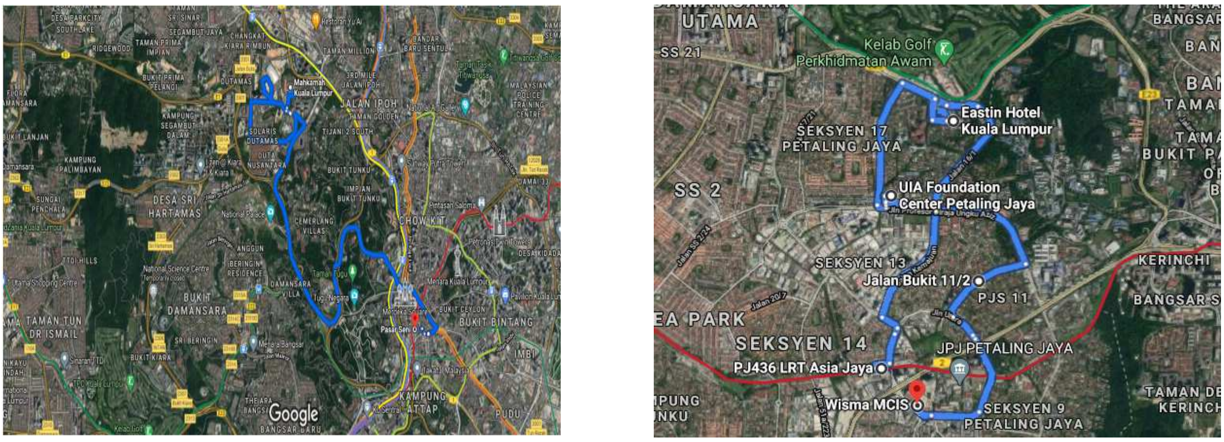
Stop ID	Stop Name	Distance (km)
Route 851		
1004342	Pasar Seni 3 (Platform A1 - A2)	0.596
1002080	Mydin Sinar Kota	0.342
1001810	Lrt Masjid Jamek	0.406
1000230	Bank Negara Malaysia	0.616
1001070	Jkr	0.277
1001411	Kerja Raya	0.327
1001173	Jln Sultan Salahuddin	0.763
1001171	Jln Sultan Salahuddin	0.316
1000597	Lembaga Peperiksaan	4.383
1001099	Jln Duta	0.309
1001101	Jln Duta	0.237
1001623	Kompleks Mahkamah Kl	0.53
1001100	Jln Duta	2.378
1001375	Kedutaan India	1.406
1000598	Duta Vista	0.933
1001172	Jln Sultan Salahuddin	2.301
1001071	Jkr (Opp)	1.205
1000488	Cimb/Lrt Mjd	0.769
1001183	Jln Tun Perak	0.169
Route T786		
1001784	Lrt Asia Jaya	0.597
1001578	Renault	0.338
1000471	Century Battery	0.461
1000650	F&N	0.403
1000467	Castell	0.349
1000603	Eastin Hotel	1.658
1002183	Pangsapuri Tiara 2	0.355
1002920	Rumah No 26	2.178
1002675	Pusat Asasi Uia	0.31
1001955	Masjid Kolej Islam Malaya	0.293
1001007	Mahsa University / Intan	0.568
1004179	Universiti Tower	0.479
1003630	Srk Alam Shah	0.386
1002142	Pangsapuri Ehsan Ria	0.11
1004272	Yayasan Salam	0.563
1003857	Tasek Tmn Jaya	0.502
1001509	Kk Mart	0.727

Table 2. Characteristics of the bus routes.



Major Attraction	851	T786
Commercial		
Mydin	✓	
Publika/Solaris	✓	
Schools/Universities		
Mahsa University		✓
Pusat Asasi UIA		✓
SRK Alam Shah		✓
Transportation		
LRT Masjid Jamek	✓	
LRT Asia Jaya		✓
Governmental Offices		
Kompleks Mahkamah KL	✓	
Bank Negara	✓	
CIMB	✓	
Residential		
Pangsapuri Tiara		✓

The data is secondary, and it is filtered in the consideration of temporal and spatial aspects. It is a 6-month data set, from June to December 2014 and it was day-to-day collected by Global Positioning System (GPS) from 6:00AM to 12:00AM. For this study, the daily data is filtered for the peak in the morning (from 7:00AM to 10:00AM) and in the evening (from 5:00PM to 8:00PM). For both routes, we study the travel time in the temporal and spatial scale. The travel time for each link is in aggregated attributes for every 10 minutes. As an exemplification for aggregated time, the travel time captured by the GPS for busses travelling along the respected link from 7.00am to 7.10am will be accumulated. It is a subsequent accumulation for a duration of 10 minutes. This study combines the aggregated time for both the morning and evening peak hour periods for each link. For instance, given bus route 851, a link simply means that from the Stop ID 1004342 to the Stop ID 1002080, and the link length is 0.596km. We will investigate the travel time for every link of both routes and identify the links that show bimodality during this period. It is observed that dual modes appear in the travel time is observed.



(a)

(b)

Fig. 1. Google map of bus routes (a) 851 and (b) T786.

3.2 Methodology

3.2.1 Mixture of Burr distribution

Mixture distribution has an established history in statistical field. It is also known as compound distribution. It has been defined to cater multimodality data. This study considers a mixture of Burr distribution to explain the travel time data. The proposal of using mixture of Burr distribution to fit the travel time data is the core contribution in this paper. See recent study for mixture of Burr distribution by (Aslam et al., 2018). To the best of our knowledge, fitting the travel time with mixture of Burr distribution is novel in transportation modelling. We aim to fit the dual peaks travel time with the mixture of Burr distribution.

θA m-finite mixture of a distribution is defined by the density with is the parameter

$$f(x) = \sum_{i=1}^{m} w_i p(x|\theta),$$

Where:

$w_i > 0\,; i = 1, \dots, m,\ \sum_{i=1}^{m} w_i = 1.$

This study considers a 2-component mixture of Burr distribution, where m = 2. The pdf is given as follows.

$$f(x) = w_1 \frac{c_1 k_1}{b_1}\left(\frac{x}{b_1}\right)^{c1-1}\left(1+\left(\frac{x}{b_1}\right)^{c_1}\right)^{-k_1-1}$$
$$+w_2 \frac{c_2 k_2}{b_2}\left(\frac{x}{b_2}\right)^{c_2-1}\left(1+\left(\frac{x}{b_2}\right)^{c_2}\right)^{-k_2-1}$$

Where $x > 0$, and c_1, c_2, k_1, k_2 are all positive. The corresponding cumulative distribution function is given by

$$F(x) = w_1\left(1-\left(1+\frac{x^{c_1}}{{b_1}^{c_1}}\right)\right)^{-k_1}$$
$$+w_2\left(1-\left(1+\frac{x^{c_2}}{{b_1}^{c_2}}\right)\right)^{-k_2}$$

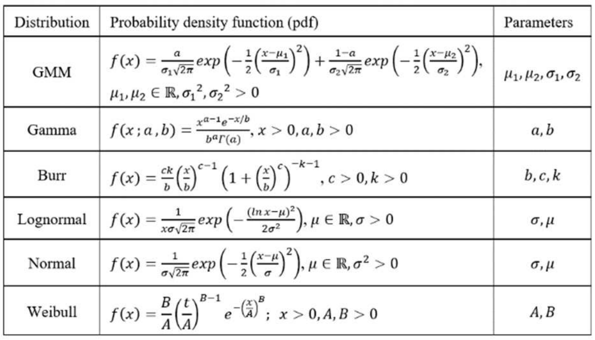
The mixture of Burr distribution is carried out via Jupyter Notebook that supports Python programming language. A few modules like lmfit and numpy packages were used. It is a semi supervised learning approach. Maximum likelihood estimation is used to estimate the parameters. The Burr mixture model is then compared to other distributions such as GMM, including some unimodal distributions, which will be elaborated in next subsection. It is observed that the mixture of Burr caters well for the travel time of some links.

3.2.2 Evaluation approach

(A) Distribution fitting

Most of the existing reviews emphasizes the fitting of unimodal distributions. This is because the travel time usually performs highly skewed distribution. However, dual mode is possible presented in travel time, especially when the daily peak-hour framework is considered, as shown by Figure 2. For comparison purpose, we fitted the data with lognormal, Weibull, gamma, normal and GMM. Table 3 tabulates the probability distribution functions and the parameters for readers' reference.

Table 3. Probability density function and the parameters.



Distribution	Probability density function (pdf)	Parameters
GMM	$f(x) = \frac{a}{\sigma_1\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) + \frac{1-a}{\sigma_2\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right),$ $\mu_1, \mu_2 \in \mathbb{R}, {\sigma_1}^2, {\sigma_2}^2 > 0$	$\mu_1, \mu_2, \sigma_1, \sigma_2$
Gamma	$f(x\,; a\,, b) = \frac{x^{a-1}e^{-x/b}}{b^a\Gamma(a)}, x > 0, a, b > 0$	a, b
Burr	$f(x) = \frac{ck}{b}\left(\frac{x}{b}\right)^{c-1}\left(1+\left(\frac{x}{b}\right)^c\right)^{-k-1}, c > 0, k > 0$	b, c, k
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left(-\frac{(ln\,x-\mu)^2}{2\sigma^2}\right), \mu \in \mathbb{R}, \sigma > 0$	σ, μ
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \mu \in \mathbb{R}, \sigma^2 > 0$	σ, μ
Weibull	$f(x) = \frac{B}{A}\left(\frac{t}{A}\right)^{B-1} e^{-\left(\frac{x}{A}\right)^B};\ x > 0, A, B > 0$	A, B

(B) Performance Evaluation

Model selection is always significant to evaluate the performance. Akaike Information Criteria (AIC) is applied here to justify the model viability. The formula of AIC is $2\boldsymbol{v}$ - 2 *ln L* where $\boldsymbol{v}$ is the number of parameters of the distribution and *ln L* is the estimated log-likelihood function.

