# Developing a Big Data Science Based Model Linked to Meteorological Data for Enhanced Applicability of Transportation Analytics

**Arnav Goenka**

*Vellore Institute of Technology, Vellore, Tamil Nadu, India*

*DOI: 10.37648/ijps.v17i01.020*

[1]*Received: 27 March 2024; Accepted: 23 June 2024; Published: 25 June 2024*

## ABSTRACT

In the current era of big data, vast amounts are generated rapidly from diverse, rich data sources. Embedded within these big data sets is valuable information and knowledge that can be uncovered using big data science techniques. Transportation data and meteorological data are prime examples of such big data. This paper presents a big data science solution for transportation analytics incorporating meteorological data. Specifically, we analyse meteorological data to examine the impact of various weather conditions (e.g., fog, rain, snow) on the on-time performance of public transit. Evaluation using real-life data collected from the Canadian city of Winnipeg demonstrates the practicality and effectiveness of our big data science solution in analysing bus delays caused by different meteorological conditions.

## INTRODUCTION

In the current era of big data, vast amounts of precise as well as uncertain data are generated rapidly from various rich data sources across numerous real-life applications. Embedded within these big data sets is valuable information and knowledge that can be uncovered through data science [1-5]. Generally, big data can be characterized by the well-known 5V's:

- Volume: This pertains to the data size, which often exceeds the capacity of commonly used software tools to capture, manage, and process within a reasonable timeframe.

- Veracity: Data quality is a critical aspect of big data, encompassing both precise and uncertain data. The reliability and trustworthiness of decision-making processes are significantly influenced by the veracity of the data being used.

- Velocity: This refers to the rate at which data is generated, collected, or flows.

- Variety: This highlights the diversity in data formats, sources, and types.

- Value: The true power of big data lies in its value. The usefulness of the data and the knowledge derived from it are key factors in its relevance and application across various fields.

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

Data science is key for finding useful information from these huge sets of data. It uses various methods including data analysis and visualization, data management, data mining, machine learning, as well as mathematical and statistical modelling; these are not just techniques but the foundations of our study allowing us to deal with large amounts of information collected from different applications and services in real life. They help manage big data; allow getting

[1] *How to cite the article:* Goenka A (June 2024); Developing a Big Data Science Based Model Linked to Meteorological Data for Enhanced Applicability of Transportation Analytics; *International Journal of Professional Studies*; Jan-Jun 2024, Vol 17, 256-266; DOI: http://doi.org/10.37648/ijps.v17i01.020

information or knowledge from well-managed data; and enable visualizing as well validating retrieved information or knowledge which brings understanding into a complex world of transport analytics.

This paper focuses on transportation and meteorological data, explicitly analyzing the on-time performance of a sustainable transportation mode—public transit bus services. According to the 2016 Canadian census [28, 29], the number of car commuters in city cores has decreased (e.g., by 28% in Montreal) over the past two decades in the eight largest Canadian census metropolitan areas (CMAs): Toronto, Vancouver, Montreal, Ottawa-Gatineau, Winnipeg, Calgary, Quebec City, and Edmonton. Conversely, the number of public transit commuters has increased (e.g., from 38% in 1996 to 55% in 2016 in Montreal) in these CMAs. This rise in public transit usage, carpooling, and active transportation modes (e.g., cycling and walking) has contributed positively to intelligent cities' environmental, social, and economic sustainability.

The on-time performance of public transit buses is a crucial factor in attracting more commuters to use buses. However, this performance can be affected by meteorological conditions. Therefore, this paper examines the impacts of various meteorological conditions (e.g., fog, rain, snow) on bus on-time performance, explicitly focusing on bus delays. The key contributions of this paper include the design and development of a big data science solution for transportation analytics that integrates meteorological data.

## OUR BIG DATA SCIENCE SOLUTION FOR TRANSPORTATION ANALYTICS

This section details our big data science solution for transportation analytics. Specifically, we focus on integrating and preprocessing big data to analyze the on-time performance (particularly lateness) of public transit buses and examine the impacts of meteorological conditions on this performance.

### Integrate and Preprocess Big Data

Our big data science solution begins by collecting and integrating two primary types of big data:

- Transportation data: On-time performance data for public transit buses.

