2022 IEEE 16th International Conference on Big Data Science and Engineering (BigDataSE) | 978-1-6654-6523-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIGDATASE56411.2022.00013

A Big Data Science Solution for Transportation Analytics with Meteorological Data

Sukhmandeep Kaur, Nikola N. Kokilev, Michael R. Kuzie, Carson K. Leung ✉,
Ben Nguyen, Adam G.M. Pazdor, Mark J.D. Shinnie
Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
✉ Carson.Leung@UManitoba.ca

***Abstract*—In the current era of big data, very large amounts of data are generating at a rapid rate from a wide variety of rich data sources. Embedded in these big data are valuable information and knowledge that can be discovered by big data science techniques. Transportation data and meteorological data are examples of big data. In this paper, we present a big data science solution for transportation analytics with meteorological data. In particular, we analyze the meteorological data to examine impact of different meteorological conditions (e.g., fog, rain, snow) on the on-time performance of public transit. Evaluation on real-life data collected from the Canadian city of Winnipeg demonstrates the practicality of our big data science solution for transportation analytics on bus delay caused by various meteorological conditions.**

Keywords*—big data, data science, data engineering, data mining, frequent pattern, association rule, transportation analytics, public transit, bus, on-time performance, bus delay, meteorological data, weather condition

I. Introduction

In the current era of big data, very large amounts of precise and uncertain data are generating at a rapid rate from a wide variety of rich data sources in numerous real-life applications. Embedded in these big data are valuable information and knowledge that can be discovered by data science [1-5]. In general, the big data can be characterized by the well-known 5V's:

- volume, which focuses on data size that usually goes beyond the ability of commonly used software tools to capture, manage and process the big data within a tolerable elapsed time;
- veracity, which focuses on data quality (e.g., precise data, imprecise or uncertain data) that may affect the certainty and/or trust for decision making;
- velocity, which focuses on data generation (or collection or flow) rates;
- variety, which focuses on data formats, sources and/or types; and
- value, which focuses on data and/or knowledge usefulness.

Examples of big data include graphs [6-8], social networks [9-11], surveillance, video and image archives, texts and documents, Internet search indices, medical and electronic health records [12-16], business transactions, web logs [17, 18], transportation data [19-23] and meteorological data [24-27]. These big data are usually generated or collected fro-m rich data sources like social media, sensors, and scientific applications.

To discover valuable information knowledge from these big data, data science takes advantages of a combination of techniques like data analysis and visualization [28-31], data management [32-34], data mining [35-42], machine learning [43-46], mathematical and statistical modelling [47]. These techniques have been widely used to handle, manage, mine and visualize the big data collected from a wide variety of real-life applications and services. For example, they have been applied to manage big data (i.e., "new oil"), to retrieve information or knowledge from the well-managed data, and to visualize and validate the retrieved information and knowledge.

In this paper, we focus on transportation data and meteorological data. In particular, we analyze on-time performance of a sustainable transportation mode—namely, public transit bus services. According to 2016 Canadian census [48, 49], the number of car commuters in the city core has decreased (e.g., by 28% in Montreal) over the past two decades in eight largest Canadian census metropolitan areas (CMAs): Toronto, Vancouver, Montreal, Ottawa-Gatineau, Winnipeg, Calgary, Quebec city, and Edmonton. In contrast, the number of public transit commuters has increased (e.g., from 38% in 1996 to 55% in 2016 in Montreal) in the eight CMAs[1,2]. This increase in public transit usage—together with carpooling and active transportation modes (e.g., cycling, walking)—has contributed positively to environmental, social and economic sustainability of smart cities.

On-time performance of public transit bus can be a factor that attracts more commuters to use buses. However, the performance can be affected by meteorological conditions. Thus, in this paper, we examine impacts of meteorological conditions (e.g., fog, rain, snow) on bus on-time performance (e.g., bus delay). Our *key contributions* of this paper include our design and development of a big data science solution for transportation analytics with meteorological data.

In terms of organization for the remainder of this paper, the next section provides background and related works. Section III describes our big data science solution system for transportation analytics. Section IV shows evaluation results on real-life open transportation data and meteorological data for the Canadian city of Winnipeg. Finally, Section V draws the conclusions.

II. Background and Related Works

There have been related works [50-54] conducted at various geographical areas to examine the lateness of public transit. In particular, many of these studies focused on how bus lateness varies at different times of the day and/or days of the week. For instance, to estimate the bus arrival time at different bus stops in northern USA, Patnaik et al. [50] applied multivariate linear regression to automatic passenger counter

[1] https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016029/98-200-x2016029-eng.pdf
[2] https://www150.statcan.gc.ca/pub/75-006-x/2019001/article/00008-eng.pdf

978-1-6654-6523-6/22/$31.00 ©2022 IEEE
DOI 10.1109/BigDataSE56411.2022.00013

(APC) data. To predict bus arrival time in the Chinese city of Shenyang, Yang et al. [51] applied the support vector machine (SVM) with a genetic algorithm (GA) to the data that capture time, weather, and road usage. In contrast, our current paper utilizes frequent pattern mining to generate frequent patterns and interesting association rules to derive new and potentially useful insights.

Moreover, to examine impacts of weather on public transport *ridership*, Zhou et al. [52] applied multivariate modeling In contrast, our current paper focuses on examining the impacts of weather on bus on-time performance (instead of bus ridership).

In addition, there have also been related works that mined frequent patterns from transit data. However, they have not yet examined how weather would affect bus performance. To elaborate, Audu et al. [53] mined frequent patterns from transit data for the Canadian city of Toronto. They used information like bus routes, time of the day, day of the week, location and reported incidents to predict bus lateness. Leung et al. [54] applied frequent pattern mining for examining impacts of COVID-19 on the bus pass-up (i.e., denial of service when the bus is full) for the transit service in the Canadian city of Winnipeg. These two related works did not examine meteorological data. In contrast, our current paper analyzes meteorological data to examine how meteorological conditions affect bus on-time performance. We demonstrates our big data science solution by mining frequent patterns and association rules from Winnipeg transit data.

