DOI: 10.63550/ICEIP.2025.32.65.002

Investigating Complex Interrelationships in the Transport System: Determining Effective Interval Widths for Modelling Traffic Characteristics

Igor Nikolaevich Pugachev[1][a], Nikolay Gennadyevich Sheshera [2][b]
Denis Yevgenyevich Grigorov [3][c],
[1]Khabarovsk Federal Research Center of the Far Eastern Branch of the Russian Academy of Sciences,,
[1] Dzerzhinkogo st., 54, Khabarovsk, Russian Federation
[2] Far Eastern State Transport University, Serysheva st., 47, Khabarovsk, Russia
[3] Khabarovsk Institute of Infocommunications SibGUTI, Lenin st., 73, Khabarovsk, Russia
ipugachev64@mail.ru, kolyaka239@mail.ru, glowfisch1996@yandex.ru

Keywords: Traffic flow intensity, Spearman's rank correlation, error probability, effective latitude intervals.

Abstract: Traffic flow intensity is a key indicator of the quality of transport services. Using modern urban road network management systems, it is possible to efficiently redistribute vehicular flows; however, this requires an understanding not only of the current traffic situation but also of the future development. This can be achieved through modeling road operating conditions. This paper proposes a comprehensive approach to studying the interrelationships among the elements of the Driver–Vehicle–Road–Environment (DVRE) system for effective modeling of traffic flow intensity variations under different weather conditions. To achieve this objective, we developed a software application using the Python programming language.

1 INTRODUCTION

Specialized agencies and research institutions conduct studies of road traffic accidents and the factors contributing to their occurrence in most industrially developed countries. It is crucial to note that traffic conditions, climatic influences, and other aspects of interaction among the elements of the Driver–Vehicle–Road–Environment (DVRE) system vary significantly not only between different countries but also within regions of a single country.

The physiological condition of the driver, which affects the driver's ability to interact with the components of the DVRE system, is influenced by a wide range of factors, including fatigue, health status, medication use, and alcohol intoxication (Khomutov D.K., 2020). Fatigue leads to reduced attention, diminished sensitivity in vision, hearing, and tactile perception, and increased reaction time. These effects represent a protective mechanism of the human body designed to ensure self-preservation and recovery through rest.

The vehicle, as a component of the DVRE system, may be analyzed from multiple perspectives: as an object of operation, maintenance, and repair; as the outcome of engineering design; and as an element of the economic and production systems involved in its creation and functional support (Levit B.Yu., Salin V.N., 2016). The reliability of its systems and components plays a critical role in ensuring road safety and effectiveness of the DVRE system.

The road infrastructure represents a vital element of the DVRE system and is characterized by various transport-operational parameters that directly influence traffic safety. These parameters include pavement smoothness, friction coefficient, carriageway width, and the presence of horizontal and vertical curves. Roads with minimal curvature and without significant inclines or declines are regarded as the least hazardous. However, in hilly or mountainous terrain, the construction of such roads

[a] https://orcid.org/0000-0003-0345-4350
[b] https://orcid.org/0009-0006-3302-5572
[c] https://orcid.org/0009-0005-4049-9488

requires substantial engineering efforts and financial investment due to the need to build bridges, overpasses, and embankments.

The environment affects all components of the DVRE system, and road infrastructure is continuously exposed to a wide range of influences—from daily and weather-related to seasonal and climatic variations. These changes can significantly alter driving conditions, which can lead to surface slipperiness or reduced visibility, both of which directly impact traffic safety.

To investigate and synthesize the patterns underlying transport-related indicators, we compiled a large dataset from various information sources, including geometric road elements, road characteristics, and meteorological and environmental conditions.

2 OPTIMIZATION OF INTERVAL WIDTHS FOR MODEL EFFICIENCY

2.1 Spearman’s rank correlation

Discretization, which involves grouping values into intervals, enhances information interpretation, improves visual clarity, and boosts model performance. Various methods are available for this purpose, including equal-width binning, percentile-based analysis, and expert judgment (Cherezov G.A., Leushin V.B., 2011).

Modern software tools enable automated selection of optimal intervals for analysis, offering a variety of variable partitioning techniques. In this study, we applied Spearman’s rank correlation method to determine the most appropriate interval structure. Since each interval represents a set of observed values, the use of standard correlation analysis methods is not appropriate in this context (Pugachev I.N., Sheshera N.G., 2024). The application of a rank-based method, in which, for example, the interval 0–200 is assigned rank 1, 200–400 is assigned rank 2, 400–600 is assigned rank 3, and so on, adequately considers the data structure and enables statistical analysis with high accuracy and reliability .