- Meteorological data: Weather conditions, including fog, rain, and snow.

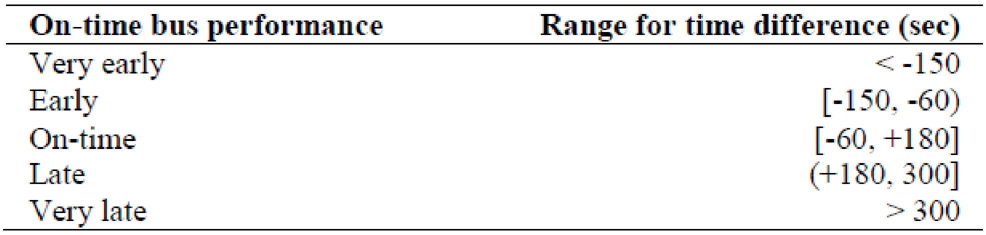
Our solution preprocesses transportation data by selecting relevant features such as bus stop number/ID, route number, day type (e.g., weekday, Saturday, Sunday, holiday), scheduled arrival time, and actual arrival time. Then we apply feature extraction to get new feature namely- deviation calculated as the difference between the actual and scheduled arrival times based on these selected features.

diff = actual arr. time - scheduled arr. Time

A positive difference indicates the bus is late, while a negative difference indicates the bus is early. In practice, achieving a 0-second difference is challenging. Therefore, to account for potential data uncertainty in the real world, any difference between -60 seconds and +180 seconds (i.e., 1 minute early to 3 minutes late) is still considered on-time. In other words:

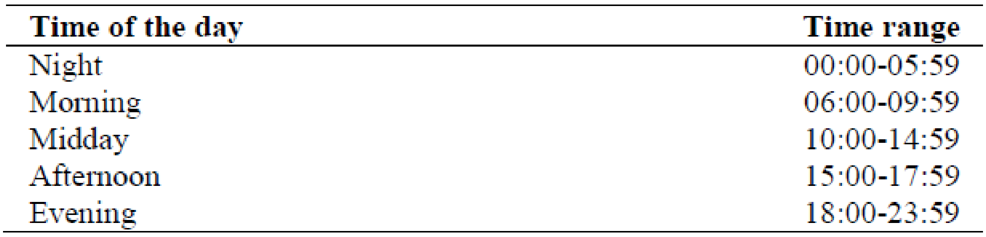
Additionally, our solution converts continuous features (e.g., the stored feature "scheduled arrival time" and the derived feature "arrival time difference") into discrete categories by binning the data. For example, scheduled arrival time throughout the day can be discretized into five categories: "morning," "midday," "afternoon," "evening," and "night." Similarly, the arrival time difference can be subdivided into categories such as "very early" and "early" for buses that arrive ahead of schedule, and "late" and "very late" for those that arrive behind schedule. See Tables I and II for further details.

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, including:

- Date/Time: Often recorded in a different time zone (e.g., Coordinated Universal Time (UTC)) than the location/city of interest. Our solution converts this recorded time into the local time in a 24-hour format for easy computation.

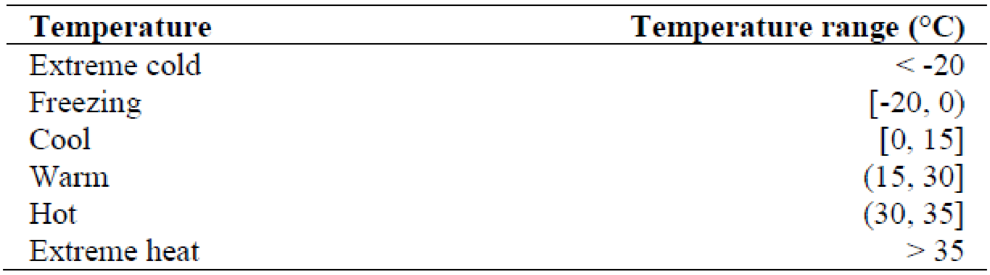
- Temperature: Usually measured in metric units (°C) in Canada. Data measured in other units (e.g., imperial unit °F) are converted to their equivalent metric units.

- Precipitation: Typically measured in metric units (mm) in Canada. Data measured in other units (e.g., imperial unit inches) are converted to their equivalent metric units. Depending on the temperature, precipitation can be in liquid form (rain) or solid form (snow).