3.3 Reliability

Standard deviation is usually carried out in transportation engineering field to measure the reliability. However, it is suitable for the data which is normally presentable. It is suggested to use other measurements such as range and interquartile for the data presents skewness. This paper applied skew-width methods by Van Lint et al. (2005) to measure the reliability. For such metrics it is considering to interpret the skewness by cumulative distribution function. The median-based Buffer Index (BI) represents an additional buffer time required in addition to the median of the travel time. For instance, considering a 95% confidence interval, with a median of 300 seconds and BI value of 0.6, an additional of 180 seconds are needed for travelers to arrive on-time. The BI is defined as:

$$BI_x = \frac{t_{95} - t_{50}}{t_{50}}$$

The skewness of the travel time is a ratio. The numerator subtracts the median from the 90th percentile, and the denominator subtract 10th percentile from the median. This ratio covers the range of 40% observations above the median, and thus it captures the tail and the skewness of the data Taylor et al. (2012). Similarly, the width of the travel time is also defined as the ratio, where the numerator takes the subtraction of the 10th percentile from the 90th percentile, and the denominator simple be the 50th percentile of the travel time. The equation of the skewness and the width travel time metrics are provided as follows.

$$\lambda_{skew} = \frac{t_{90}-t_{50}}{t_{50}-t_{10}}$$

(1)

$$\lambda_{var} = \frac{t_{90}-t_{10}}{t_{50}} \qquad (2)$$

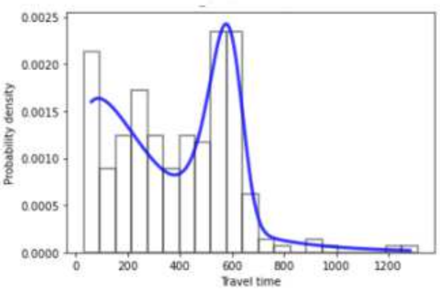
4. Results and discussion

4.1 Distributional fitting

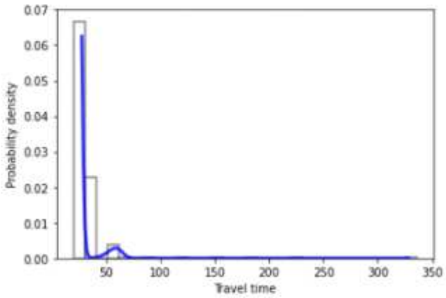
The link travel time is fitted by the (1) mixture of Burr distributions. Some existing distributions have been fitted with the data for comparison purpose. The distributions considered here are the well-know (2) Gaussian Mixture Model (GMM), the unimodal (3) Burr, (4) lognormal, (5) Weibull, (6) gamma and (7) normal distributions. We run the fittings for all links of both routes, but some results have been presented here due to the sake of brevity.

We reported the findings from the data fitting. First, it is found that the pattern of all links is outperformed by the mixture models, compared to the unimodal distributions. In addition to that, it is further justified by AIC that the mixture of Burr distribution obtained slightly lower than the GMM for certain links. The analysis shows that the hypothesis of performing mixture of Burr distributions in explaining the link travel time is plausible, as the lower AIC indicates a better fit for the model. We first discuss the route 851. The results show that for some links where the mixture of Burr distribution performed slightly better than GMM. For instances, the link 1000488 and 1001071 gives lower AIC values, which the AIC indicator suggests that the mixture of Burr is more presentable. The bimodality of the links presented is predicted due to the route characteristics which operates around several tourist attractions spots such as Merdeka Square, Tugu Negara and the Sultan Abdul Samad Building. These national monumental are located along the busy street of this bus link. For some links, the mixture of Burr distribution is observed to be competent with the GMM. For instance, for the link 1004342, the AIC values of MBD and GMM are 32.11 and 31.20 respectively. This is the starting point of the bus route situated at a busy area with LRT facilities and surrounded by the 4 signalized intersection street. The bus operating along this 0.6 kilometre link travels approaching Bus id 1002080. As for link 1002080, it has exhibited an evidence of bimodality because of its shorter length and the influence that the bus stop is located within a commercial area such as the mini market Mydin. Susilawati (2013) and Ma et al. (2016) states that bimodality usually present in shorter links of the routes. Result has proven that the travel time are not fitted by unimodal Burr distribution with finite parameters, and this concluded that the mixture models has captured a better result. It is enlightened to observe that for both links 1001183 and 1000597, GMM fails to carry out the distributional fitting due to the parameter constraint. Link 1000597 is the longest link in route T786, measuring 4.4 kilometres. This part of the route lies at the Board of Examination, Ministry of education and is interconnected with busy roads in Hartamas. Together with that is a mosque and an institute situated nearby. Burr Mixture model ranks as the best model to fit the data of link 1000597. Furthermore, it has also the advantage to cater the travel time for both short and long links. Results has also shown that Burr Mixture model fits nicely to the travel time data for Route T786.

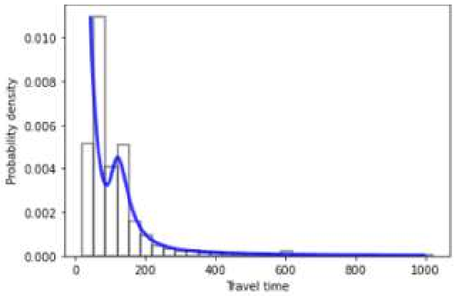
Route T786 is a shorter route compared to 851 which consists of links passing through educational institutes and residential areas, which the coverage is in suburban region. From the analysis it provides strong justification that the mixture of Burr distribution is a plausible model in the travel time analysis. MBD performs well to explain the pattern of both routes spatially (urban and suburban) and temporally (peak hours). See Table 4 for the performance evaluation and parameter estimation of the distributions. Also, the travel time of three links, i.e. 1002080, 1002142 and 1001509 has been presented together with MBD. Figure 2 shows that the MBD fits nicely to the travel time data. Our analysis provides strong justification that the mixture of Burr distribution is a plausible model in the travel time analysis.



(a)



(b)



(c)

Fig. 2. MBD fittings for links (a) 1002080, (b) 1002142, and (c) 1001509.

Table 4. Performance evaluation and parameter estimation.

Bus ID	Distribution	Parameter estimation	Log Likelihood	AIC
Route: 851				
1001071	MBD	$w_1 = 0.11, b_1 = 785.08, c_1 = 10, k_1 = 10$ $w_2 = 0.89, b_2 = 770.52, c_2 = 1.50, k_2 = 10$	-16.99	**33.97**
	GMM	$w_1 = 0.7, \mu_1 = 63.23, \sigma_1 = 14.80$ $w_2 = 0.3, \mu_2 = 129.52, \sigma_2 = 34.57$	-17.46	34.92
	Burr	$b = 74.48, c = 3.04, k = 0.72$	-2401.45	4808.9
	lognormal	$\mu = 4.51, \sigma = 0.69$	-2412.51	4829.02
	Weibull	$A = 130.25, B = 1.28$	-2484.9	4973.8
	gamma	$a = 1.96, b = 60.61$	-2460.59	4925.18
	normal	$\mu = 119.18, \sigma = 115.4$	-2676.23	5356.46
1002080	MBD	$w_1 = 0.65, b_1 = 999.99, c_1 = 1.32, k_1 = 5.35$ $w_2 = 0.35, b_2 = 738.38, c_2 = 9.99, k_2 = 10$	-16.49	32.97
	GMM	$w_1 = 0.005, \mu_1 = 63.58, \sigma_1 = 5.46$ $w_2 = 0.095, \mu_2 = 365, \sigma_2 = 266.18$	-15.18	**30.36**
	Burr	data are not fitted by Burr	-	-
	lognormal	$\mu = 5.61, \sigma = 0.89$	-16067.2	32138.4
	Weibull	$A = 406.39, B = 1.56$	-15769.2	31542.4
	gamma	$a = 1.87, b = 196.70$	-15830.2	31664.4
	normal	$\mu = 366.88, \sigma = 196.70$	15938.4	-31872.8
1000488	MBD	$w_1 = 0.32, b_1 = 35.62, c_1 = 9.99, k_1 = 0.1$ $w_2 = 0.68, b_2 = 60.23, c_2 = 9.99, k_2 = 0.2$	-15.83	**31.66**
	GMM	$w_1 = 0.37, \mu_1 = 68.99, \sigma_1 = 31.04$ $w_2 = 0.63, \mu_2 = 185.64, \sigma_2 = 118$	-17.07	34.14
	Burr	$b = 334, c = 1.62, k = 3.14$	-3697.43	7400.86
	lognormal	$\mu = 4.86, \sigma = 0.82$	-3692.71	7389.42
	Weibull	$A = 192.81, B = 1.35$	-3703.62	7411.24
	gamma	$a = 1.77, b = 99.5$	-3695	7394
	normal	$\mu = 175.81, \sigma = 99.5$	-3854.25	7712.5