III. Our Big Data Science Solution for Transportation Analytics

In this section, we describe our big data science solution for transportation analytics. In particular, we integrate and preprocess big data and analyze on-time performance (especially, lateness) of public transit bus and examine impacts of meteorological conditions on the performance.

A. *Ingegrate and Preprocess Big Data*

Our big data science solution first collects and integrates two key types of big data:

- Transportation data—especially, on-time performance data for public transit bus; and
- Meteorological data—especially, weather conditions (e.g., fog, rain, snow).

For *transportation data*, our solution preprocesses the data by selecting relevant features. These include bus stop number/ID, route number, day type (e.g., weekday, Saturday, Sunday, holiday), scheduled arrival time, and actual arrival time. Based on the selected features, our solution then computes the difference (i.e., deviation) between the actual and scheduled arrival times:

$$\text{diff} = \text{actual arr. time} - \text{scheduled arr. time} \qquad (1)$$

A positive difference indicates the bus is late, whereas a negative difference indicates the bus is early. In practice, it can be challenging to get a 0s difference. Hence, taking into account of potential data uncertainty in real world, any difference between -60s and +180s (i.e., 1 minute early to 3 minutes late) is still considered to be *on-time*. In other words:

$$\text{perf.} = \begin{cases} \text{early} & \text{if diff} < -60\text{s} \\ \text{on time} & \text{if} -60\text{s} \le \text{diff} \le +180\text{s} \\ \text{late} & \text{if diff} > +180\text{s} \end{cases} \qquad (2)$$

Moreover, our solution converts continuous features (e.g., stored feature "scheduled arrival time", derived feature "arrival time difference") into discrete categories by binning the data. For instance, scheduled arrival time (of a day) can be discretized into five categories "morning", "midday", "afternoon", "evening" and "night". As for arrival time difference, early bus can be further subdivided and discretized into categories "very early" and "early". Similarly, late bus can be further subdivided and discretized into categories "late" and "very late". See Tables I and II.

TABLE I. Sample Categories for Bus Arrival Difference

On-time bus performance	Range for time difference (sec)
Very early	< -150
Early	[-150, -60)
On-time	[-60, +180]
Late	(+180, 300]
Very late	> 300

TABLE II. Sample Categories for Time of the Day

Time of the day	Time range
Night	00:00-05:59
Morning	06:00-09:59
Midday	10:00-14:59
Afternoon	15:00-17:59
Evening	18:00-23:59

For *meteorological data*, our solution preprocesses the data by selecting relevant features. These include:

- date/time, which can sometimes be recorded in a time zone (e.g., Coordinated Universal Time (UTC)) different from the location/city of interest. In this case, our solution converts the recorded time into the local time in the 24-hour time format for easy computation;
- temperature, which is usually measured in the *metric* units of °C in Canada. Any data measured in other units (e.g., imperial unit of °F) can be converted into its equivalent metric units;
- precipitation, which is usually measured in the *metric* units of mm in Canada. Any data measured in other units (e.g., imperial unit of inches) can be converted into its equivalent metric units. Depending on the temperature, precipitation can be in liquid form (as rain) or frozen solid form (as snow); and
- other related weather data (e.g., categorical description of the weather conditions like "fog", "rain, "snow", " clear sky").

Similar to the preprocessing of transportation data, our solution also converts continuous features into discrete categories by binning the data. Here, it uses the same binning procedure (see Table II) to preprocess date/time in meteorological data in the same way that is preprocesses the transportation data. See Tables III and IV for the discretized categories for features "temperature" and "precipitation".

TABLE III. SAMPLE CATEGORIES FOR TEMPERATURE

Temperature	Temperature range (°C)
Extreme cold	< -20
Freezing	[-20, 0)
Cool	[0, 15]
Warm	(15, 30]
Hot	(30, 35]
Extreme heat	> 35

TABLE IV. SAMPLE CATEGORIES FOR PRECIPITATION

Precipitation	Precipitation range (mm)
No	≤ 0.0001
Negligible	(0.0001, 1]
1-5 mm	(1, 5]
5-10 mm	(5, 10]
10-15 mm	(10, 15]
15-20 mm	(15, 20]
> 20 mm	> 20

B. Analyze Big Data

Once the big data are integrated and preprocessed, our solution analyzes the integrated and preprocessed transportation and meteorological data to examine the impacts of weather conditions on the public transit bus on-time performance. In particular, it conducts frequent pattern mining to find frequently occurring patterns consisting of weather conditions and their corresponding bus performance. Given that we aim to examine the impact of weather conditions on the bus performance, we express our preference that each pattern must include at least one weather condition and must contain a bus performance category. By incorporating this preference, our solution reduces the search space by removing those patterns not satisfying the expressed preference. Moreover, infrequent patterns are removed. Here, a pattern is *frequent* if its frequency ≥ user-specified frequency threshold. As a preview, {fog, morning, on time} is an example of frequent patterns.

After mining frequent patterns, our solution forms association rules from the mined patterns. These rules are of the form "antecedent A ⟶ consequent C". Given that we aim to examine the impact of weather conditions on the bus performance, we express our preference that:

- antecedent of a rule must capture at least one weather condition, and
- consequent of a rule must capture a bus performance category.

By incorporating this preference, our solution reduces the search space by removing those association rules not satisfying the expressed preference. Consequently, the resulting rules reveal associative relationship between the weather conditions and their corresponding bus performance. Here, an association rule is *interesting* if its frequency ≥ user-specified frequency threshold *minfreq* and its confidence ≥ user-specified confidence threshold *minconf*:

$$\mathrm{freq}(\mathrm{A} \rightarrow \mathrm{C}) \geq \mathrm{minfreq} \tag{3}$$

$$\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{C}) = \frac{\mathrm{freq}(\mathrm{A}\rightarrow\mathrm{C})}{\mathrm{freq}(\mathrm{A})} \geq \mathrm{minconf} \tag{4}$$

As a preview, "{fog, morning} ⟶ on time" is an example of interesting association rules.