Spearman’s rank correlation is expressed as:

$$\rho = 1 - \frac{6\sum_{i=1}^{m} d_i^2}{n(n^2-1)}, \qquad (1)$$

where d_i is the difference between the ranks, defined as $d_i = R(X_i) - R(Y_i), n$, and n is the number of observations used to calculate the coefficient.

Spearman’s rank correlation method is a statistical approach designed to assess the strength and direction of the relationship between two variables represented as ranked data. This method serves as an alternative to Pearson’s correlation coefficient and is particularly useful when the data do not conform to a normal distribution or contain outliers.

During the analysis, each variable is assigned rank values in ascending or descending order, after which the researcher calculates the differences between the corresponding ranks. The resulting Spearman’s rank correlation coefficient characterizes the degree of dependence between these rank differences, allowing the identification of statistically significant relationships within the sample.

The coefficient ranges from −1 to +1. A value of +1 indicates a positive monotonic relationship, in which an increase in one variable is accompanied by an increase in the other. A coefficient of −1 indicates a negative monotonic relationship, where an increase in one variable corresponds to a decrease in the other (Zhu J., Ma L., Ni M., Li Z., 2021; Kosheleva N.N., 2012). A correlation coefficient of 0 indicates the absence of a monotonic relationship between the variables.

One of the key advantages of Spearman’s rank correlation over Pearson’s correlation is its robustness to outliers, as well as the absence of a requirement for the data to be normally distributed. This method is widely used by researchers in statistical analysis and can be implemented using various software tools, including Python, R and SPSS (Liquet B., Riou J., 2019).

2.2 Determination of Effective Interval Widths

As part of a study focused on analyzing the influence of temperature-related factors (including air temperature, soil temperature, and dew point) on traffic flow intensity, we conducted a procedure to determine the optimal interval widths for all considered variables (Shper V.L., 2021). The following values were established (Pugachev I.N., Sheshera N.G., Grigorov D.E., 2024):

Traffic flow intensity: 100, 200, 250, 300, 400, 500, 600 vehicles/hour.

Temperature indicators (air temperature, soil temperature, and dew point): 5, 10, 15, 20, 25 °C.

The impact of temperature conditions on traffic flow may depend on additional factors such as the day of the week, time of day, and precipitation. In particular, a decrease in traffic intensity is typically observed on weekends, due to the lack of necessity for regular commuting (Hill J.B., 2024). Moreover, weather conditions may influence individuals' emotional states, which in turn affects their decision to leave home.

Considering these factors, the analysis of dependencies was conducted not only for all possible combinations of interval widths for traffic intensity and temperature indicators, but also with consideration of different days of the week and time periods. As a result, a total of 126,000 different combinations and corresponding correlation coefficients were obtained (Pugachev I.N., Grigorov D.Y., Skripko P.B., Sheshera N.G., 2024).

As the primary dataset on traffic flow intensity, we used a sample containing information on traffic conditions and meteorological parameters at control point 1311 (Figure 1), located at the intersection of Muravyova-Amurskogo Street and Dzerzhinskogo Street in Khabarovsk. This dataset comprises 8,115 records collected during the operational period of the video surveillance system.

We developed a Python-based software application to perform this analysis (Figure 2).

For analysis, four iterable lists were generated, containing various parameters such as days of the week, time intervals, and possible classification methods for both temperature conditions and traffic flow intensity. As the primary dataset on traffic flow intensity, a sample was used containing information on traffic conditions and meteorological parameters at control point 1311, located at the intersection of Muravyova-Amurskogo Street and Dzerzhinskogo Street in the city of Khabarovsk.