1004342	MBD	$w_1 = 0.88, b_1 = 96.40, c_1 = 10, k_1 = 0.05$ $w_2 = 0.12, b_2 = 63.99, c_2 = 10, k_2 = 9.88$	-16.06	**32.11**
	GMM	$w_1 = 0.04, \mu_1 = 50.69, \sigma_1 = 2.37$ $w_2 = 0.96, \mu_2 = 121.37, \sigma_2 = 220.23$	-15.60	31.20
	Burr	$b = 1550.6, c = 1.15, k = 11.49$	-5216.94	10439.88
	lognormal	$\mu = 4.78, \sigma = 1$	-5190.22	10384.44
	Weibull	$A = 195.92, B = 1.09$	-5217.3	10438.6
	gamma	$a = 1.22, b = 154.70$	-5213.36	10430.72
	normal	$\mu = 188.72, \sigma = 175.85$	-5514.12	11032.24
1001183	MBD	$w_1 = 0.52, b_1 = 56.46, c_1 = 4.76, k_1 = 1.10$ $w_2 = 0.48, b_2 = 118.34, c_2 = 5.87, k_2 = 0.39$	-13.40	**26.80**
	GMM	Data are not fitted with GMM	-	-
	Burr	Data are not fitted with Burr	-	-
	lognormal	$\mu = 4.45, \sigma = 0.81$	-1070.33	2144.66
	Weibull	$A = 129.99, B = 1.24$	-1087.45	2178.9
	gamma	$a = 1.62, b = 73.93$	-1082.71	2169.42
	normal	$\mu = 120.41, \sigma = 110.48$	-1156.9	2317.8
1000597	MBD	$w_1 = 0.58, b_1 = 43.32, c_1 = 9.99, k_1 = 0.06$ $w_2 = 0.42, b_2 = 766.52, c_2 = 10, k_2 = 10$	-14.50	**28.99**
	GMM	data are not fitted with GMM	-	-
	Burr	data are not fitted with Burr	-	-
	lognormal	$\mu = 5.44, \sigma = 1.09$	-152.327	308.654
	Weibull	$A = 381.09, B = 1.21$	-150.887	305.774
	gamma	$a = 1.28, b = 279.75$	-151.023	306.046
	normal	$\mu = 358.73, \sigma = 282.857$	-154.905	313.81
Route: T786				
1001007	MBD	$w_1 = 6.45 \times 10^{-9}, b_1 = 388.68, c_1 = 0.19, k_1 = 9.93$ $w_2 = 0.99, b_2 = 71.41, c_2 = 5.99, k_2 = 0.33$	-18.08	**48.16**
	GMM	invalid	-	-
	Burr	$b = 66.23, c = 3.74, k = 0.45$	-16528.20	33062.40
	lognormal	$\mu = 4.64, \sigma = 0.73$	-16650.80	33305.60
	Weibull	$A = 151.33, B = 1.23$	-17140.60	34285.20
	gamma	$a = 1.79, b = 78.10$	-17003.70	34011.40
	normal	$\mu = 139.89\ , \sigma = 142.23$	-18516.60	37037.20
1002142	MBD	$w_1 = 0.95, b_1 = 28.14, c_1 = 10, k_1 = 3.85$ $w_2 = 0.05, b_2 = 74.84, c_2 = 9.99, k_2 = 10$	-18.38	48.75
	GMM	$w_1 = 0.95, \mu_1 = 23.88, \sigma_1 = 14.07$ $w_2 = 0.05, \mu_2 = 53.45, \sigma_2 = 4.42$	-18.58	**45.16**
	Burr	$b = 29.48, c = 116.52, k = 0.10$	-1365.68	2737.36
	lognormal	$\mu = 3.46, \sigma = 0.24$	-2118.24	4240.48
	Weibull	$A = 37.66, B = 1.86$	-2528.44	5060.88
	gamma	$a = 10.01, b = 3.35$	-2281.82	4567.64
	normal	$\mu = 33.58, \sigma = 19.54$	-2673.88	5351.76
1001509	MBD	$w_1 = 0.78, b_1 = 26.91, c_1 = 9.99, k_1 = 0.1$ $w_2 = 0.22, b_2 = 118.67, c_2 = 9.99, k_2 = 0.44$	-19.59	51.18
	GMM	$w_1 = 0.54, \mu_1 = 31.07, \sigma_1 = 10.64$ $w_2 = 0.46, \mu_2 = 120.16, \sigma_2 = 46.73$	-17.37	**42.73**
	Burr	$b = 53.29, c = 3.03, k = 0.60$	-11127.90	22261.80
	lognormal	$\mu = 4.29, \sigma = 0.72$	-11133.70	22271.40
	Weibull	$A = 106.84, B = 1.26$	-11437.90	22879.80
	gamma	$a = 1.87, b = 52.44$	-11339.90	22683.80
	normal	$\mu = 98.09, \sigma = 95.01$	-12339.70	24683.40

4.2 Interpretation and reliability metrics

Equation (1) and (2) are calculated based on percentile values for the travel time distribution. Equation (1) with more than one indicates that users has greater delay with respect to the travel time median. In general, as Equation (1) increases, it is more likely for a user to experience high travel time compared to the median. Equation (2) indicates the spread of travel time distribution. The larger the value, the higher the difference between two extreme values compared to the median. Large values of both parameters give an indication of low reliability, which means the uncertainty of the travel time is increased.

It is reported that the link 1001509 of route T786 and the link 1001071 of route 851 have poor reliability. The buffer time is 4.74 and 3.45 respectively, indicating that more additional travel time is required relatively to median. For instance, the travel time median for link 1001509 is about 1.23 minutes. The buffer time of 4.74 simply means that 4.74 times more than 1.23 minutes, which equivalently an additional of 5.83 minutes is needed for a traveler to travel the link. The analysis can be served as an indicator, which is very useful to the users to understand better the travel time and its reliability for each route link. It is believed that one of the contributing factors for link 1001509 could be the interconnectivity of this link to busy roads in Hartamas. Besides that, link 1001509 is the last station of route 851. The result could be caused by the congestions at the station. As for link 1001071, it is observed that the 1.205 km section is in between a stretch of two junctions. The bottleneck junction along this link which connects a street from where Asia School of Business-Residence could affect the flow of the traffic. Besides that, with the governmental offices (Public Works, Bahagian Perundingan Pengurusan Aset, JKR Malaysia) around might give rise to the low travel time reliability during peak hours. On the contrary, link 1002080 from route 851 and link 1002142 from route T786 had revealed the best travel time reliability among the links. This is in consideration of the fact that link 1002080 operates along a short stretch of a main road, travelling on a straight 0.3 km road with only 1 signalized intersection. To add on, there was no any disruptions flow caused by junctions or bottlenecks. The same condition had been observed for link 1002142. The link travels along a 0.1 km road with no signalized intersection in between. It is evident that these could be the factors that contribute to the high travel time reliability for these two links.

The reliability analysis has been analyzed for all links for both routes. Table 5 tabulates parts of results due to sake of brevity. The finding implies that for the links which connected by signalized intersections and junctions may get lower reliability. Therefore, it is reasonable to give tolerance of the travel time according to the buffer index for each respective links, i.e. an additional of 2 minutes bus arrival time is expected at the link 1004342.

Table 5. Performance evaluation and parameter estimation.

Link	λ_{var}	λ_{skew}	BI
Route 851			
1001071	4.26	2.14	3.45
1002080	0.71	1.40	0.74
1004342	2.56	2.89	2.08
1001183	3.29	2.25	2.69
1000597	1.00	1.77	1.05
Route T786			
1002142	1.00	0.36	0.34
1001509	3.64	2.76	4.74

5. Concluding remark

Since the quick and massive developments over the past few decades, road congestion has been a major issue in Klang Valley, especially the areas that located in the center, such as KL downtown and its connectivity of Petaling Jaya. The issue remains unsolved and the road traffic condition is getting deteriorated with improper traffic design as time goes by. Delay in bus travel time may cause inconvenience to the passengers, for the bus service company, losses is expected if passenger reduction happened due to the uncertainty of the travel time. Proper time management can be well done with a feasible solution on a reliable travel time estimation. This paper provided an insight for the travel time pattern of bus routes T786 and 851, which covers suburban and urban region respectively. The investigation started with the diagnostic of the travel time pattern, by using a mixture of Burr distribution. The mixture of Burr captured the travel time in temporal-spatial aspects. The MBD well performs to estimate the dual modes travel time for the links of both routes, with different route characteristics is taken into the consideration. Results showed that the travel time of some links is well fitted by the distribution, and it is competitive with the existing GMM. Certainly, the unimodal distributions (Burr, Lognormal, Weibull, Gamma and Normal) provide poor fitting when the data presents dual mode. Subsequently, we study the travel time variability with skew-width approach. Results indicated that low reliability is observed in the urban route. This may be affected by the attraction spots, the number of intersections and the length of the links. For bus route in suburban region, it is expected that a delay of 5 minutes is possible in the last station. The findings of this paper serve an important information for the users and the bus service company. Also, the information is a factor in order to carry out a relevant study of benefit-cost analysis in the near future.

6. Acknowledgement

The authors would like to acknowledge RapidKL for the assistance with data collection. This study is supported by Internal Research Grant (IRG) 2021 of Sunway University with the grant code GRTIN-IRG-122-2021, and the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS), Project No.: FRGS/1/2018/TK08/UTAR/02/1.

7. References

Aslam, Muhammad, Tahir, Minhas & Hussain, Zawar. (2018). Reliability analysis of 3-component mixture of distributions. Scientia Iranica. 25. 10.24200/sci.2017.4441.

Büchel, B. & Corman, F. (2020). Review on Statistical Modeling of Travel Time Variability for Road-Based Public Transport. Front. Built Environ. 6:70. doi: 10.3389/fbuil.2020.00070.

Comi, A. Nuzzolo, A. Brinchi, S. & Verghini, R. (2017). Bus travel time variability: some experimental evidences. Transportation Research Procedia 27, pp101-108.

Comi, A., Zhuk, M., Kovalyshyn, V., & Hilevych, V. (2020). Investigating bus travel time and predictive models: a time series-based approach. Transportation Research Procedia, 45, 692–699. doi:10.1016/j.trpro.2020.02.109.

Chalumuri, R.S. & Yasuo, A. (2014). Modelling travel time distribution under various uncertainties on Hanshin expressway of Japan. Eur. Transp. Res. Rev. 6, 85–92. https://doi.org/10.1007/s12544-013-0111-3

Fan, W. & Gurmu, Z (2015). Dynamic Travel Time Prediction Models for Buses Using Only GPS Data. International Journal of Transportation Science and Technology 4(4), pp 353-366, ISSN 2046-0430, https://doi.org/10.1016/S2046-0430(16)30168-x.

Guo, J., Li, C., Qin, X., Huang, W., Wei, Y., & Cao, J. (2019). Analyzing distributions for travel time data collected using radio frequency identification technique in urban road networks. Science China Technological Sciences, 62(1), 106-120. doi: 10.1007/s11431-018-9267-4.

Guessous, Y., Aron, M., Bhouri, N., & Cohen, S. (2014). Estimating Travel Time Distribution under Different Traffic Conditions. Transportation Research Procedia, 3, 339-348. doi:https://doi.org/10.1016/j.trpro.2014.10.014

Kieu, Le Minh, Bhaskar, Ashish, & Chung, Edward (2015) Public transport travel time variability definitions and monitoring. Journal of Transportation Engineering, 141(1), Article number: 04014068 1-9.

Liyanage, S. & Dia, H. (2020). An agent-based simulation approach for evaluating the performance of on-demand bus services. Sustainability 12, 4117. doi:10.3390/su12104117

Low, V. J. M., Khoo, H. L. & Khoo, W. C. (2021).