- Other Related Weather Data: Includes categorical descriptions of weather conditions like "fog," "rain," "snow," "clear sky."

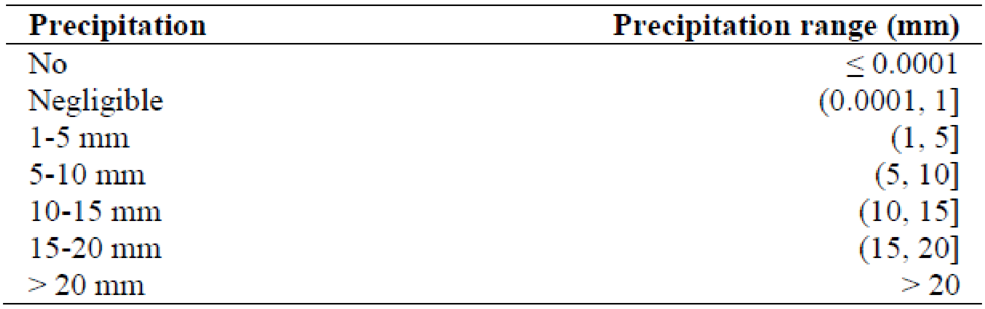
Similar to the preprocessing of transportation data, our solution converts continuous features into discrete categories by binning the data. The same binning procedure (see Table II) used for date/time in transportation data is applied to meteorological data. See Tables III and IV for the discretized categories for features such as "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 |

**Analyse Big Data**

Once the big data are integrated and pre-processed, our solution analyses the transportation and meteorological data to examine the impacts of weather conditions on public transit bus on-time performance. The analysis process involves several key steps:

1. Frequent Pattern Mining:

- Our solution conducts frequent pattern mining to identify recurring patterns, including weather conditions and their corresponding bus performance.

- To focus on the impact of weather conditions on bus performance, we specify that each pattern must include at least one weather condition and a bus performance category.

- This preference reduces the search space by eliminating patterns that do not meet these criteria.

- Additionally, infrequent patterns are removed based on a user-specified frequency threshold. For example, {fog, morning, on time} could be an infrequent pattern.

2. Forming Association Rules:

- After mining frequent patterns, our solution generates association rules from these patterns. These rules take the form "antecedent A → consequent C".

- To maintain focus on the impact of weather conditions on bus performance, we specify that:

- The antecedent of a rule must include at least one weather condition.

- The consequence of a rule must include a bus performance category.

- This preference further reduces the search space by eliminating association rules that do not meet these criteria.

The resulting rules reveal the associative relationships between weather conditions and bus performance categories. An association rule is considered interesting if it meets a user-specified frequency threshold (min freq) and a confidence threshold (min conf).

By incorporating these steps, our solution effectively examines the relationship between meteorological conditions and public transit bus performance, providing valuable insights into how weather conditions impact bus punctuality.

freq(A -> C) > minfreq

## EVALUATION

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

- Winnipeg Transit on-time performance data: This data captures public transit bus on-time performance (i.e., early, on-time, late) and is collected by the Global Positioning System (GPS) on-board computers equipped on all 640 buses. These buses cover 50,000 km, carry more than 48 million passengers, and operate for 1.5 million hours on 87 bus routes annually.

- Hourly or daily historical weather and climate data: This crucial data includes features such as temperature, total hourly precipitation amount (with a resolution of 0.1 mm), and 41 weather phenomena/events (e.g., rain, snow, fog, sky condition). For this paper, we focused on historical weather and climate data collected from two weather stations—Winnipeg International Airport (Station ID 5023227) and The Forks in downtown Winnipeg (Station ID 5023262). This data is of utmost importance as it provides the necessary context for the transit on-time performance data.