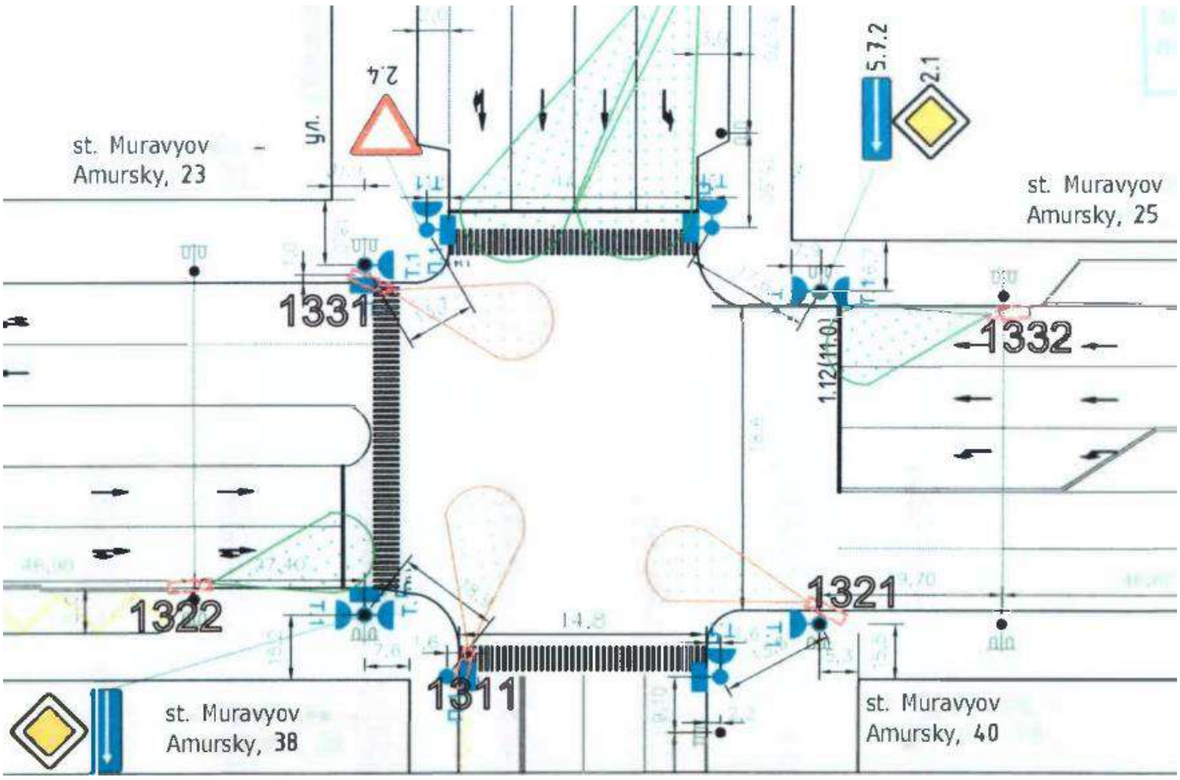

Figure 1: Schematic layout of the intersection with installed lane control checkpoints (1311, 1321, 1331).

This dataset consists of 8,115 records collected during the operational period of the video surveillance system (specify the time frame, if applicable).

For analysis, we generated four iterable lists containing various parameters such as days of the week, time intervals, and classification methods for both temperature conditions and traffic flow intensity (Pudova N.V., Nikitin V.V., 2004).

2.3 Threshold Criterion for Effective Interval Widths

The Spearman error probability statistic (p-value) reflects the probability of obtaining the observed value of the rank correlation coefficient, or a more extreme value (closer to 1 or −1), assuming that the null hypothesis is true. In statistical analysis, researchers generally accept that if the p-value falls below a predetermined significance level (e.g., 0.05), the null hypothesis is rejected, and the identified correlation is deemed statistically significant (Balasyan, B.G., Matveeva, T.A., Svetlichnaya, V.B., 2019).

It is important to emphasize that the p-value does not indicate the strength of the correlation itself, but rather the level of statistical significance of the result. The smaller the p-value, the higher the likelihood that the observed correlation is not due to random chance and is indeed present in the general population.