Statistical modelling of bus travel time with Burr distribution. ITM Web of Conference 36, 01011. IC-MSA. https://doi.org/10.1051/itmconf/20213601011

Van Lint, J. W. C., & Van Zuylen, H. J. (2005). Monitoring and Predicting Freeway Travel Time Reliability: Using Width and Skew of Day-to-Day Travel Time Distribution. Transportation Research Record, 1917(1),54–62. https://doi.org/10.1177/0361198105191700107

Ma, Z., Ferreira, L., Mesbah, M., & Zhu, S. (2016). Modeling distributions of travel time variability for bus operations. Journal of Advanced Transportation, 50(1), 6-24. doi:10.1002/atr.1314

Mazloumi, E., Currie, G., & Rose, G. (2010). Using GPS data to gain insight into public transport travel time variability. Journal of Transportation Engineering, 136(7), 623-631.

Moosavi, S. M. H., Ismail, A. & Yuen, C. W. (2020). Using simulation model as a tool for analysing bus service reliability and implementing improvement strategies. Plos One. https://doi.org/10.1371/journal.pone.0232799

Md Noor, R., Seong Yik, N., Kolandaisamy, R., Ahmedy, I., Hossain, M.A., Alvin Yau, K., Md Shah, W. & Nandy, T. (2020). Predict Arrival Time by Using Machine Learning Algorithm to Promote Utilization of Urban Smart Bus. doi: 10.20944/preprints202002.0197.v1

Polus, A. (1979). A study of travel time and reliability on arterial routes. Transportation, 8(2), 141-151. doi:10.1007/BF00167196

Shariat Mohaymany, A., Ganjkhanloo, A., Bahaabadi, M. R., & Sayyad, A. (2019). Analysis of Travel Time Distribution for Varying Length of Time Interval. AUT Journal of Civil Engineering, 4(2), 10-10. doi:10.22060/ajce.2019.15596.5545

Shi, C. Y., Chen, B. Y. & Li, Q. (2017). Estimation of travel time distribution in urban road networks using low-frequency floating car data. ISPRS International Journal of Geo-Information, 6, 253, doi: 10.3390/ijgi6080253

Susilawati, S., Taylor, M. A. P., & Somenahalli, S. V. C. (2013). Distributions of travel time variability on urban roads. Journal of Advanced Transportation, 47(8), 720-736. doi:10.1002/atr.192

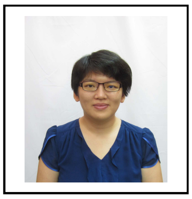
Taylor, M. (1980). Travel time variability – the case of two public modes. Transportation Science 16(4): 507-521. http://dx.doi.org/10.1287/trsc.16.4.507

Taylor, M., & Susilawati, S. (2012). Modelling Travel Time Reliability with the Burr Distribution. Procedia - Social and Behavioral Sciences, 54, 75–83. doi:10.1016/j.sbspro.2012.09.727

Wong, A. C. K. (2009). Travel time prediction model for regional bus transit. Master’s thesis. Available from tspace.library.utoronto.ca

Sun, W., Hu, Z & Hong, L. J. (2018). Gaussian Mixture Model-Based Random a Search for Continuous Optimization via Simulation, Winter Simulation Conference (WSC), Gothenburg, Sweden, 2018, pp. 2003-2014, doi: 10.1109/WSC.2018.8632380

AUTHOR BIOGRAPHIES



Wooi Chen Khoo is a senior lecturer at Sunway University in Malaysia since 2017. Khoo received her Ph.D degree in Statistics in the field of applied probability and application from University of Malaya. She also holds BSc and MSc degrees in Mathematics from University of Science Malaysia. She is currently the University representative of data management. She deals with the Department of Statistics of Malaysia (DOSM). Her areas of interests are in applied probability and the application, parameter estimation, time series, statistical Inference and data analytics.

DOI: 10.6977/IJoSI.202109_6(5).0006

A Text Analytics Approach to Study Python Questions Posted on Stack Overflow

Lee Yong Meng, Soo Yin Yi, Gan Keng Hoon* and Nur-Hana Samsudin
School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, MALAYSIA
* Corresponding author E-mail: khgan@usm.my

(Received July 13, 2021; Final version received 1 September 2021; Accepted 9 September 2021.)

Abstract

Stack Overflow (SO) is one of the largest discussion platforms for programmers to communicate their ideas and thoughts related to various topics like software development and data analysis. Many programmers are actively contributing to this platform and discuss about Python programming language. To better study the topics related to Python questions posted on the platform, a text analytics approach incorporating text preprocessing steps and Latent Dirichlet Allocation (LDA) topic modelling algorithm is proposed. The two main objectives of this study are: to discover and compare the topics of the questions about Python programming language posted on SO from 2008 to 2016, and to analyze questions about Python programming language with high votes posted on SO from 2008 to 2016 using topic modelling technique with a suitable number of topics. From the study, we find that the topics of the Python questions posted on Stack Overflow have gradually shifted towards those related to data modelling and analysis from 2008 to 2016. Furthermore, the study also shows that a suitable number of topics using the topic modelling technique yield a high coherence score concerning the topic model in use, which is important to extract more meaningful topics from the collection of Python questions.

Keywords: Stack Overflow, text processing, text analytics, topic modelling.

1. Introduction

Stack Overflow (SO) is one of the largest open-source software platforms for programmers to ask and discuss programming questions. This platform includes voting, badging and user reputation systems to ensure that the questions and answers posted on the platform are meaningful or relevant to its users. Therefore, SO ecosystem encourages many programmers to not only help each other solve their programming questions voluntarily, but also to showcase their ability in programming problem solving and seeking a better job (Xu et al., 2020). Nevertheless, with its rise in popularity, issues such as duplication of questions (Wang et al., 2020) and the quality of the answers in response to the questions on the platform (Meldrum et al., 2020) greatly affect the browsing experience by programmers when searching for answers through this platform.

Many programming questions have been posted on SO since its official launch in 2008. These include questions related to different programming languages such as C language, Python, Java, and R, to name a few. Specifically, Python and R are the two programming languages most highly associated with the questions related to data analysis posted on the platform. This is reasonable because there are many existing libraries and packages useful for data analysis in both Python and R. This kind of information, which can be extracted using text analytics approaches, can serve for various usages. For example, it can be used by the programming language development team to identify the aspects of the language that are most relevant to these topics so that they can work on improving the language in terms of syntax, features, and even documentations. Besides, it can also be used as a guideline for the programming language course team to identify the important topics to be covered in the content of their courses to meet the requirements of the learners.

Several recent studies have been conducted to analyze the questions and answers (Q&A) about computer programming and software development posted on SO. These include several works performed to extract and classify the topics of discussion on SO related to mobile application development for different platforms, such as Android, iOS, and Window Phones (Ahmad et al., 2019; Beyer et al., 2020; Fontão, et al., 2018). In these studies, the analyses are performed without splitting those discussions according to years to discover the trend or to compare the change of the topics discussed over several years. Furthermore, it is also important to identify the important topics from the questions of a selected programming language on the platform that are most relevant to the users. In this context, a relevant question means any question posted on SO that receives many votes from users who find it helpful for them through the voting system implemented on this platform.

In this study, a text analytics approach involving text preprocessing steps and topic modelling algorithm will be used to analyze the questions related to Python programming language posted on SO. The two main objectives of this study are listed below:

✓ To discover and analyze the topics of the questions about Python programming language posted on SO from 2008 to 2016 to identify and compare the topics being discussed in each year.

✓ To analyze questions about Python programming language with high votes posted on SO from 2008 to 2016 using topic modelling technique with a suitable number of topics.

This paper is structured as follows: Section 2 discusses the related works, including the application of topic modelling algorithms on SO and other forums. Section 3 describes the proposed solution to address the two main objectives of this study, including the dataset description and preprocessing steps, the text preprocessing steps, and the topic modelling approach. Then, Section 4 presents the analysis of findings from the proposed solution. Finally, Section 5 discusses the work done and some limitations of this study and Section 6 summarizes the key points discussed throughout this paper.

2. Related Works

As a large open-source software platform, the Q&A posted on SO contains a fruitful source of information that can be studied and analyzed to understand the topics being discussed by programmers from time to time. Text preprocessing techniques and topic modelling algorithm such as Latent Dirichlet Allocation (LDA) is used in several recent works to study the textual data extracted from the Q&A available on SO for various purposes. For example, LDA algorithm is utilized by researchers to extract the topics of the questions related to various programming languages for further analysis. The study by Ali and Linstead (2020) focuses on several programming languages such as Python, JavaScript, C++, and R to word cloud discover the topics related to these programming languages that have been exhausted for 10 years. Topic exhaustion is a term describing the occurrence where the number of questions related to a topic posted on SO decreases and it takes a longer waiting time to obtain an answer for those questions over the years. The study by Chakraborty et al. (2021) is conducted to identify difficult topics for questions related to new programming languages such as Go, Swift and Rust posted on SO. It is found that topics related to “data” and “data structure” are difficult topics regardless of programming languages. Another study is conducted using the LDA algorithm by Marçal et al. (2020) to identify skill gaps between college and workspace by analyzing topics of the questions related to Computer Science posted on SO.

Topic modelling techniques are also used by researchers to analyze the topics of discussions on different Q&A websites. Stack Exchange, being a network of multiple online Q&A websites in a vast variety of fields (including SO – the main Q&A website under Stack Exchange dedicated for programmers), is often the choice of many researchers to analyze the trends of popular topics being discussed among communities in different fields. For example, the threads from Data Science Stack Exchange (and Reddit) are analyzed using the LDA algorithm by Karbasian and Johri (2020) to identify not only the important Data Science topics and useful examples relevant for teaching Data Science courses, but also various topics related to professional developments. On the other hand, the LDA algorithm is used by Tamla et al. (2019) to identify the thoughts and needs of serious games (SG) developers from their discussions on GameDev Stack Exchange.

On top of that, the LDA algorithm is also commonly used for topic extraction of different fields in several other online forums and social media networks. A combination of Twitter and Reddit datasets are used by Curiskis et al. (2020) to extract the topics being discussed by users in online social networks (OSNs) using the LDA topic modelling algorithm. Besides, the LDA topic modelling algorithm is used to analyze the content specific to eating disorder on Reddit (Moessner et al., 2018) and to extract patient knowledge through their narratives on a patient forum (Dirkson, et al., 2019), respectively. Another application of the LDA topic modelling algorithm is presented by Jaworska and Nanda (2018) to examine the change of topics over time in the large corpus of corporate social responsibility (CSR) reports from the oil sector. The analysis shows that the popular topics of the CSR reports have shifted from the topics related to "climate change" to those related to "human rights".