IV. EVALUATION

To evaluate the effectiveness of our big data science solution for transportation analytics with meteorological data, we applied our solution to real-life data. Specifically, it integrated the following two types of data:

- Winnipeg Transit *on-time performance data* (archive[3] and recent data[4]), which capture public transit bus on-time performance (i.e., early, on-time, late). These data were collected by the Global Positioning System (GPS) on-board computer equipped on all 640 buses running 50,000 km to carry more than 48M passengers over a total of 1.5M hours on 87 bus routes per year[5];
- hourly or daily historical *weather and climate data*[6], which capture features like temperature, total hourly precipitation amount (with a resolution of 0.1 mm) and 41 weather phenomena/events[7] (e.g., rain, snow, fog, sky condition). In this paper, we focused on the historical weather and climate data collected from two weather stations—namely, Winnipeg International Airport (Station ID 5023227) and The Forks in downtown Winnipeg (Station ID 5023262).

Once our solution integrated these data, it preprocessed the integrated data as described in Section III. We first visualized the distribution of data so as to get a better understand of the data. For instance, Fig. 1 shows the correlations between the temperature (ranging from -40°C to +30°C) and bus lateness (ranging -7,500s = 125 minutes early to +10,000s ≈ 167 minutes late).

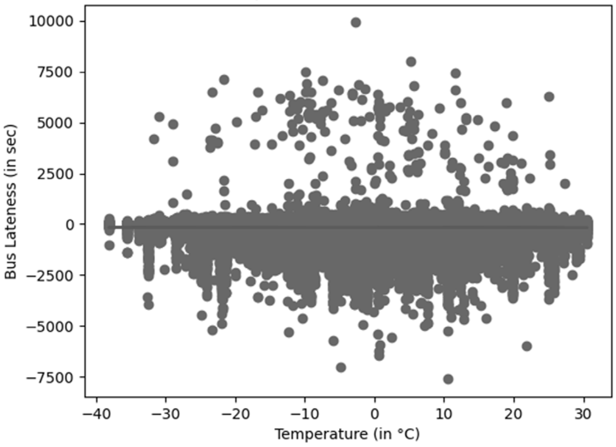

Fig. 1. Temperature vs. bus lateness (i.e., difference between scheduled and actual bus arrival times)

Then, Fig. 2 shows the correlations between the snow (ranging from 0.0mm to 1.2mm at an increment of 0.1mm)

[3] https://data.winnipeg.ca/Transit/Transit-On-Time-Performance-Data-Archive/cymk-nyei

[4] https://data.winnipeg.ca/Transit/Recent-Transit-On-Time-Performance-Data/gp3k-am4u

[5] https://db.winnipegtransit.com/en/about-us/interestingtransitfacts/

[6] https://climate.weather.gc.ca/historical_data/search_historic_data_e.html

[7] https://climate.weather.gc.ca/glossary_e.html#weather

and bus lateness (with the same range of 125 minutes early to 167 minutes late).

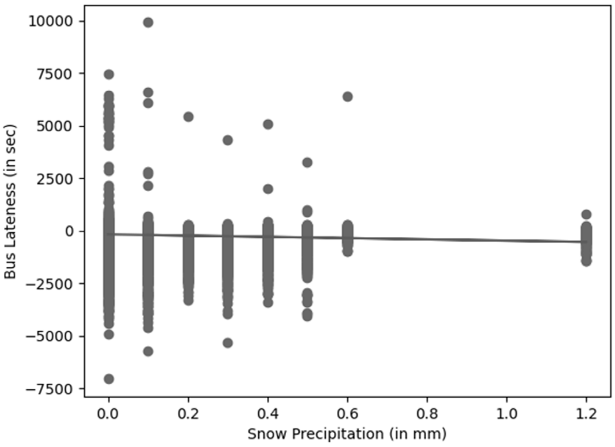

Fig. 2. Solid precipitation (i.e., snow) vs. bus lateness

Similarly, Fig. 3 shows the correlations between the rain (ranging from 0mm to 20mm) and bus lateness (ranging -8,000s ≈ 133 minutes early to +8,000s ≈ 133 minutes late).

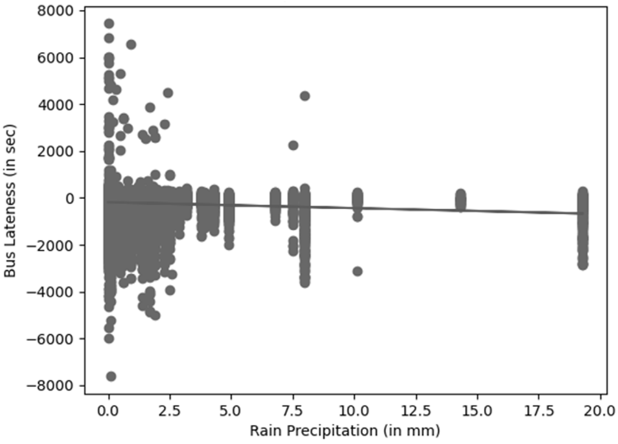

Fig. 3. Liquid precipitation (i.e., rain) vs. bus lateness

A. Impacts of Meteorological Conditions on Bus Lateness

Afterwards, we continued to apply our big data science solution to mine and analyze the preprocessed integrated to discover frequent patterns and interesting association rules for transportation analytics. We first examined the impacts of the four meteorological conditions—namely, clear sky, fog, rain, and snow—on public transit bus on-time performance. With four meteorological conditions (i.e., clear sky, fog, rain, and snow) and five bus on-time performance indicators (i.e., very early, early, on time, late, and very late), our solution returned 20 association rules as shown in Table V. As the number of days with a meteorological condition can be different from the number of days with another meteorological condition (e.g., the number of days with clear sky can be different from the numbers of foggy, rainy and/or snowy days), we normalized frequency of the five rules within each of the four meteorological conditions. As such, the confidence of the rules is identical to the frequency of the rules. Mathematically, for association rules of the form "antecedent → consequent", if their frequencies are normalized within each antecedent group, then frequency freq(antecedent) = 1. As such, the confidence $\mathrm{conf}(\mathrm{A} \rightarrow \mathrm{C}) = \frac{\mathrm{freq}(\mathrm{A}\rightarrow\mathrm{C})}{\mathrm{freq}(\mathrm{A})} = \mathrm{freq}(\mathrm{A} \rightarrow \mathrm{C})$.