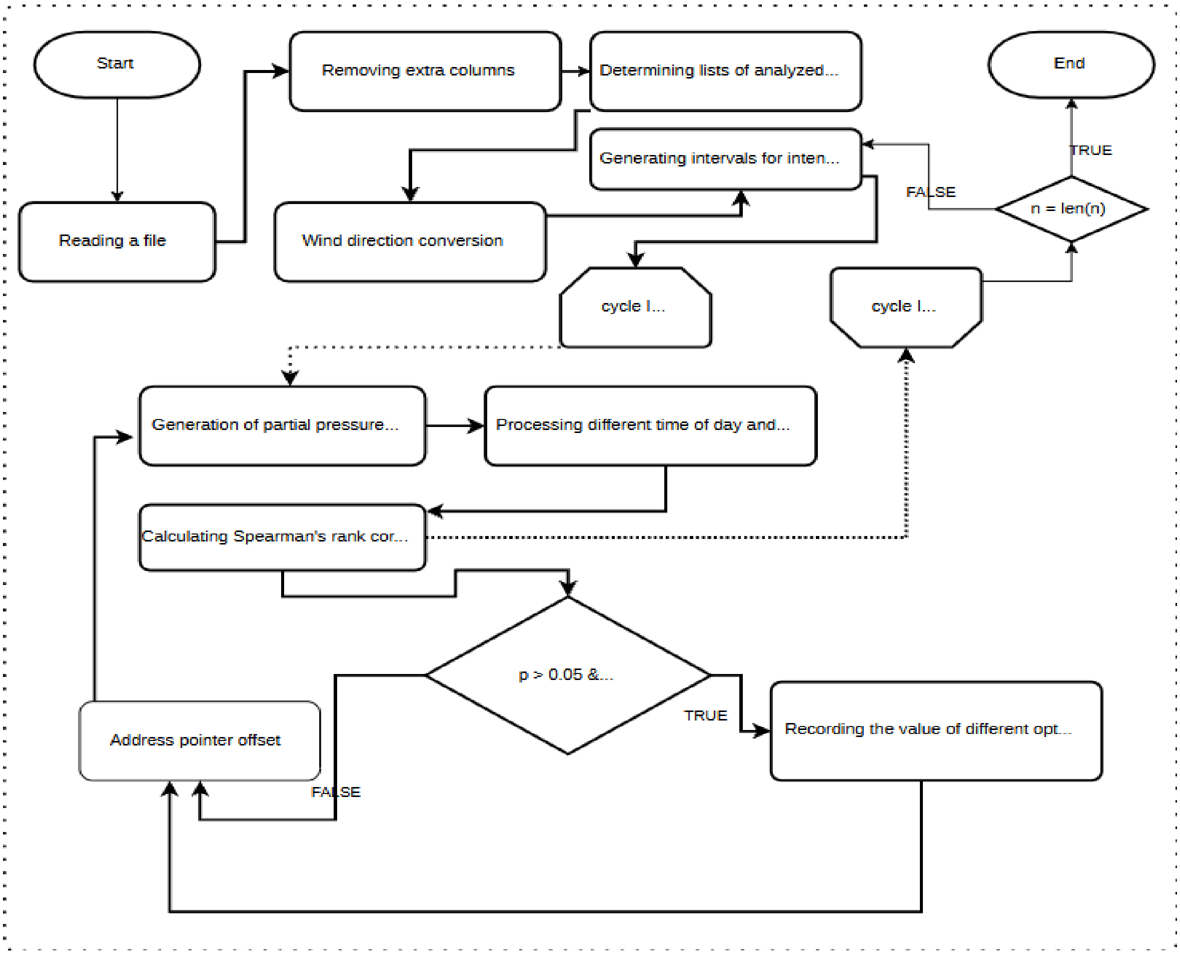

Figure 2: Block diagram of the algorithm for analyzing interval widths using Spearman's rank correlation.

To ensure the integrity of the experiment, the parameter «Precipitation» was included in the analysis, but only if its value was zero. In this study, precipitation is considered an anomalous factor that may distort the results; therefore, its exclusion from the dataset is deemed justified (Pugachev I.N., Sheshera N.G., Grigorov D.E., 2024).

Within each combination of interval widths for traffic intensity and temperature indicators, the researchers assigned corresponding ranks, followed by correlation analysis. In the next step, one of the parameters (specify which parameter) was modified, and the Spearman rank correlation was recalculated. This procedure allowed a systematic examination of all possible combinations of interval widths and the identification of the most significant dependencies. To select the most meaningful relationships, we developed a criterion whereby only those records were included in the final sample for which the correlation coefficient ranged from 0.7 to 1 or from −0.7 to −1, and the p-value did not exceed 0.05. This approach enabled the exclusion of statistically insignificant combinations, reducing the original dataset from 126,000 possible combinations to 243 statistically significant cases.