The works done by researchers using the topic modelling approaches have opened the door for countless possibilities for future studies to analyze texts from different Q&A websites and forums. In this study, the LDA topic modelling algorithm is used to analyze Python questions posted on SO in two ways: first, to identify and compare the topics discussed by programmers on the platform for different years; and second, to extract the topics of the Python questions with high votes on the platform.

3. Proposed Solution

To achieve the two objectives in this study, we propose a text analytics solution that utilizes text processing techniques and the LDA topic modelling algorithm to study and identify the topics of the Python questions posted on SO from 2008 to 2016.

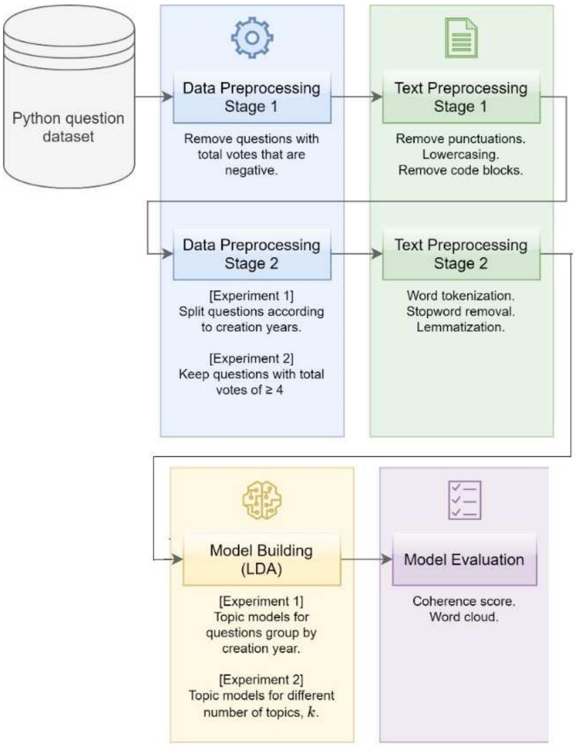

Python question dataset
Data Preprocessing Stage 1
Remove questions with total votes that are negative.
Text Preprocessing Stage 1
Remove punctuations. Lowercasing. Remove code blocks.
Data Preprocessing Stage 2
[Experiment 1] Split questions according to creation years.
[Experiment 2] Keep questions with total votes of ≥ 4
Text Preprocessing Stage 2
Word tokenization. Stopword removal. Lemmatization.
Model Building (LDA)
[Experiment 1] Topic models for questions group by creation year.
[Experiment 2] Topic models for different number of topics, k.
Model Evaluation
Coherence score. Word cloud.

Fig. 1. Overall framework of the proposed solution.

There are two different experiment sets designed to analyze these Python questions. In the first experiment set, the selected Python questions are split into different groups according to the creation year of these questions on the platform. Then, the topic models are built for the respective group of questions to extract the topics of Python questions created in each year. Finally, a comparison is made between topics extracted from different groups of Python questions to investigate any changes in topics of Python questions posted on SO from 2008 to 2016. In the second experiment set, the set of Python questions receiving at least 4 votes from SO users from 2008 to 2016 are studied to identify the relevant topics of Python questions among SO users. In this experiment set, different numbers of topics, k, are tested to identify the optimal k to build a topic model, which covers different topics of the Python questions receiving high votes on the platform.

This study is conducted using the Python programming language. Fig. 1 shows the overall framework of the proposed solution. The description and preprocessing steps of the dataset used, the text preprocessing steps, the topic modelling techniques used for model building, and the model evaluation used in the proposed solution are more thoroughly discussed next in the following subsections.

3.1 Dataset Description and Preprocessing

The dataset used in this study is the Python questions dataset (Overflow, 2019) which can be retrieved from the Kaggle website. This dataset consists of three CSV files, namely: "Answers.csv", "Questions.csv" and "Tags.csv". For this purpose, our work focuses on "Questions.csv", the CSV file that stores 607,282 Python questions posted on SO from August 2008 to October 2016. Table 1. Data description of the Python questions dataset.

summarizes the details and descriptions of each attribute in the Python questions dataset.

Table 1. Data description of the Python questions dataset.

Attribute name	Attribute type	Attribute description
Id	Categorical	The unique identifier of Python questions posted on SO.
OwnerUserId	Categorical	The unique identifier of SO users who post the Python questions.
CreationDate	Date and time	The recorded date and time at which the Python questions are posted SO.
Score	Numerical	Total votes received by the Python questions on SO.
Title	Text	A user-generated title for the Python questions on SO.
Body	Text	Description of the Python questions containing normal text and code blocks, written in HTML script.

Since SO is an English-only Q&A website dedicated to the global community, the texts used in the Python questions dataset are very unlikely to contain non-English words that are not understood by SO users. In addition, SO users can choose to downvote any questions which are not posted in the English language. Therefore, no special care is taken to filter out Python questions not written in the English language from the dataset before the analysis.

In this study, the main source of textual data used for the analysis comes from the "Body" attribute. This is because the "Body" attribute contains more detailed descriptions than the "Title" attribute about the Python questions posted by SO users. There are some questions with poorly defined titles that do not describe the questions well, such as "Introducing Python" and "Most possible pairs". Furthermore, the descriptions of the questions provided are usually complete English sentences which can better represent the frequency of each word being used in the written English language. The data preprocessing steps, which are crucial to ensure the quality and validity of the data used to build the topic models, are conducted at two different stages. The first stage is conducted once on the original Python questions dataset to reduce the amount of data and select relevant features to be used for both experiment sets. On the other hand, the data preprocessing steps during the second stage are conducted specifically for each experiment sets by using different attributes.

a) Data Preprocessing: Stage 1

During this stage, the "Id" and "OwnerUserId" attributes are removed from the dataset because these attributes do not help in our study to analyze Python questions posted on SO. Next, Python questions receiving negative scores (i.e., more downvotes than upvotes) by SO users are also removed from the dataset. Then, to prepare the data for the first experiment set, a new attribute "CreationYear" is derived from the "CreationDate" attribute from the dataset. This attribute stores only the year when the Python questions are posted on SO to enable the grouping of each question into its respective year group. Finally, the selected attributes of the Python questions dataset for further analysis are "Score", "Body" and "CreationYear".

b) Data Preprocessing: Stage 2

For the first experiment set, the "CreationYear" attribute is used to split the Python questions into different groups according to the creation year of each question. On the other hand, for the second experiment set, Python questions with "Score" below 4 are filtered out. After performing this step, the number of Python questions used for further analysis has reduced to only 74,195 questions, which is sufficient for the analysis in this study.

3.2 Text Preprocessing

a) Text Preprocessing: Stage 1

During this stage, several text preprocessing steps are used to prepare the textual data for both experiment sets, including the removal of punctuations, lowercasing the texts, and the removal of code blocks. The first two steps: removing punctuations and lowercasing the texts are crucial to removing the unnecessary parts from the text which can introduce noise to the analysis of the Python questions.

Furthermore, special treatment is required to preprocess the descriptions of the Python questions stored in the "Body" attribute due to several reasons. First, the descriptions of the Python questions are written in HTML script, for instance, each paragraph is enclosed within "<p>" and "</p>" tag pairs. Second, including source code of Python or any programming language in the descriptions of the Python questions (enclosed within "<code>" and "</code>" HTML tag pairs) will certainly affect the result of the analysis in this study. Therefore, steps are taken to first remove the source code from the descriptions. Then, the descriptions of the Python questions are transformed from HTML script to normal English text without HTML tags. This whole process is performed using the BeautifulSoup library from the bs4 module in Python.

b) Text Preprocessing: Stage 2

The text preprocessing steps performed during this stage are a series of steps within a text normalization pipeline. In a text normalization pipeline, all the texts are converted from human-readable texts, including slang words and informal texts, into their corresponding machine-readable forms (Rahate and Chandak, 2019). The steps included in this pipeline are word tokenization, stop words removal and lemmatization. In natural language processing (NLP), tokenization is defined as the task to split a stream of characters into words and punctuations. More specifically, there are two types of tokenization methods used in NLP, which are word tokenization and sentences tokenization. Word tokenization is used to separate words via unique space character whereas sentences tokenization is used to perform tokenization based on sentence boundaries and one of the examples is punctuations. In this study, word tokenization is the method performed to split the words in the description of each Python question into individual units of words or tokens. In Python, word tokenization is performed using "simple_preprocess" utility functions implemented in the Gensim library.

Another text preprocessing step is stop word removal. In NLP, stops words or noise words are the words that contain little information that is not required in the analysis process (Kaur and Buttar, 2018). Therefore, stop words are often removed from the text corpus to improve the efficiency with little influence on the results of NLP tasks. Stop words are usually the most common words in a language. Some stop words in the English language are "I", "am", "is", "are", "this" and "that". In this study, the stop words removed from the list of tokenized words are English stop words listed in the NLTK library in Python.

Finally, lemmatization is the technique used to complete the text normalization pipeline. Lemmatization is performed to convert the tokenized words into their base form or dictionary form. For example, the words "use", "used", "uses" and "using" are all conjugated verbs that are derived from and will be converted to their base form "use" after lemmatization. One benefit of performing lemmatization is that it helps reduce the impact of inflection on English words, such as treating derived words in a text corpus as different words, to the results generated by the NLP models. The lemmatization in Python is performed with the core English language model using the SpaCy library, a popular NLP library along with NLTK.

3.3 Model Building

Topic modelling is one of the applications in text analytics used for studying and identifying the underlying key topics of texts and documents, which are often referred to as the combination of different topics (Curiskis et al., 2020). In simple terms, the goal of a topic modelling task is to extract different "topics" hidden within the given texts and documents through a topic model. A topic model is a generative model driven by the probability framework to help identify such topics. In general, a topic is associated with different words and phrases from the texts and documents that tend to occur together. In other words, similar words or phrases tend to be grouped within the same topic.