TABLE V. ASSOCIATION RULES REVEALING THE IMPACTS OF THE FOUR METEOROLOGICAL CONDITIONS ON BUS ON-TIME PERFORMANCE

Rule	Frequency = confidence
clear sky → on time	0.48250
clear sky → very early	0.26336
clear sky → early	0.18052
clear sky → late	0.06822
clear sky → very late	0.00540
fog → on time	0.44128
fog → very early	0.32342
fog → early	0.17818
fog → late	0.05292
fog → very late	0.00420
rain → on time	0.41346
rain → very early	0.35116
rain → early	0.16868
rain → late	0.06216
rain → very late	0.00454
snow → on time	0.41030
snow → very early	0.35368
snow → early	0.17804
snow → late	0.05406
snow → very late	0.00392

Table V shows most buses arrived *on time* regardless of the meteorological conditions. For example:

- Among the days with clear sky, 48.250% of buses arrived on time. Similarly, among the foggy days, 44.128%, of buses arrived on time. Among rainy or snowy days, 41.346% and 41.030% of buses arrived on time, respectively.
- Within each meteorological condition group, the second and third most frequent bus performance indicators were very early and early.
- Moreover, within each group, < 7% of buses arrived late and < 0.6% of buses arrived very late.

B. Impacts of Time Intervals of a Clear Day on Bus Lateness

Next, we examined whether the impacts of meteorological conditions differ at various time intervals of a day. Specifically, our solution preprocessed the data by binning time into five time intervals of a day—namely, morning, midday, afternoon, evening, and night—according to in Table II. With four meteorological conditions (i.e., clear sky, fog, rain, and snow), five time interval, and five bus on-time performance indicators (i.e., very early, early, on time, late, and very late), our solution discovered 100 association rules. Table VI shows 25 of them, which were observed on days with clear sky (i.e., clear days). Among the five time intervals:

- morning was the busiest time intervals (with the most number of bus activities/performance transactions). It accounted for 35.10% of the transactions.
- The next few busiest time intervals were evening, midday, afternoon, and then night. They respectively accounted for 24.95%, 17.75% 12.12%, and 9.97% of the activities.

Recall from Table V that, the most frequent bus performance indicator on clear days was "on time". Within each time interval, bus performance was similar, though varied slightly. Specifically, the most frequent bus performance indicator was on time in four of the five time intervals. The exception was in the afternoon when the most frequent bus performance indicator was "very early", whereas the "on time" was the second-most frequent bus performance indicator. Hence, it is encouraging to discover that “on time” and “very early” were the two most frequent bus performance indicators on the days when sky was clear. With the “early” being the third-most frequent bus performance indicator, public transit buses were observed to perform very well for being on time or (very) early.

TABLE VI. ASSOCIATION RULES REVEALING THE IMPACTS OF CLEAR SKY AT FIVE TIME INTERVALS OF A DAY ON BUS ON-TIME PERFORMANCE

Rule	Frequency	Confidence
{clear sky, morning} → on time	0.1933	0.5503
{clear sky, morning} → very early	0.0689	0.1963
{clear sky, morning} → early	0.0640	0.1823
{clear sky, morning} → late	0.0228	0.0651
{clear sky, morning} → very late	0.0020	0.0058
	0.3510	
{clear sky, midday} → on time	0.0842	0.4741
{clear sky, midday} → very early	0.0475	0.2675
{clear sky, midday} → early	0.0348	0.1961
{clear sky, midday} → late	0.0105	0.0593
{clear sky, midday} → very late	0.0005	0.0028
	0.1775	
{clear sky, afternoon} → **very early**	0.0544	0.4485
{clear sky, afternoon} → **on time**	0.0409	0.3371
{clear sky, afternoon} → early	0.0183	0.1510
{clear sky, afternoon} → late	0.0072	0.0594
{clear sky, afternoon} → very late	0.0004	0.0037
	0.1212	
{clear sky, evening} → on time	0.1063	0.4260
{clear sky, evening} → very early	0.0836	0.3350
{clear sky, evening} → early	0.0434	0.1739
{clear sky, evening} → late	0.0153	0.0612
{clear sky, evening} → very late	0.0009	0.0036
	0.2495	
{clear sky, night} → on time	0.0576	0.5763
{clear sky, night} → very early	0.0149	0.1494
{clear sky, night} → early	0.0136	0.1366
{clear sky, night} → late	0.0122	0.1228
{clear sky, night} → very late	0.0014	0.0148
	0.0997	

Observed from Table VI, the top-3 frequent rules are “{clear sky, morning} → on time”, “{clear sky, evening} → on time”, and “{clear sky, midday} → on time”. They reveal that, when the sky was clear in the morning, evening or midday, buses arrived on time. When the sky was clear in the afternoon, the top-2 frequent rules are “{clear sky, afternoon} → very early” and “{clear sky, afternoon} → on time”. They reveal that, when the sky was clear in the afternoon, buses were more likely to arrive very early than on time. This suggests riders to come to the bus stops early so as to avoid missing their buses that arrive (and depart) very early.