Table 1: Fragment of Spearman’s Rank Correlation Data Between Traffic Flow Intensity Interval Widths and Temperature Conditions (Air, Soil, and Dew Point).

p_air	p_soil	p_dew	rho_air	rho_soil	rho_dew	day	time	t_air	t_soil	t_dew	int
4.6E-10	2.17E-09	4.95E-10	-0.79	-0.77	-0.79	7	17	5	5	5	100
4.6E-10	2.17E-09	3.27E-09	-0.79	-0.77	-0.77	7	17	5	5	10	100
4.6E-10	2.17E-09	1.66E-09	-0.79	-0.77	-0.78	7	17	5	5	15	100
4.6E-10	2.17E-09	2.07E-07	-0.79	-0.77	-0.7	7	17	5	5	25	100
4.6E-10	1.38E-08	4.95E-10	-0.79	-0.75	-0.79	7	17	5	10	5	100
1.4E-10	1.59E-09	2.27E-08	-0.8	-0.78	-0.74	7	17	10	5	10	200
1.4E-10	1.59E-09	4.87E-08	-0.8	-0.78	-0.73	7	17	10	5	15	200
1.4E-10	1.02E-08	5.42E-09	-0.8	-0.75	-0.76	7	17	10	10	5	200
1.4E-10	1.02E-08	2.27E-08	-0.8	-0.75	-0.74	7	17	10	10	10	200
4.6E-10	3.21E-09	3.27E-09	-0.79	-0.77	-0.77	7	17	5	25	10	100
7.1E-10	1.73E-07	1.16E-08	-0.81	-0.73	-0.77	6	17	25	10	20	300
7.1E-10	1.73E-07	5.9E-07	-0.81	-0.73	-0.71	6	17	25	10	25	300
7.1E-10	1.75E-07	5.53E-08	-0.81	-0.73	-0.75	6	17	25	15	5	300
7.1E-10	1.75E-07	2.8E-08	-0.81	-0.73	-0.76	6	17	25	15	10	300
7.1E-10	1.75E-07	1.96E-08	-0.81	-0.73	-0.77	6	17	25	15	15	300
9.8E-09	1.65E-08	1.23E-07	-0.75	-0.74	-0.71	7	17	15	10	5	400
9.8E-09	1.65E-08	2.26E-07	-0.75	-0.74	-0.7	7	17	15	10	15	400
8.4E-08	1.37E-07	5.48E-08	0.738	0.73	0.74	1	8	15	10	20	500
9.9E-07	1.22E-07	2.06E-08	0.729	0.77	0.79	5	8	15	10	20	500
3.39E-07	2.17E-07	4.02E-06	-0.712	-0.70	-0.74	7	17	10	5	5	250
3.39E-07	2.17E-07	3.8E-06	0.712	-0.70	-0.74	7	17	10	5	10	250
3.39E-07	2.4E-07	4.02E-06	0.712	-0.70	-0.74	7	17	10	10	5	250
3.39E-07	2.4E-07	3.8E-06	0.712	-0.70	-0.74	7	17	10	10	10	250
6E-08	2.17E-07	4.02E-06	-0.724	-0.70	-0.74	7	17	25	5	5	250

3 PRELIMINARY RESULTS

For traffic flow intensity, the following intervals were selected:

- 60 options with a width of 100 vehicles/hour;
- 75 options with a width of 200 vehicles/hour;
- 21 options with a width of 250 vehicles/hour;
- 47 options with a width of 300 vehicles/hour;
- 40 options with a width of 400 vehicles/hour.

The objective of this study is to determine the optimal interval widths that provide the most accurate forecasting of traffic flow intensity. As a result of generating various parameter combinations, 243 statistically significant options were identified, reflecting the relationship between the temperature conditions and traffic intensity. However, for further analysis, a single most suitable option must be selected, given that the dependent variable is traffic flow intensity (Tanaseychuk A.V., Lemeshko B.Yu., 2009).

For a more detailed analysis, an interval width of 100 vehicles/hour is considered; however, in practice, the most common interval is 300 vehicles/hour. To enhance the reliability of the results, we prefer this option. It is important to note that the study of relationships between traffic flow intensity and meteorological factors is not limited to this aspect, as there are other meteorological factors for which the 300 vehicles/hour interval could be ineffective.

The concept of selecting the traffic flow intensity interval width based on temperature conditions is outlined as follows (Pugachev I.N., Sheshera N.G., Grigorov D.E., 2024). Suppose that all independent variables show a significant correlation with the traffic flow intensity at the 250 vehicles/hour interval. In the city of Khabarovsk, where traffic intensity at one control point can reach 5000 vehicles/hour, this interval provides a sufficient level of granularity for analysis. In this case, the intensity intervals can be defined as follows: 0–250 vehicles/hour; 251–500 vehicles/hour; 501–750 vehicles/hour; …; 4751–5000 vehicles/hour.