One of the popular algorithms used in topic modelling is the Latent Dirichlet Allocation (LDA) model. LDA works by assuming that there is a mixture of different topics within the texts and documents (Alghamdi and Alfalqi, 2015). There is a probability distributed over each topic, measuring the likelihood that a word appears in each topic. LDA algorithm then assigns these words to the topic based on the probability that these words appear in the corresponding topics. In the end, the list of the most probable words in each topic indicates the context of the topics. In this study, the LDA model is built using the Gensim library in Python.

3.4 Model Evaluation

Due to its unsupervised nature, a topic modelling task is often used for exploratory analysis. The dataset is not split into training and test dataset during the topic model building process.

Therefore, the challenge in evaluating the model performance of a topic modelling task is similar to those unsupervised learning tasks, in which, there is no ground truth label used for evaluating the performance of a topic model. This is different from another text analytics task called topic classification, which is supervised learning tasks.

In this study, two approaches are used to evaluate the topics generated for each topic model, namely: the inspection using word cloud and coherence score. Evaluation by inspecting the word cloud of each topic is treated as an informal approach, whereas the coherence score, which relates to the computation of the similarity of words within each topic, is a formal approach to evaluate the performance of a topic model.

a) Word Cloud

Word cloud is a visual representation that captures and displays a list of words from a document, which means a “bag of words”, and their corresponding frequencies. Visually speaking, the more frequent a word appears in the document, the bigger the size of the word in the word cloud. In this context, we treat each topic as a document, in which, the frequency of each word in the word cloud is simply the probability that the word appears in that topic. With this convention, we can visually study the list of most common words that appear in topics generated by the topic models. In Python, word clouds can be generated using the “wordcloud” library.

The advantages of using word clouds to visually represent the topics generated are, it is intuitive and engaging. Humans are visual creatures, in such, there is a region in the human brain specialized for processing visual elements. Therefore, the information delivered through a word cloud can be easily perceived by humans in general. However, the application of word cloud to evaluate the performance of a topic model could be subjective. This means different people might perceive the word cloud differently due to the differences in expertise and cognitive ability among different people. Therefore, word cloud is used as an informal approach to evaluate the performance of a topic model.

b) Coherence Score

Another approach that can be used to evaluate the performance of a topic model is by measuring the coherence score of a topic. Coherence score, which is also referred to as the topic coherence measures, is obtained by computing the similarities of most probable words in each topic semantically (Röder et al., 2015). A high coherence score implies that there is a high degree of similarities among words within the same topic, and thus the topic is said to be more coherent. Therefore, the words are more associated with each other, which implies that the topic is relevant and not merely because the same words appear to be the high scoring words across different topics.

The coherence score for one topic can be calculated using the following formula, Eq. 1:

$$score = \sum_{i<j} sim(v_i, v_j) \qquad (1)$$

where vi and vj are two words in the selected topic such that i and j are both integer values not more than the total number of words in the topic, and sim(vi,vj) is the similarity function used to calculate the word similarity between vi and vj. After obtaining the coherence score for each topic, the coherence score for the topic model is calculated by taking the average value of the coherence scores for all topics generated by the topic model.

A model performance evaluated using numerical measures is often perceived as a more formal approach. Therefore, it is used for our purpose of evaluating the performance of the topic models in our proposed solution. In Python, the coherence score can be calculated by calling the “get_coherence” method of an instance of “CoherenceModel” object in the Gensim library. The “coherence” and “topn” parameters are set to their default values, “c_v” and “20” respectively.

4. Analysis of Findings

a) Experiment 1: Comparing the topics of Python questions in different years

The number of topics, k must be specified before building topic models using LDA algorithms. In this case, we choose $k = 5$ for the first experiment set to identify the 5 topics hidden within the descriptions of Python questions for different years. These coherence scores of the topic models from 2008 to 2016 are plotted in a bar chart in 錯誤! 找不到參照來源。.

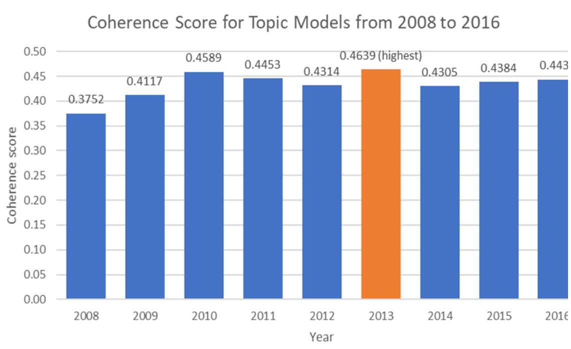


Fig. 2. Bar chart showing the coherence scores for topic models from 2008 to 2016.

From Fig. 2, the coherent score of the topics extracted from the Python questions in 2008 is the lowest among the scores of those in other years. This might be because the data only contains the Python questions posted in SO from August to December 2008. Therefore, the coherence score of the topics in 2018 might be susceptible to undesirable behaviour such as noise and scarcity in the textual data. The word clouds for each topic of Python questions posted on three selected years: 2008 (the first year), 2013 (the year with the highest coherence score) and 2016 (the last year), are shown in Fig. 3, 4 and 5 respectively.

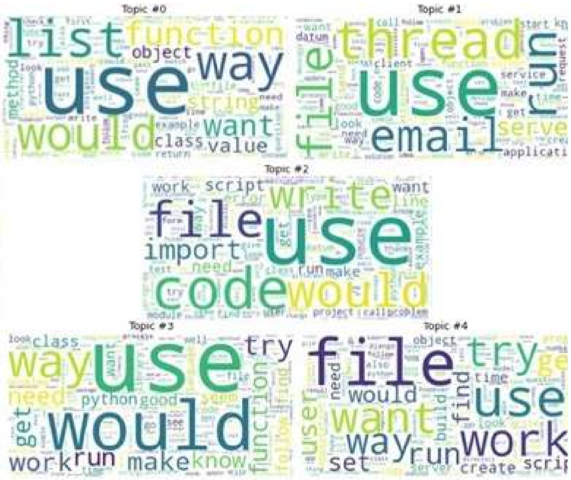


Fig. 3. Word clouds for topics extracted from Python questions in 2008.

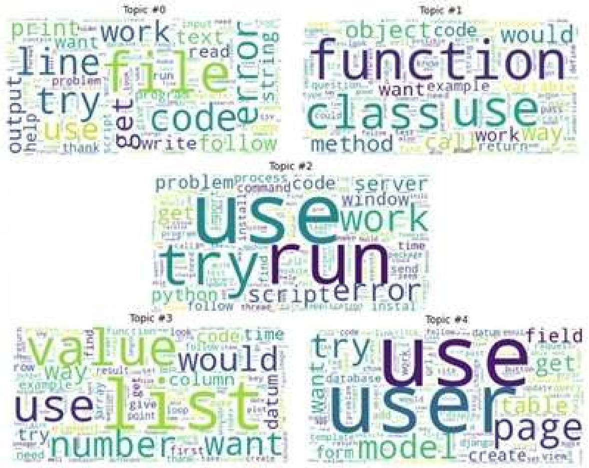


Fig. 4. Word clouds for topics extracted from Python questions in 2013.

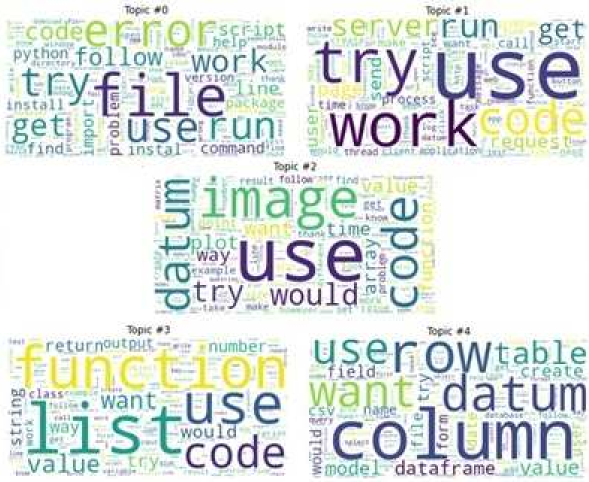


Fig. 5. Word clouds for topics extracted from Python questions in 2016.

By visually inspecting each word cloud generated from the extracted topics within the same year, the same words might appear multiple times in different topics extracted from the topic model. For example, in 2008 (see 錯誤! 找不到參照來源。), the words “file” and “way” appear as words with high probability score in three of the extracted topics. In 2016 (see Fig. 5), two of the words: “try” and “code”, are also common words in at least three extracted topics in that year. This implies that the topics extracted from the Python questions posted on SO within one year are quite similar to each other, even though the list of high scoring words in each topic might differ slightly from one topic to another.

Comparing the topics extracted from the Python questions across different years, there is a gradual shift towards the keywords such as “table”, “datum”, “row” and “column” from 2008 to 2016. These words are not among the high score words in any topics extracted in 2008. This shift might be due to the emergence of Python as an important programming language mainly used by software engineers or data scientists to perform their daily tasks involving transforming and preprocessing data stored in table format. On the other hand, words such as “write” and “import” only appear as high score words in one of the topics in 2008. This could imply that the questions related to importing libraries and writing files have been resolved in earlier days. Users might have already gathered enough information from earlier questions to solve similar problems without having to post new questions to the platform. Therefore, it can be said that there is no one-to-one correspondence between the topics extracted from one year to another because some topics in the past might not stay relevant today and they might eventually be replaced by newer topics in later years.

Finally, the word “use” seems to have a high probability score in all the topics extracted from Python questions for any given year. Therefore, the word “use” might be one of the domain-specific stop words which should be removed from the texts and documents. Table 2. List of high scoring words for each topic extracted from Python questions in 2008, 2013 and 2016.

summarizes the topics extracted from Python questions in 2008, 2013 and 2016, their most probable labels and the corresponding words in each topic. Note that the word “use” is removed from the lists in the table because it appears as one of the six words with the highest probability score in all topics from different years.

b) Experiment 2: Showing the topic for question with high scores

First, a baseline topic model is first built by using the LDA algorithm, with the number of topics, k= 5 specified. The topic model with 5 topics is chosen as the baseline model because we have also set the number of topics, k = 5 for all topic models in the first experiment set. The results from the first experiment set show that topic models with number of topics, k = 5 can generate good results in this study. However, it is also important to test different number of topics which yields the best model performance measured by the coherence score. Therefore, a hyperparameter tuning procedure is performed to search for the optimal value of k from a list of numbers from 2 to 10. The coherence scores for each topic model with the different number of topics are plotted in a bar chart in Fig. 6.