Also observed from Table VI, when the sky was clear at night, it was 57.63% confident that buses arrived on time. In other words, among all buses operating on nights with clear sky (i.e., clear nights), 57.63% of them arrived on time. Similarly, when the sky was clear in the morning, it was 55.03% confident that buses arrived on time.

C. Impacts of Time Intervals of a Day with Other Meteorological Conditions on Bus Lateness

In Section IV-B, we examined the impacts of time intervals of clear days. In this section, we examine the impacts of time intervals of days with other meteorological conditions (e.g., foggy days, rainy days, snowy days). Recall from Table VI that clear sky occurred more frequently in the morning than evening, midday, afternoon and night. In other words, buses ran more frequently on clear mornings than clear evenings, clear middays, and clear afternoons. Buses ran the least frequently in clear nights. Table VII reveals that, although fog also occurred frequently in the morning, it occurred more frequently in midday and evening than afternoon or night. This makes sense because, intuitively, fog is usually formed when warm moist air passes over a cool surface. There were higher chances of having cool surface in the morning than in the afternoon.

TABLE VII. FREQUENCY RANKS OF PATTERNS DISCOVERED AT FIVE TIME INTERVALS OF A DAY UNDER DIFFERENT METEOROLOGICAL CONDITIONS

Time interval	**Rank (and frequency)**			
	Clear sky	***Fog***	***Rain***	***Snow***
morning	1st 0.3510	1st 0.4969	4th 0.2092	1st 0.3848
midday	3rd 0.1775	2nd 0.2088	2nd 0.2494	2nd 0.2130
afternoon	4th 0.1212	4th 0.1145	1st 0.2583	4th 0.1686
evening	2nd 0.2495	3rd 0.1196	3rd 0.2254	3rd 0.1857
night	5th 0.0997	5th 0.0593	5th 0.0566	5th 0.0468

Frequency rank for rainy and snowy days were also observed to be different in Table VII For example, rain occurred more frequently in the afternoon than midday, evening, morning, and night. In other words, more buses ran on rainy afternoons than rainy middays, rainy evenings, and rainy mornings. Buses ran the least frequently in rainy nights. As another example, snow occurred more frequently in the morning than midday, evening, afternoon, and night. In other words, more buses ran on snowy mornings than snowy middays, snowy evenings, snowy afternoons, and snowy nights. Two other interesting observations from Table VII are:

- The ranked orders for both foggy and snowy conditions were identical: More frequent in the morning than midday, evening, afternoon, and night.
- Although the most frequent time intervals were different among the four meteorological conditions (with most of them—except rain—frequently occurring in the mornings, whereas afternoon was the most frequent time interval on rainy days), they all agree on the least frequent time interval—namely, night. This is probably due to reduced numbers of services at night (i.e., passed midnight).

Among the four meteorological conditions, the top frequent rule for foggy days is “{fog, morning} → on time”, which accounted for 22.68% of all bus services on foggy days. The top-3 frequent rule for rainy days are “{rain, afternoon} → very early”, “{rain, midday} → on time”, and “{rain, morning} → on time”, which accounted for 14.28%, 11.71% and 10.09% of all bus services on rainy days. The top frequent rule for snowy days is “{snow, morning} → on time”, which accounted for 16.94% of all bus services on snowy days. See Table VIII.

Despite the difference in frequency rank orders among the four meteorological conditions, the consequent of rules within each group (i.e., each of the five time interval groups for each meteorological conditions) was mostly consistent: A majority of them led to “on time”, “very early”, “early”, “late” and “very late” as ranked bus performance indicators in these groups. Exceptions are:

- Afternoons, in which “on time” and “very early” were swapped on the ranked bus performance indicator list. In other words, more very early buses (ranked as the most frequent patterns) were observed in the afternoons (regardless whether it was foggy, rainy or snowy afternoons) than on time buses (ranked as the second-most frequent patterns).

- Foggy and rainy nights, in which "very early" and "early" were swapped on the ranked bus performance indicator list. In other words, more early buses (ranked as the second-most frequent patterns) were observed at foggy or rainy nights than very early buses (ranked as the third-most frequent patterns).

TABLE VIII. SAMPLE ASSOCIATION RULES REVEALING THE IMPACTS OF THREE OTHER METEORLOGICAL CONDITIONS AT FIVE TIME INTERVALS OF A DAY ON BUS ON-TIME PERFORMANCE

Rule	Frequency	Confidence
{fog, morning} → on time	0.2268	0.4563
{fog, morning} → very early	0.1521	0.3061
{fog, morning} → early	0.0918	0.1847
{fog, morning} → late	0.0241	0.0484
{fog, morning} → very late	0.0021	0.0043
	0.4969	
{snow, morning} → on time	0.1694	0.4400
{snow, morning} → very early	0.1194	0.3103
{snow, morning} → early	0.0761	0.1977
{snow, morning} → late	0.0183	0.0477
{snow, morning} → very late	0.0016	0.0041
	0.3848	
{rain, afternoon} → **very early**	0.1428	0.5526
{rain, afternoon} → **on time**	0.0679	0.2628
{rain, afternoon} → early	0.0341	0.1319
{rain, afternoon} → late	0.0126	0.0488
{rain, afternoon} → very late	0.0009	0.0037
	0.2583	
{fog, night} → on time	0.0341	0.5750
{fog, night} → **early**	0.0103	0.1732
{fog, night} → **very early**	0.0081	0.1362
{fog, night} → late	0.0062	0.1053
{fog, night} → very late	0.0006	0.0100
	0.0593	

D. *Impacts of Temperature on Bus Lateness Under Different Meteorological Conditions*

In Sections IV-B and IV-C, we examined the impacts of time intervals under four different meteorological conditions (e.g., foggy days, rainy days, snowy days). In this section, we examine the impacts of temperature under these four meteorological conditions. See Table IX, from which we made some interesting observations:

- At the first glance, the frequency ranks for the four meteorological conditions seem to be different. However, their frequency ranks were consistent: Buses were observed to be more frequent in *freezing* temperature than *cool, extreme cold*, and *warm* temperatures. For example, 86.26% of buses operating on snowy days ran under a freezing temperature. Similarly, 58.68% of buses operating on foggy days ran under a freezing temperature, and 30.38% of buses operating on clear days ran under a freezing temperature. As rain turns into snow under freezing temperature, the most frequent temperature under which buses run was cool temperature: 72.50% of buses operating on rainy days ran under a cool temperature.
- As fog is usually formed when warm moist air passes over a cool surface, no or insufficient warm moist air passes under an extreme cold temperature.
- Under freezing or extreme cold temperatures (i.e., below 0°C), rain turns into snow (i.e., from its liquid precipitation form into its solid precipitation form).
- Under warm temperature (i.e., above +15°C), snowmelts.