Once our solution integrated these datasets, we preprocessed the data as described in Section III. Initially, we visualized the data distribution to understand the dataset better. For instance,

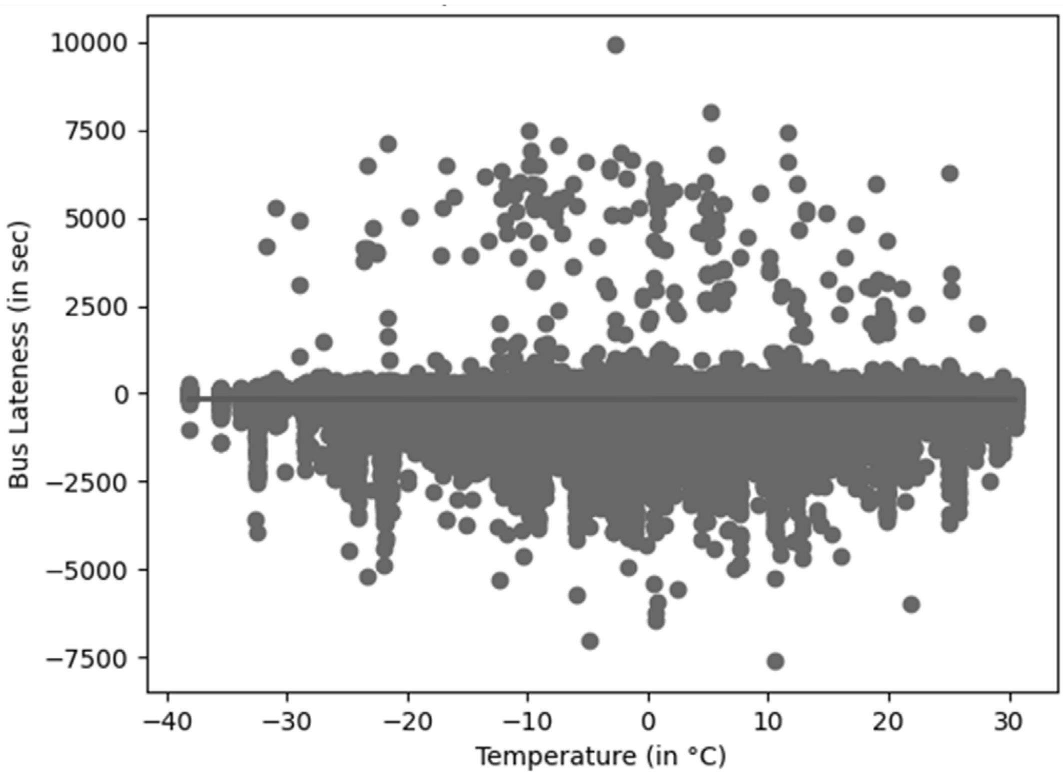


Fig. 1. Temperature vs. bus lateness (i.e., difference between scheduled and actual bus arrival times)

Then, Fig. 2 shows the correlations between the snow (ranging from 0.0mm to 1.2mm at an increment of 0.1mm) and bus lateness (with the same range of 125 minutes early to 167 minutes late).