When analyzing the relationship between variables to assess the influence of one variable on another, it is often sufficient to apply pairwise correlation analysis. However, this study examines the cumulative effect of temperature parameters on traffic flow intensity. Since independent variables such as air temperature, soil temperature, and dew point are influenced by common factors (e.g., seasonal changes) and simultaneously affect the dependent variable, their analysis requires a comprehensive approach that considers various conditions of their impact (Pugachev I.N., Sheshera N.G., Grigorov D.E., 2024).

When selecting the traffic intensity interval width, for example, 250 vehicles/hour, all other options are excluded. Next, among the 21 possible options, the second most significant factor is determined—the width of the temperature interval:

- 12 options with a width of 5 °C;
- 6 options with a width of 10 °C;
- 2 options with a width of 15 °C;
- 1 option with a width of 20 °C.

When analyzing soil temperature, the most common intervals are also 5 °C, 10 °C, 15 °C, and 20 °C. To ensure a more detailed analysis, the smallest width of 5 °C is selected (e.g., -40…-35.1; -35…-30.1; -30…-25.1; -25.2…0.1, etc.).

Using a similar method, by fixing the soil temperature interval step at 5 °C and excluding all other options, the interval width for soil temperature is determined, followed by that for dew point temperature.

As a result, the following optimal interval widths were identified:

- traffic flow intensity – 250 vehicles/hour;
- air temperature – 5 °C;
- soil temperature – 5 °C;
- dew point temperature – 5 °C.

Thus, the proposed method allows for the systematic determination of the most significant interval widths for forecasting traffic flow intensity, taking into account the influence of temperature factors (Navoy D.V., Kapskiy D.V., Filippova N.A., Pugachev I.N., 2024). The impact of all other features on traffic flow intensity can be determined through the generation of interval widths and pairwise correlation. If there is a need to group data and determine an effective interval width, Spearman's rank correlation can be applied in the unique value generator, following the algorithm provided above and using the existing program structure.

4 PAIRWISE RANK CORRELATION OF SPEARMAN FOR DETERMINING EFFECTIVE INTERVAL WIDTHS OF THE INFLUENCE OF CONCURRENT WEATHER FACTORS ON CHANGES IN TRAFFIC FLOW INTENSITY

Similarly, in studies examining the influence of temperature regimes on traffic flow intensity, the previously developed software concept was utilized to generate various combinations. Only the lists of interval width values were substituted.

For the partial pressure of water vapor, the following list of interval widths was adopted: 5, 10, 15. The best results of the pairwise Spearman rank correlation and their corresponding effective interval widths are presented in Table 2.

Table 2: Significant Results of Pairwise Spearman Rank Correlation Between Partial Pressure of Water Vapor and Traffic Flow Intensity.

p_pressure	rho_pressure	int	pressure
3.07E-07	-0.72227	100	5
1.56E-09	-0.77606	100	5
3.61E-08	-0.73204	200	5
3.60E-07	-0.71204	250	5
2.77E-07	-0.72411	300	5

For relative humidity, the following list of interval widths was adopted: 10, 20, 30, 50. The best pairwise Spearman rank correlation coefficient when analyzing the influence of relative humidity on traffic flow intensity was -0.65, with an error probability of 1.64E-05, at an intensity interval width of 250 and a relative humidity of 50.

The influence of this variable is weak, but it is present, and given the negligible error probability, the data is accepted (Zguralskaya E.N., 2018).

For saturation deficit, the following list of interval widths was adopted: 5, 10, 15. The best results of the pairwise Spearman rank correlation and their corresponding effective interval widths are presented in Table 3.

Table 3: Significant Results of Pairwise Spearman Rank Correlation Between Saturation Deficit Interval Widths and Traffic Flow Intensity.

p_saturation	rho_saturation	int	saturation
1.72E-06	-0.66295	100	5
5.12E-07	0.668812	100	5
2.04E-06	-0.6594	200	5
3.56E-06	0.629839	200	5
2.56E-05	0.669839	250	5
2.56E-05	0.669839	250	5
5.82E-06	0.619	300	5
3.34E-07	-0.69462	400	5
9.91E-06	0.606839	400	5
1.54E-05	0.607768	500	5

For station-level pressure, the following list of interval widths was adopted: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50. The best results of the pairwise Spearman rank correlation and their corresponding effective interval widths are presented in Table 4.

Table 4: Significant Results of Pairwise Spearman Rank Correlation Between Station-Level Pressure Interval Widths and Traffic Flow Intensity.