TABLE IX. FREQUENCY RANKS OF PATTERNS DISCOVERED AT FOUR TEMPERATURE RANGES UNDER DIFFERENT METEOROLOGICAL CONDITIONS

Temperature range	Rank (and frequency)			
	Clear sky	***Fog***	***Rain***	***Snow***
freezing	1st 0.3038	1st 0.5868	N/A	1st 0.8626
cool	2nd 0.2741	2nd 0.3731	1st 0.7250	2nd 0.0948
extreme cold	3rd 0.2208	N/A	N/A	3rd 0.0420
warm	4th 0.1986	3rd 0.0393	2nd 0.2746	N/A

Again, we applied our solution to mine frequent patterns and association rules. With four meteorological conditions (i.e., clear sky, fog, rain, and snow), four temperature ranges, and five bus on-time performance indicators (i.e., very early, early, on time, late, and very late), our solution could discover up to 80 association rules. We then grouped these rules by the antecedent of the rules to form 16 groups. As pointed out earlier, some of these 16 groups did not exist (e.g., no snow at warm temperature). Within each of the existing 12 groups (with an exception of one group), the most popular consequent was "on time". This indicates that buses was on time at different temperature ranges under any of the four meteorological conditions. The exception occurred under an extreme cold temperature on snowy days, in which the top-2 frequent consequents were "very early" and "on time". This means that, under an extreme cold temperature on snowy days, more buses were observed to arrive very early than on time. See Table X.

TABLE X. SAMPLE ASSOCIATION RULES REVEALING THE IMPACTS OF FOUR METEORLOGICAL CONDITIONS AT FOUR TEMPERATURE RANGES ON BUS ON-TIME PERFORMANCE

Rule	Frequency	Confidence
{snow, freezing} → on time	0.3553	0.4118
{snow, freezing} → very early	0.3041	0.3525
{snow, freezing} → early	0.1532	0.1776
{snow, freezing} → late	0.0466	0.0540
{snow, freezing} → very late	0.0034	0.0039
	0.8626	
{rain, cool} → on time	0.2941	0.4055
{rain, cool} → very early	0.2660	0.3668
{rain, cool} → early	0.1205	0.1661
{rain, cool} → late	0.0415	0.0573
{rain, cool} → very late	0.0029	0.0039
	0.7250	
{fog, freezing} → on time	0.2405	0.4098
{fog, freezing} → very early	0.2172	0.3700
{fog, freezing} → early	0.1005	0.1711
{fog, freezing} → late	0.0266	0.0454
{fog, freezing} → very late	0.0020	0.0035
	0.5868	
{clear sky, freezing} → on time	0.1530	0.5036
{clear sky, freezing} → very early	0.0727	0.2392
{clear sky, freezing} → early	0.0536	0.1763
{clear sky, freezing} → late	0.0228	0.0750
{clear sky, freezing} → very late	0.0017	0.0057
	0.3038	
{snow, extreme cold} → **very early**	0.0211	0.5002
{snow, extreme cold} → **on time**	0.0129	0.3063
{snow, extreme cold} → early	0.0060	0.1436
{snow, extreme cold} → late	0.0019	0.0459
{snow, extreme cold} → very late	0.0001	0.0037
	0.0420	

E. *Impacts of Precipitation on Bus Lateness Under Different Meteorological Conditions*

In Section IV-D, we examined the impacts of temperature under these four meteorological conditions. In this section, we examine the impacts of precipitation under these four meteorological conditions. See Table XI, from which we made some interesting observations:

- The frequency ranks for the four meteorological conditions were consistent: Buses were observed to be more frequent under *no* precipitation than *negligible* precipitation, *1-5mm, 5-10mm, 10-15mm,* and *15-20mm* of precipitation. In other words, *the frequency decreases when the amount of precipitation increases*.
- With clear sky, there is no precipitation.
- No significant precipitation was observed on foggy days.
- No heavy precipitation (above 5mm) was observed on snowy days.
- On rainy days, some small amount of precipitation (even negligible precipitation) was observed.

TABLE XI. FREQUENCY RANKS OF PATTERNS DISCOVERED AT SIX PRECIPITATION RANGES UNDER DIFFERENT METEOROLOGICAL CONDITIONS

Precipitation range	**Rank (and frequency)**			
	Clear sky	***Fog***	***Rain***	***Snow***
No	1st 1.0000	1st 0.9506	N/A	1st 0.7918
Negligible	N/A	2nd 0.0493	1st 0.6789	2nd 0.2081
1-5mm	N/A	N/A	2nd 0.2886	3rd 0.2037
5-10mm	N/A	N/A	3rd 0.1515	N/A
10-15mm	N/A	N/A	4th 0.0113	N/A
15-20mm	N/A	N/A	5th 0.0021	N/A