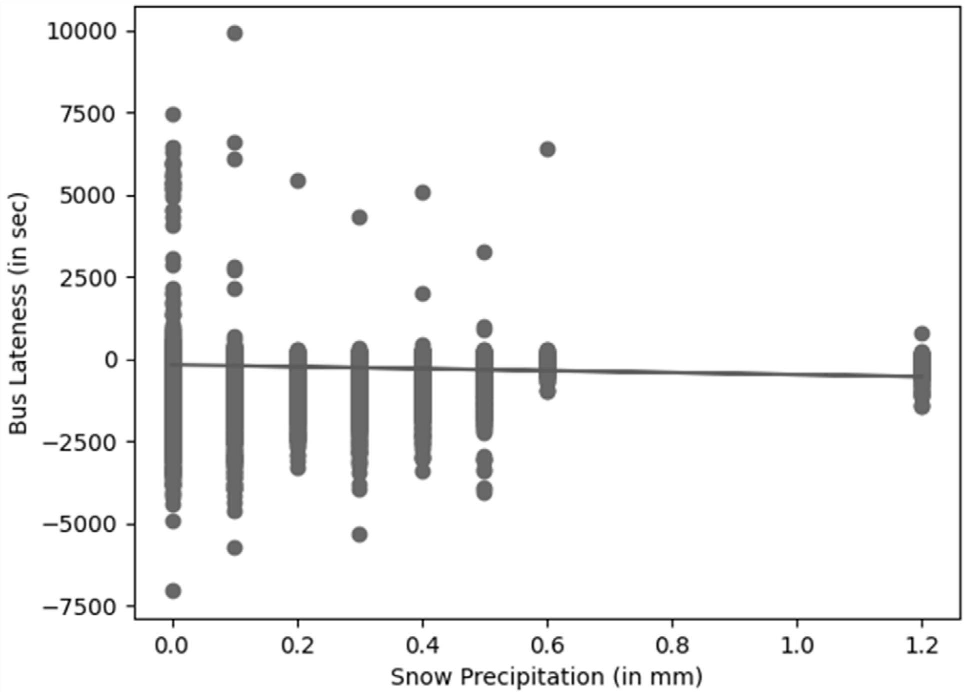


Fig. 2. Solid precipitation (i.e., snow) vs. bus lateness

Similarly, Fig. 3 shows the correlations between the rain

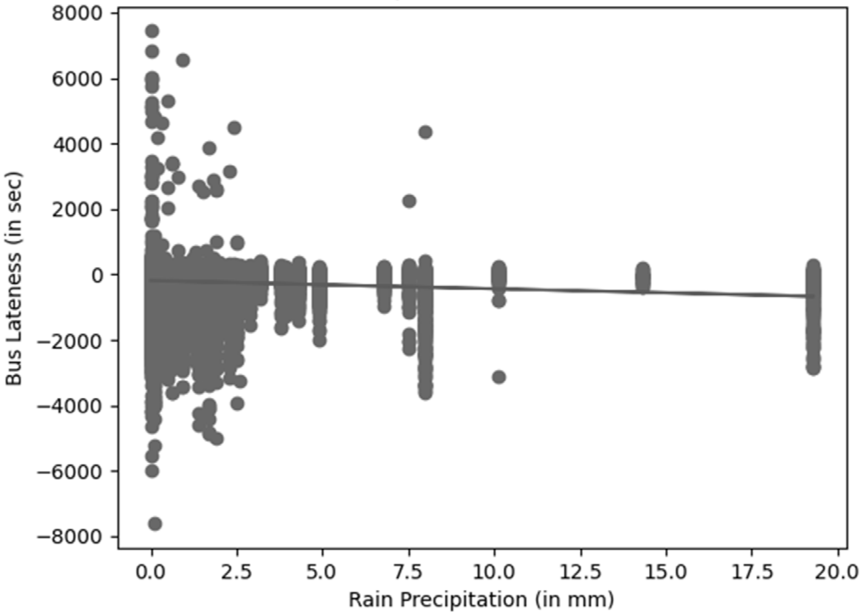


Fig. 3. Liquid precipitation (i.e., rain) vs. bus lateness

**Impacts of Meteorological Conditions on Bus Lateness**

Next, we applied our big data science solution to mine and analyze the preprocessed and integrated data to uncover frequent patterns and interesting association rules for transportation analytics. We specifically examined the impact of four meteorological conditions—clear sky, fog, rain, and snow—on the on-time performance of public transit buses.

Our analysis revealed 20 association rules, which are summarized in Table V. These rules are based on the four meteorological conditions and five bus on-time performance indicators: very early, early, on time, late, and very late. To account for variations in the number of days corresponding to each meteorological condition (e.g., clear skies may occur more frequently than fog, rain, or snow), we normalized the frequency of the five performance indicators within each meteorological condition. Consequently, the confidence of the rules directly reflects the frequency of occurrences under each weather condition.

Here’s a brief overview of the findings:

- Clear Sky: Generally associated with the least bus lateness.