P stations	rho stations	stations	int	p stations	rho stations	stations	int
0	1	30	100	0	1	35	200
0	1	30	100	0	1	30	250
0	1	35	100	0	1	30	250
0	1	35	100	0	1	30	250
0	1	35	100	0	1	35	250
0	1	30	200	0	1	35	250

For sea-level pressure, the following list of interval widths was adopted: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50. The best results of the pairwise Spearman rank correlation and their corresponding effective interval widths are presented in Table 5.

Table 5: Significant Results of Pairwise Spearman Rank Correlation Between Sea-Level Pressure Interval Widths and Traffic Flow Intensity.

p_sea	rho_sea	sea	int	p_sea	rho_sea	sea	int
0	1	45	100	0	1	45	250
0	1	45	100	2.34E-07	-0.7271	20	500
0	1	45	200	2.34E-07	-0.7271	30	500

When analyzing the influence of wind speed on traffic flow intensity, it is essential to consider various combinations of qualitative features, not only those related to the day of the week or time of day, but also, in this case, wind direction. Wind direction is typically measured in degrees from true north, where 0° indicates north, 90° indicates east, 180° indicates south, and 270° indicates west, respectively. In combination with wind speed, it can have a significant impact on traffic flow intensity.

If the wind blows in the direction of traffic flow (tailwind), it facilitates vehicle movement by reducing air resistance and improving fuel efficiency. This effect is particularly noticeable on highways and at high speeds. Conversely, if the wind blows against the direction of traffic flow (headwind), it creates additional resistance, slows down vehicles, and increases fuel consumption (Pugachev I.N., Sheshera N.G., Grigorov D.E., 2024). This can be especially problematic for heavy vehicles such as trucks or buses, having a large frontal aerodynamic surface area.

In addition, strong or gusty winds can create hazardous driving conditions, particularly for high-profile vehicles such as trucks, buses, or trailers, as they may cause lateral displacement or loss of control. This may necessitate the implementation of speed restrictions or temporary road closures by traffic authorities.

To analyze the influence of wind speed on traffic flow intensity, eight wind direction groups were defined, and the following interval widths were adopted:

- N (North): 337.5°–22.5° = 1;
- NE (Northeast): 22.5°–67.5° = 2;
- E (East): 67.5°–112.5° = 3;
- SE (Southeast): 112.5°–157.5° = 4;
- S (South): 157.5°–202.5° = 5;
- SW (Southwest): 202.5°–247.5° = 6;
- W (West): 247.5°–292.5° = 7;
- NW (Northwest): 292.5°–337.5° = 8.

During the wind speed data preprocessing stage, each value was reassigned to the corresponding group number. For example, a value of 125° was recoded as 4, 254° as 7, and 55° as 2, and so on. This transformation was necessary due to the challenges associated with correctly defining degree-based intervals. The degree scale begins at north and increases clockwise. East corresponds to 90°, south to 180°, and west to 270°. The northern direction is defined as the range from 337.5° to 22.5°, which is difficult to implement using pandas libraries due to the lack of a continuous numeric sequence from negative to positive values.

A list of intervals for the new wind direction data was created: 1, 2, 3, 4, 5, 6, 7, 8. Additional functions were integrated into the previously developed software to enable the generation of wind speed intervals and wind direction groups.

The best results of the pairwise Spearman rank correlation reflecting the influence of wind speed on traffic flow intensity, along with the corresponding effective interval widths, are presented in Table 6.

Table 6: Fragment of Significant Results of Pairwise Spearman Rank Correlation Between Wind Speed and Direction, and Traffic Flow Intensity.

Wind Speed (P)	Wind Speed	Wind Direction	Wind Speed	Value
0.0355	0.74	2	2	100
0.0337	-0.74	2	2	100
0.0067	-0.96	3	3	200
0.0387	0.77	2	2	250
0.0001	-0.84	2	2	250
0.0072	-0.85	3	3	300
0.0030	-0.73	4	4	400
0.0240	-0.77	4	4	400
0.0165	-0.80	2	2	500
0.0098	-0.73	2	2	500

A distinctive feature of analyzing the influence of precipitation on traffic flow intensity is the generally low density of such relationships. Typically, in a given dataset, the proportion of records containing precipitation events may be less than 10%.

Therefore, to clarify the dependency, it was decided to filter the sample. The selection criterion was the presence of any precipitation greater than 0.

The following list of interval widths was adopted for precipitation: 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15. The corresponding function was implemented.