Here, we applied our solution to mine frequent patterns and association rules. With four meteorological conditions (i.e., clear sky, fog, rain, and snow), six precipitation ranges, and five bus on-time performance indicators (i.e., very early, early, on time, late, and very late), our solution could discover up to 120 association rules. We then grouped these rules by the antecedent of the rules to form 24 groups. As pointed out earlier, some of these 24 groups did not exist (e.g., no snow at warm temperature). Within each of the existing 11 groups (with an exception of three groups), the most popular consequent was "on time". This indicates that buses was on time at different precipitation ranges under any of the four meteorological conditions. See Table XII. The exceptions were:

- With negligible or 1-5mm of precipitation on snowy days, the top-2 frequent consequents were "very early" and "on time". This means that, with some but small amount (≤ 5mm of precipitation) on snowy days, more buses were observed to arrive very early than on time.
- With 5-10mm of precipitation on snowy days, the top-2 most frequent consequents were "early" and "late". In particular, 49.72% and 30.86% of buses operating under 5-10mm of precipitation on snowy days were early and late, respectively. Then, the next frequent consequent was "on time", which accounted for 11.82% of buses operating under that antecedent conditions. The two least frequent consequents were "very late" and "very early", which accounted for 7.02% and 3.55% respectively. This ranking was totally different from what we have observed so far.

TABLE XII. SAMPLE ASSOCIATION RULES REVEALING THE IMPACTS OF FOUR METEORLOGICAL CONDITIONS AT SIX PRECIPITATION RANGES ON BUS ON-TIME PERFORMANCE

Rule	**Frequency**	**Confidence**
{clear sky, no precip} → on time	0.4820	0.4820
{clear sky, no precip} → very early	0.2638	0.2638
{clear sky, no precip} → early	0.1806	0.1806
{clear sky, no precip} → late	0.0680	0.0680
{clear sky, no precip} → very late	0.0053	0.0053
	1.0000	
{fog, no precip} → on time	0.4222	0.4441
{fog, no precip} → very early	0.3047	0.3205
{fog, no precip} → early	0.1701	0.1789
{fog, no precip} → late	0.0496	0.0521
{fog, no precip} → very late	0.0039	0.0041
	0.9506	
{snow, no precip} → on time	0.3367	0.4252
{snow, no precip} → very early	0.2625	0.3315
{snow, no precip} → early	0.1436	0.1814
{snow, no precip} → late	0.0456	0.0576
{snow, no precip} → very late	0.0032	0.0041
	0.7918	
{rain, negl precip} → on time	0.2833	0.4172
{rain, negl precip} → very early	0.2365	0.3483
{rain, negl precip} → early	0.1156	0.1703
{rain, negl precip} → late	0.0407	0.0600
{rain, negl precip} → very late	0.0027	0.0040
	0.6789	
{snow, negl precip} → **very early**	0.0018	0.4246
{snow, negl precip} → **on time**	0.0013	0.3150
{snow, negl precip} → early	0.0009	0.2100
{snow, negl precip} → late	0.0002	0.0456
	0.2081	
{rain, 5-10mm} → **early**	0.0102	0.4972
{rain, 5-10mm} → **late**	0.0063	0.3086
{rain, 5-10mm} → **on time**	0.0024	0.1182
{rain, 5-10mm} → **very late**	0.0014	0.0702
{rain, 5-10mm} → **very early**	0.0010	0.0355
	0.0215	

V. CONCLUSIONS

In this paper, we presented a big data science solution for transportation analytics with meteorological data. In particular, it integrates, preprocesses, mines and analyzes the data to examine impacts of temperature and/or precipitation on various time intervals of a day under different meteorological conditions (e.g., clear sky, fog, rain, snow) on the on-time performance of public transit buses. Evaluation on real-life data collected from the Canadian city of Winnipeg demonstrates the practicality of our big data science solution for transportation analytics with meteorological data. *Ongoing and future work* include further exploration on impacts of combinations of meteorological features on bus performance, and on adaptation of our solution to transportation analytics of other transportation modes.

ACKNOWLEDGMENT

This work is partially supported by NSERC (Canada) & University of Manitoba.

REFERENCES

[1] S.H. Ahmed, et al., "Guest editorial introduction to the special issue on data science for intelligent transportation systems. IEEE TITS 23(9), 2022, 16484-16491.

[2] D. Deng, et al., "Spatial-temporal data science of COVID-19 data," IEEE BigDataSE 2021, 7-14.

[3] K.E. Dierckens, et al., "A data science and engineering solution for fast k-means clustering of big data," IEEE TrustCom-BigDataSE-ICESS 2017, 925-932.

[4] C.K. Leung, et al., "Big data science on COVID-19 data," IEEE BigDataSE 2022, 14-21.

[5] U. Qamar, M.S. Raza, Data Science Concepts and Techniques with Applications. Springer, 2020.

[6] M.T. Alam, et al., "Mining high utility subgraphs," IEEE ICDM Workshops 2021, 566-573.

[7] M.E.S. Chowdhury, et al., "A new approach for mining correlated frequent subgraphs," ACM TMIS 13(1), 2022, 9:1-9:28.

[8] R.R. Haque, et al., "UFreS: a new technique for discovering frequent subgraph patterns in uncertain graph databases," IEEE ICBK 2021, 253-260.

[9] D. Choudhery, C.K. Leung, "Social media mining: prediction of box office revenue," IDEAS 2017, 20-29.

[10] C.C.J. Hryhoruk, C.K. Leung, "Compressing and mining social network data," IEEE/ACM ASONAM 2021, 545-552.

[11] C.K. Leung, S.P. Singh, "A mathematical model for friend discovery from dynamic social graphs," IEEE/ACM ASONAM 2021, 569-576.

[12] J. De Guia, et al., "DeepGx: deep learning using gene expression for cancer classification," IEEE/ACM ASONAM 2019, 913-920.

[13] D.L.X. Fung, et al., "Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19," BMC J. Transl. Med. 19, 2021, 318:1-318:18.

[14] C.K. Leung, C. Zhao, "Big data intelligence solution for health analytics of COVID-19 data with spatial hierarchy," IEEE DataCom 2021, 13-20.