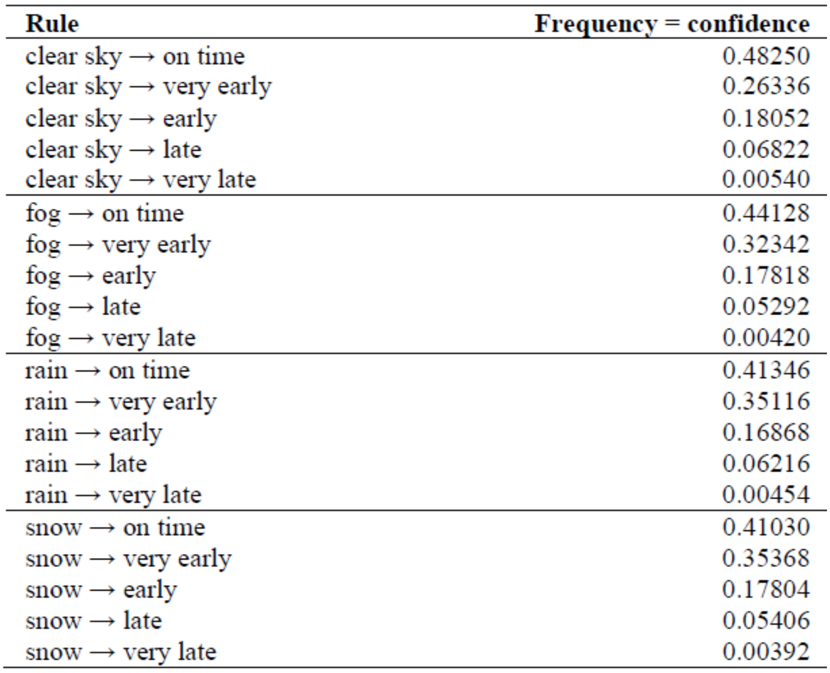
- Fog: Often linked to increased bus delays.

- Rain: Typically results in moderate delays.

- Snow: Most strongly associated with both significant delays and instances of buses being very late.

These patterns and rules help identify how different weather conditions affect bus punctuality, offering insights that can be used to improve transit scheduling and planning.

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 illustrates that, despite varying meteorological conditions, most buses arrived on time. The data reveals the following insights:

- On-Time Arrivals:

  - Clear Sky: 48.250% of buses arrived on time.

  - Fog: 44.128% of buses were on time.

  - Rain: 41.346% of buses met the schedule.

  - Snow: 41.030% of buses arrived on time.

- Most Common Performance Indicators:

  - For each meteorological condition, the most frequent bus performance categories were "very early" and "early."

- Minimal Delays:

  - Late: Less than 7% of buses were late under any weather condition.

  - Very Late: Fewer than 0.6% of buses were very late regardless of the weather.

These observations indicate that public transit buses generally maintain punctuality across different weather conditions, with only a small percentage of buses experiencing delays.

### Impacts of Time Intervals on Bus Lateness on Clear Days

In the next phase of our comprehensive analysis, we meticulously explored whether the effects of meteorological conditions varied across different time intervals throughout the day. Our approach involved thorough preprocessing of the data by categorizing time into five distinct intervals: morning, midday, afternoon, evening, and night, as detailed in Table II.

With data on four meteorological conditions (clear sky, fog, rain, and snow), five-time intervals, and five performance indicators (very early, early, on time, late, and very late), our solution identified 100 association rules. Table VI displays 25 rules, specifically from days with clear skies.

Key Findings from Clear Days:

- Busiest Time Intervals:

- Morning: The most active period, accounting for 35.10% of bus activities.

- Evening: The second busiest, with 24.95% of activities.

- Midday: Following, with 17.75% of activities.

- Afternoon: The fourth busiest, at 12.12%.

- Night: The least busy, with 9.97% of activities.

- Bus Performance Indicators:

- Most Frequent Indicator: On clear days, "on time" was the most common performance indicator, consistent across four of the five-time intervals.