Table 2. List of high scoring words for each topic extracted from Python questions in 2008, 2013 and 2016.

Year	Topic	Label	Words
2008	#0	Python syntax	List, way, would, function, want, string
	#1	Server application	Thread, run, file, email, server, application
	#2	File I/O, import	Code, file, would, write, import, script
	#3	Other	Would, way, try, work, make, need
	#4	File application	File, work, want, try, way. run
2013	#0	File application	File, try, code, error, line, get
	#1	Functional, object-oriented	Function, class, object, would, way, call
	#2	Server application	Run, try, work, error, script, server
	#3	List, Array	List, value, number, would, want, way
	#4	Data model, table form	User, page, try, model, table, get
2016	#0	File application	File, error, try, run, work, get
	#1	Server application	Work, try, code, run, server, get
	#2	Plotting data	Image, code, datum, try, would, plot
	#3	Python Syntax	List, function, code, value, want, try
	#4	Data model, table form	Column, row, datum, want, table, value

Figure 6 shows that $k = 8$ yields the best topic model with the highest coherence score of 0.4482. Therefore, another topic model is built by using the LDA algorithm with the number of topics, $k = 8$ specified. The word clouds for each topic of Python questions with a high number of upvotes for topic models with 5 and 8 topics are shown in Fig. 7 and Fig. 8 respectively.

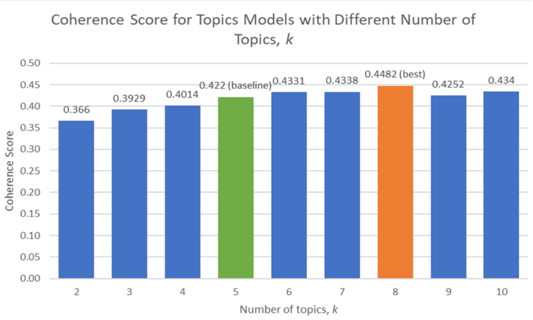


Fig. 6. Bar chart showing the coherence scores for topic models from 2008 to 2016.

By comparing the two sets of word clouds in Fig. 7 and 8 respectively, it can be observed that some of the topics relevant to the Python programmers are not significant in the baseline topic model. For example, topics related to "import" and "column" are not significant in extracted topics of the baseline topic model. These two words do not have their topics. Instead, they can be found in some other topics extracted using the same model. On the other hand, the words "import" and "column" are among the highest-scoring words in Topic #4 in Topic #5 respectively, extracted using the best topic model. The associated words in Topic #4 include: "file", "instal" and "package", indicating that this topic is related to the installation and importing of Python packages. Whereas in Topic #5, the associated high scoring words include: "model", "datum", "row" and "table", which might be a topic related to the data analysis: data model or data stored in table forms.

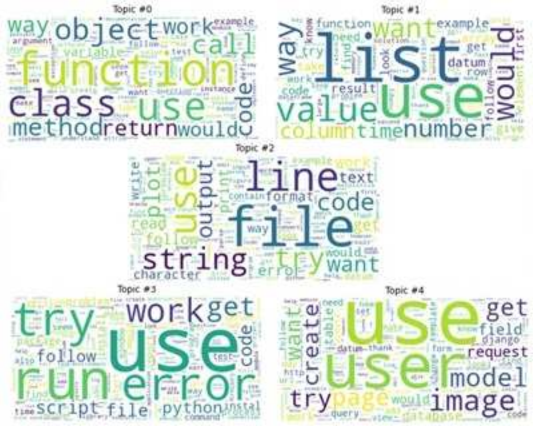


Fig. 7. Word clouds for topics extracted from Python questions with a high score using the best topic model – topic model with 5 topics.

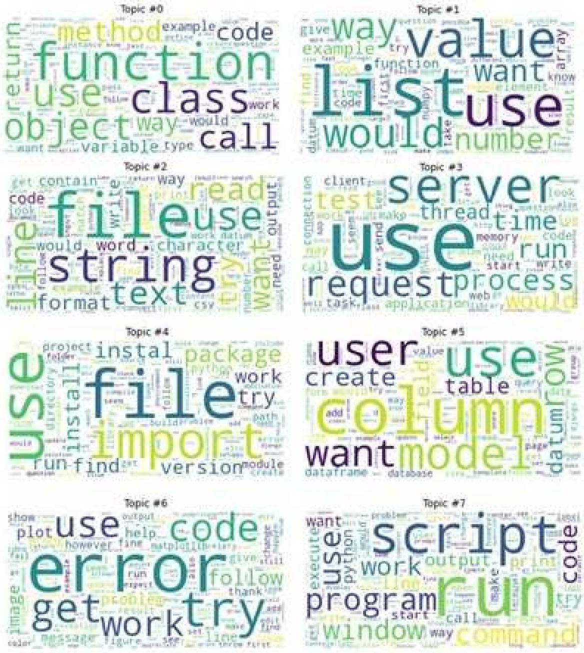


Fig. 8. Word clouds for topics extracted from Python questions with a high score using baseline topic model – topic model with 8 topics.

It is also observed that the high scoring words within the topics extracted using the best topic model (k = 8, which has a higher coherence score) are more associated with each other, as compared to those extracted using the baseline topic model (k = 5, with a lower coherence score). The list of topics extracted from both topic models, the most probable labels and the corresponding words of each topic are summarized in Table 3. Again, the word "use" that appear in all the topics is removed from the list of words in this table.

5. Discussions

Based on the experiment sets, it is shown that topic modelling is useful in extracting the topics from the description of the Python questions posted on SO. The topics are also manually labelled according to the high scoring words within each topic. In this section, we will discuss some improvements which can be performed on the study when extracting the information from the description of the Python questions with our topic modelling approach.

Table 3. List of topics extracted from the baseline and the best topic models, the most probable labels and the corresponding words for each topic.

Model	Topic	Label	Words
Baseline	#0	Functional, object-oriented	function, class, object, call, method, code
	#1	List, array	List, value, would, way, want, number
	#2	Text, string	File, line, string, try, code, plot
	#3	Python script	Run, try, error, work, get, script
	#4	Others	User, image, model, try, get, page
Best	#0	Functional, object-oriented	Function, class, object, call, method, return
	#1	List, array	List, value, would, way, number, want
	#2	File I/O	File, string, line, text, read, want
	#3	Server application	Server, request, process, run, time, test
	#4	Package import and installation	File, import, package, instal, try, version
	#5	Data model, table form	Column, model, user, want, row, create
	#6	Handling error	Error, try, get, code, work, follow
	#7	Python script and command	Run, script, program, command, window, work

First, during the text preprocessing steps, only the most common English stop words are removed. That means, the stop word removal only handles the most frequent words in general English language such as “a”, “the”, “I”, “me”, “by” and “was”. There is no additional step performed to collect the domain-specific stop words before performing stop word removal. Therefore, the experiment results might be influenced by these words to a certain degree. For example, the word “use” that appears as the high scoring words in every topic might be a domain-specific stop word.

Second, the hyperparameter tuning process in the second experiment set only involves changing the number of topics to a limited range of values (from 2 to 10) to obtain the best topic model. With this, there is a high chance that some better topic models (which might yield even higher topic coherence score using the same dataset) are missed out. However, by adding more hyperparameters, the computational resources required to complete the hyperparameter tuning process will increase exponentially.

The future works include efforts to collect domain-specific stop words and exclude them from the analysis of the Python questions posted on SO. Besides, a more thorough hyperparameter tuning process can be performed over a wider range of number of topics (say up to 50 topics), or by including more hyperparameters to the process to search for the model settings that yield the best topic model to the dataset. On top of that, this solution can be adapted to perform text analytics on the questions of other programming languages posted on SO, such as R, C++ and Java.

6. Conclusion

In this study, we apply topic modelling, a text analytics approach to study the Python questions posted on SO from 2008 to 2016. Specifically, we study the description of these questions because the description contains more semantic information than the title of these questions. Due to the unstructured nature of textual data, we perform a series of text preprocessing steps on the descriptions of the Python questions such as removing punctuations and changing the texts into lowercase, transforming the HTML script to normal text, tokenization, stop word removal and lemmatization. Then, two experiment sets are performed on the preprocessed texts. First, the questions are grouped into years and then a topic model is built for each group using the LDA algorithm. The extracted topics are then compared across different years to identify the trend and changing topics of questions over years. Second, the questions with a score of at least 4 are used to build another topic model to identify the topics that cover these questions. In both experiment sets, the evaluation criteria used are the inspection through the word clouds (informal approach) and the computation of the coherence score of each topic model (formal approach).

Through the results obtained from the first experiment set, it is observed that there is a gradual shift to the topics of the Python questions posted on SO from 2008 to 2016. The topic about the data model and table becomes more prominent over the years. For example, there is one topic with keywords such as “table”, “datum”, “row” and “column” generated by the topic model in 2016. These keywords are not significant in any topics generated by the topic models from earlier years. At the same time, the topic with keywords related to file input and output such as “code”, "file", "import" and "script" become less significant over the years.

On the other hand, the results obtained from the second experiment set shows that a suitable number of topics to the topic model built using the LDA algorithm yields extraction of more meaningful topics from Python questions with high votes posted on SO. In this study, the topic model with k = 8 yields the highest coherence score. The topics related to Python package installation and importing, and data models are more prominent in this topic model as compared to the baseline model in this experiment set (k = 5) with a lower coherence score. Therefore, the topic model with the right number of topics that yields a higher coherence score, the topic model is more effective in extracting relevant topics from texts and documents.

Several limitations of this study are also identified and discussed. First, the stop word removal process does not include domain-specific stop words. Second, the hyperparameter tuning process in the second experiment set only involves changing the number of topics to a limited range of values from 2 to 10 due to limited computational resources. Therefore, the future works of this study include collection of domain-specific stop words and exclude these stop words from the analysis and conducting a more thorough hyperparameter tuning process to identify the best topic model to extract information from Python questions posted on SO.