[15] Q. Liu, et al., "A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images," IEEE Access 8, 2020, 213718-213728.

[16] J. Zammit, et al., "Semi-supervised COVID-19 CT image segmentation using deep generative models," BMC Bioinformatics 23 (Supplement 7), 2022, 343:1-343:15.

[17] C.C.J. Hryhoruk, C.K. Leung, "Interpretable mining of influential patterns from sparse web," IEEE/WIC/ACM WI-IAT 2021, 532-537.

[18] C.K. Leung, et al., "A web intelligence solution to support recommendations from the web," IEEE/WIC/ACM WI-IAT 2021 Companion, 160-167.

[19] C.C.J. Hryhoruk, et al., "Smart city transportation data analytics with conceptual models and knowledge graphs," IEEE SmartWorld 2021, 455-462.

[20] M.D. Jackson, et al., "A Bayesian framework for supporting predictive analytics over big transportation data," IEEE COMPSAC 2021, 332-337.

[21] J. Kim, et al., "A regression-based data science solution for transportation analytics," IEEE IRI 2022, 55-60.

[22] M. Kolisnyk, et al., "Analysis of multi-dimensional road accident data for disaster management in smart cities," IEEE IRI 2022, 43-48.

[23] C.K. Leung, et al., "Conceptual modeling and smart computing for big transportation data," IEEE BigComp 2021, 260-267.

[24] R.C. Camara, et al., "Fuzzy logic-based data analytics on predicting the effect of hurricanes on the stock market," FUZZ-IEEE 2018, 576-583.

[25] T.S. Cox, et al., "An accurate model for hurricane trajectory prediction," IEEE COMPSAC 2018, vol. 2, 534-539.

[26] B. Nguyen, et al., "A data science solution for mining weather data and transportation data for smart cities," IEEE COMPSAC 2022, 1672-1677.

[27] C. Silva, F. Martins, "Traffic flow prediction using public transport and weather data: a medium sized city case study," WorldCIST 2020, vol. 2, 381-390.

[28] W. Lee, et al. (eds.), Big Data Analyses, Services, and Smart Data, 2021.

[29] C.K. Leung, et al., "Big data analytics of social network data: Who cares most about you on Facebook?" Highlighting the Importance of Big Data Management and Analysis for Various Applications, 2018, 1-15.

[30] C.K. Leung, et al., "Smart data analytics on COVID-19 data," IEEE iThings-GreenCom-CPSCom-SmartData-Cybermatics 2021, 372-379.

[31] K.H. Yoo, et al., "Big data analysis and visualization: challenges and solutions," Applied Sciences 12(16), 2022, 8248:1-8248:5.

[32] C.S. Eom, et al., "Effective privacy preserving data publishing by vectorization," Information Sciences 527, 2020, 311-328.

[33] A.M. Olawoyin, et al., "Privacy-preserving publishing and visualization of spatial-temporal information," IEEE BigData 2021, 5420-5429.

[34] K. Phiwhorm, et al., "Adaptive multiple imputations of missing values using the class center," Journal of Big Data 9, 2022, 52:1-52:25.

[35] M.T. Alam, et al., "Discriminating frequent pattern based supervised graph embedding for classification. PAKDD 2021, Part II, 16-28.

[36] M.T. Alam, et al., "Mining frequent patterns from hypergraph databases," PAKDD 2021, Part II, 3-15.

[37] T.J. Czubryt, et al., "Q-VIPER: quantitative vertical bitwise algorithm to mine frequent patterns," DaWaK 2022, 219-233.

[38] S.Z. Ishita, et al., "New approaches for mining regular high utility sequential patterns," Applied Intelligence 52, 2022, 3781-3806.

[39] E.W. Madill, et al., "Enhanced sliding window-based periodic pattern mining from dynamic streams," DaWaK 2022, 234-240.

[40] M.M. Rahman, et al., "Mining weighted frequent sequences in uncertain databases," Information Sciences 479, 2019, 76-100.

[41] K.K. Roy, et al., "Mining sequential patterns in uncertain databases using hierarchical index structure," PAKDD 2021, Part II, 29-41.

[42] M.E. Shahmi et al., "A new approach for mining correlated frequent subgraphs," ACM TMIS 13(1), 2022, 9:1-9:28.

[43] R. Froese, et al., "The border k-means clustering algorithm for one dimensional data," IEEE BigComp 2022, 35-42.

[44] C.K. Leung, et al., "Machine learning and OLAP on big COVID-19 data," IEEE BigData 2020, 5118-5127.

[45] E. Madill, et al., "ScaleSFL: a sharding solution for blockchain-based federated learning," ACM BSCI 2022, 95-106.

[46] A.M. Olawoyin, et al., "Open data lake to support machine learning on Arctic big data," IEEE BigData 2021, 5215-5224.

[47] C.K. Leung, S.P. Singh, "A mathematical model for friend discovery from dynamic social graphs," IEEE/ACM ASONAM 2021, 569-576.

[48] J. Gilmore, et al., "Commuters using sustainable transportation in census metropolitan areas," Statistics Canada, 2017.

[49] K. Savage, "Results from the 2016 census: commuting within Canada's largest cities," Statistics Canada, 2019.

[50] J. Patnaik, et al., "Estimation of bus arrival times using APC data," Journal of Public Transportation 7(1), 2004, 1–20.

[51] M. Yang, et al., "Bus arrival time prediction using support vector machine with genetic algorithm," Neural Network World 26(3), 2016, 205–217.

[52] M. Zhou, et al., "Impacts of weather on public transport ridership: results from mining data from different sources," Transportation Research Part C 75, 2017, 17-29.

[53] A.A. Audu, et al., "An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city," CISIS 2019, 224-236.

[54] C.K. Leung, et al., "Data mining on open public transit data for transportation analytics during pre-COVID-19 era and COVID-19 era," INCoS 2020, 133-144.