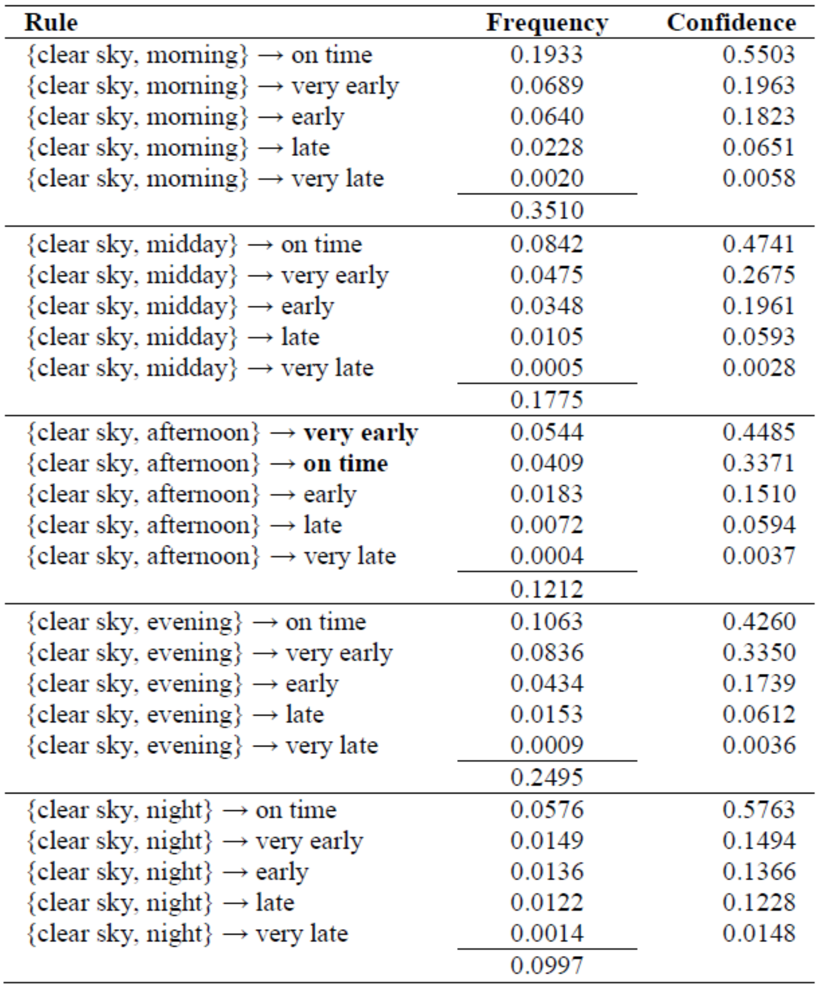
- Afternoon Exception: In the afternoon, "very early" was the most frequent performance indicator, with "on time" being the second most frequent.

- General Observation: Across all time intervals on clear days, "on time" and "very early" were the most common indicators, with "early" as the third most frequent.

These observations suggest that public transit buses performed exceptionally well on clear days, with "on time" and "very early" being the predominant performance outcomes throughout the day, reinforcing the reliability of the public transit system.

**Impacts of Time Intervals on Bus Lateness on Clear Days**

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 | |

Table VI reveals that the top three most frequent association rules on clear days are:

{Clear Sky, Morning} → On Time

{Clear Sky, Evening} → On Time

{Clear Sky, Midday} → On Time

These rules indicate that buses would most likely arrive on time in the morning, evening, and midday during clear sky conditions.

In contrast, during the clear afternoon period, the two most frequent rules were:

{Clear Sky, Afternoon} → Very Early

{Clear Sky, Afternoon} → On Time

These findings have important implications for bus riders. For instance, in the afternoon on clear days, buses were more likely to arrive very early rather than just on time. This suggests that riders should consider arriving at the bus stop a bit earlier in the afternoon to avoid missing a bus that might arrive ahead of schedule.

Additionally, the analysis shows that there was a 57.63% confidence that buses would arrive on time on clear nights. Similarly, there was a 55.03% confidence in buses arriving on time during clear mornings. This data underscores that buses performed exceptionally well in punctuality during these times on clear days.

These findings highlight that, under clear sky conditions, buses were generally punctual across different times of the day, with a notable tendency for earlier arrivals in the afternoon.

## CONCLUSIONS

In this paper, we introduced a comprehensive big data science solution designed for transportation analytics using meteorological data. Our approach effectively integrates, preprocesses, mines, and analyses large datasets to explore how weather conditions—such as clear skies, fog, rain, and snow—affect public transit bus on-time performance across various times of the day.

Our evaluation, based on real-life data from the Canadian city of Winnipeg, underscores the practicality and effectiveness of our solution. By dissecting the impacts of temperature and precipitation on bus punctuality, we have shown how our method can unearth actionable insights into public transit performance.

Looking ahead, we plan to extend our research to investigate the effects of more complex combinations of meteorological features on bus performance. Additionally, we aim to adapt our big data science solution to analyse other modes of transportation, broadening its applicability and impact in transportation analytics.

## 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] J. Gilmore, et al., “Commuters using sustainable transportation in census metropolitan areas,” Statistics Canada, 2017.

[29] K. Savage, “Results from the 2016 census: commuting within Canada’s largest cities,” Statistics Canada, 2019.