References

Ahmad, A., Feng, C., Li, K., Asim, S. M., & Sun, T. (2019). Toward empirically investigating non-functional requirements of iOS developers on stack overflow. IEEE Access, 7, 61145–61169.

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6.

Ali, R. H., & Linstead, E. (2020). Modeling Topic Exhaustion for Programming Languages on StackOverflow. SEKE, (pp. 400–405).

Beyer, S., Macho, C., Di Penta, M., & Pinzger, M. (2020). What kind of questions do developers ask on Stack Overflow? A comparison of automated approaches to classify posts into question categories. Empirical Software Engineering, 25, 2258–2301.

Chakraborty, P., Shahriyar, R., Iqbal, A., & Uddin, G. (2021). How do developers discuss and support new programming languages in technical Q&A site? An empirical study of Go, Swift, and Rust in Stack Overflow. Information and Software Technology, 137, 106603.

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Information Processing & Management, 57, 102034.

Dirkson, A. R., Verberne, S., Kraaij, W., Jorge, A. M., Campos, R., Jatowt, A., & Bhatia, S. (2019). Narrative detection in online patient communities. Proceedings of Text2Story—Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019), (pp. 21–28).

Fontão, A., Ábia, B., Wiese, I., Estácio, B., Quinta, M., dos Santos, R. P., & Dias-Neto, A. C. (2018). Supporting governance of mobile application developers from mining and analyzing technical questions in stack overflow. Journal of Software Engineering Research and Development, 6, 1–34.

Jaworska, S., & Nanda, A. (2018). Doing well by talking good: A topic modelling-assisted discourse study of corporate social responsibility. Applied Linguistics, 39, 373–399.

Karbasian, H., & Johri, A. (2020). Insights for curriculum development: Identifying emerging data science topics through analysis of Q&A communities. Proceedings of the 51st ACM Technical Symposium on Computer Science Education, (pp. 192–198).

Kaur, J., & Buttar, P. K. (2018). A systematic review on stopword removal algorithms. International Journal on Future Revolution in Computer Science & Communication Engineering, 4, 207–210.

Marçal, I., Garcia, R. E., Eler, D., & Correia, R. C. (2020). A Strategy to Enhance Computer Science Teaching Material Using Topic Modelling: Towards Overcoming The Gap Between College And Workplace Skills. Proceedings of the 51st ACM Technical Symposium on Computer Science Education, (pp. 366–371).

Meldrum, S., Licorish, S. A., Owen, C. A., & Savarimuthu, B. T. (2020). Understanding stack overflow code quality: A recommendation of caution. Science of Computer Programming, 199, 102516.

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. International Journal of Eating Disorders, 51, 656–667.

Overflow, S. (2019, 10). Python Questions from Stack Overflow. Python Questions from Stack Overflow. Retrieved from https://www.kaggle.com/stackoverflow/pythonquestions

Rahate, P. M., & Chandak, M. (2019). Text Normalization and Its Role in Speech Synthesis. International Journal of Engineering and Advanced Technology Special Issue, 8, 115–122. doi:10.35940/ijeat.e1029.0785s319

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining, (pp. 399–408).

Tamla, P., Böhm, T., Nawroth, C., Hemmje, M., & Fuchs, M. (2019). What Do Serious Games Developers Search Online? A Study of GameDev StackExchange. CERC, (pp. 131–142).

Wang, L., Zhang, L., & Jiang, J. (2020). Duplicate question detection with deep learning in stack overflow. IEEE Access, 8, 25964–25975.

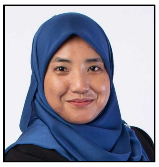
Xu, L., Nian, T., & Cabral, L. (2020). What makes geeks tick? a study of stack overflow careers. Management Science, 66, 587–604.

AUTHOR BIOGRAPHIES

Lee Yong Meng is currently pursuing his Master's degree in Data Science and Analytics from the School of Computer Sciences, Universiti Sains Malaysia (USM). He received his B.App.Sc. degree in mathematical modelling from the School of Mathematical Sciences, USM in 2016. He is also a Graduate Technologist with the Malaysia Board of Technologists (MBOT).

Soo Yin Yi is a postgrad student who is currently pursuing Master Data Science and Analysis at Universiti Sains Malaysia. He is currently working as a Data Analyst in Keysight Tecnologies. He also holds a Bachelor Degree in Logistics from Universiti Utara Malaysia.

Gan Keng Hoon is a senior lecturer in School of Computer Sciences, Universiti Sains Malaysia. She received her Ph. D. degree from Universiti of Malaya (UM) in 2013. She is current the Program Manager of Research Ecosystem and Innovation at the School of Compter Sciences. Her domains of specialization include information retrieval, structured retrieval, structured document representation and query optimization. She has initiated a research platform SIIR (Semantics in Information Retrieval @ ir.cs.usm.my) which is a research initiative related to semantically enhanced information retrieval, and its related applications.



Nur-Hana Samsudin was born in Kuala Lumpur, Malaysia. She received her Ph.D. degree in Computer Science from University of Birmingham, United Kingdom in 2017. Currently she is a Senior Lecturer at Universiti Sains Malaysia in Penang Malaysia. She currently holds one patent and one copyright besides producing research paper since her Master studies. Her interest covers in Natural Language Processing, Speech Processing, sustainable under-resourced language studies and polyglot speech synthesis.

INSTRUCTIONS TO AUTHORS

Submission of Papers

The International Journal of Systematic Innovation is a refereed journal publishing original papers four times a year in all areas of SI. Papers for publication should be submitted online to the IJoSI website (http://www.ijosi.org) In order to preserve the anonymity of authorship, authors shall prepare two files (in MS Word format or PDF) for each submission. The first file is the electronic copy of the paper without author's (authors') name(s) and affiliation(s). The second file contains the author's (authors') name(s), affiliation(s), and email address(es) on a single page. Since the Journal is blind refereed, authors should not include any reference to themselves, their affiliations or their sponsorships in the body of the paper or on figures and computer outputs. Credits and acknowledgement can be given in the final accepted version of the paper.

Editorial Policy

Submission of a paper implies that it has neither been published previously nor submitted for publication elsewhere. After the paper has been accepted, the corresponding author will be responsible for page formatting, page proof and signing off for printing on behalf of other co-authors. The corresponding author will receive one hardcopy issue in which the paper is published free of charge.

Manuscript Preparation

The following points should be observed when preparing a manuscript besides being consistent in style, spelling, and the use of abbreviations. Authors are encouraged to download manuscript template from the IJoSI website, http://www.ijosi.org.

1. *Language.* Paper should be written in English except in some special issues where Chinese maybe acceptable. Each paper should contain an abstract not exceeding 200 words. In addition, three to five keywords should be provided.
2. *Manuscripts.* Paper should be typed, single-column, double-spaced, on standard white paper margins: top = 25mm, bottom = 30mm, side = 20mm. (The format of the final paper prints will have the similar format except that double-column and single space will be used.)
3. *Title and Author.* The title should be concise, informative, and it should appear on top of the first page of the paper in capital letters. Author information should not appear on the title page; it should be provided on a separate information sheet that contains the title, the author's (authors') name(s), affiliation(s), e-mail address(es).
4. *Headings.* Section headings as well as headings for subsections should start front the left-hand margin.
5. *Mathematical Expressions.* All mathematical expressions should be typed using Equation Editor of MS Word. Numbers in parenthesis shall be provided for equations or other mathematical expressions that are referred to in the paper and be aligned to the right margin of the page.
6. *Tables and Figures.* Once a paper is accepted, the corresponding author should promptly supply original copies of all drawings and/or tables. They must be clear for printing. All should come with proper numbering, titles, and descriptive captions. Figure (or table) numbering and its subsequent caption must be below the figure (or table) itself and as typed as the text.
7. *References.* Display only those references cited in the text. References should be listed and sequenced alphabetically by the surname of the first author at the end of the paper. References cited in the text should appear as the corresponding numbers in square bracket with or without the authors' names in front. For example

Altshuller, G.,1998. *40 Principles: TRIZ Keys to Technical Innovation*, Technical Innovation Center.

Sheu, D. D., 2007. Body of Knowledge for Classical TRIZ, *the TRIZ Journal*, 1(4), 27-34.

IJoSI

The International Journal of Systematic Innovation
Journal Order Form

Organization Or Individual Name	
Postal address for delivery	
Person to contact	Name: e-mail: Position: School/Company:
Order Information	**I would like to order ___ copy(ies) of the** *International Journal of Systematic Innovation*: **Period Start: 1st/ 2nd half ____, Year:______(Starting 2010)** **Period End : 1st/ 2nd half ____, Year:____** **Price:** **Institutions: US $150 (yearly) / NT 4,500 (In Taiwan only)** **Individuals: US $50 (yearly) / NT 1500 (In Taiwan only)** (Local postage included. International postage extra) **E-mail to**: IJoSI@systematic-innovation.org or fax: +886-3-572-3210 Air mail desired □ (If checked, we will quote the additional cost for your consent)
Total amount due	**US$**

Payment Methods:

1. **Credit Card (Fill up the following information and e-mail/ facsimile this form to The Journal office indicated below)**
2. **Bank transfer**
   Account: The Society of Systematic Innovation
   Bank Name: Mega International Commercial BANK
   Account No: 020-53-144-930
   SWIFT Code: ICBCTWTP020
   Bank code：017-0206
   Bank Address: No. 1, Xin'an Rd., East Dist., Hsinchu City 300, Taiwan (R.O.C.)

VISA / Master/ JCB/ AMERICAN Cardholder Authorization for Journal Order

Card Holder Information

Card Holder Name	(as it appears on card)		
Full Name (Last, First Middle)			
Expiration Date	/ (month / year)	Card Type	□ VISA □ MASTER □ JCB
Card Number	□□□□-□□□□-□□□□-□□□□	Security Code	□□□
Amount Authorized		Special Messages	
Full Address (Incl. Street, City, State, Country and Postal code)			

Please Sign your name here ________________________ (same as the signature on your card)

The Society of Systematic Innovation
5 F, #350, Sec. 2, Guanfu Rd,
Hsinchu, Taiwan, 30071, R.O.C.