The best results of the pairwise Spearman rank correlation and their respective effective interval widths are presented in Table 7.

Table 7: Significant Results of Pairwise Spearman Rank Correlation Between Precipitation and Traffic Flow Intensity.

p_ precipitation	rho_ precipitation	interval_ precipitation	int	p_ precipitation	rho_ precipitation	interval_ precipitation	int
0.023158	0.777462	2	100	0.00815	0.720962	7	200
0.00715	0.728982	2	100	0.00815	0.720962	2	250
0.033777	-0.74541	2	100	0.00815	0.720962	2	250
0.023465	0.776419	3	100	0.00815	0.720962	3	250
0.044961	-0.7178	4	100	0.003209	0.772804	2	300
0.005092	0.748602	7	100	0.002461	0.785544	7	300
0.005092	0.748602	8	100	0.002461	0.785544	8	300
0.005092	0.748602	9	100	0.002461	0.785544	9	300
0.005092	0.748602	10	100	0.002461	0.785544	10	300
0.001768	0.800327	2	200	0.001996	0.795046	2	400
0.020312	-0.78756	2	200	0.045256	-0.71714	2	400
0.008936	0.715165	3	200	0.00717	0.728811	4	400
0.005663	0.742635	4	200	0.00717	0.728811	5	400
0.005663	0.742635	5	200	0.032104	-0.75	2	500

5 CONCLUSION

Grouping and sorting are essential operations in data processing. They are performed to facilitate more convenient and systematic analysis and presentation of information. Within the framework of this study, which is based on the application of Spearman's rank correlation method, a procedure was carried out to determine the optimal interval widths of independent variables that influence the dependent variable—traffic flow intensity. Since traffic intensity is the key aspect of the analysis, the primary objective was to identify the most informative intervals of this indicator, taking into account the correlation coefficient and the level of statistical significance (p-value).

An analysis of the dataset, which included data grouped by day of the week, time intervals, and temperature parameters (air temperature, soil temperature, and dew point), made it possible to identify 243 statistically significant combinations, including:

- 60 cases with an interval width of 100 vehicles/hour;
- 75 cases with an interval width of 200 vehicles/hour;

- 21 case with an interval width of 250 vehicles/hour;
- 47 cases with an interval width of 300 vehicles/hour;
- 40 cases with an interval width of 400 vehicles/hour.

The most frequently occurring interval width was 200 vehicles/hour, which could suggest its potential suitability for further analysis. However, in subsequent studies of relationships with meteorological and environmental factors, this interval did not demonstrate sufficient informativeness. The analysis revealed that the only interval width that maintained statistical significance across all considered weather and environmental conditions was 250 vehicles/hour. This interval was therefore identified as optimal for analyzing traffic flow intensity.

The determination of interval widths for the remaining features was carried out with a fixed traffic intensity interval width of 250 vehicles/hour. Preference was given to those values that occurred most frequently within the specified range and had a correlation coefficient of ≥ 0.7 at a significance level of $p < 0.05$. In exceptional cases, a threshold of ≥ 0.6 was permitted. In the event of an equal number of observations within the 250 vehicles/hour interval, priority was given to the smallest interval width, which ensured a more detailed representation of the data for further analysis. Based on the results of the correlation analysis and the established selection criteria, the following optimal interval widths were determined for the quantitative features (Pugachev I.N., Grigorov D.E., Sheshera N.G., 2024):

- Traffic flow intensity (N): 250 vehicles/hour;
- Air temperature (t_air): 5 °C;
- Soil temperature (t_soil): 5 °C;
- Dew point temperature (DP): 5 °C;
- Partial pressure of water vapor (p_s): 5 Pa;
- Relative humidity (φ): 50%;
- Saturation deficit (d): 5 g/m³;
- Atmospheric pressure at station level (P_station): 30 hPa;
- Atmospheric pressure at sea level (P_sea): 45 hPa;
- Wind speed: 2 m/s;
- Precipitation: 2 mm.

Thus, the conducted analysis made it possible to identify the most significant intervals for forecasting traffic flow intensity, taking into account the influence of meteorological factors (Galimyanov F.A., 2020). These findings can be used for further modeling and optimization of traffic flows. The results of this study formed the basis of a larger project aimed at forecasting traffic flow intensity. (Pugachev I.N., Evtyukov S.S., Sheshera N.G., Grigorov D.E., 2024). Looking ahead, it is worth presenting intermediate results obtained using the effective interval widths identified in this work (Figure 3).

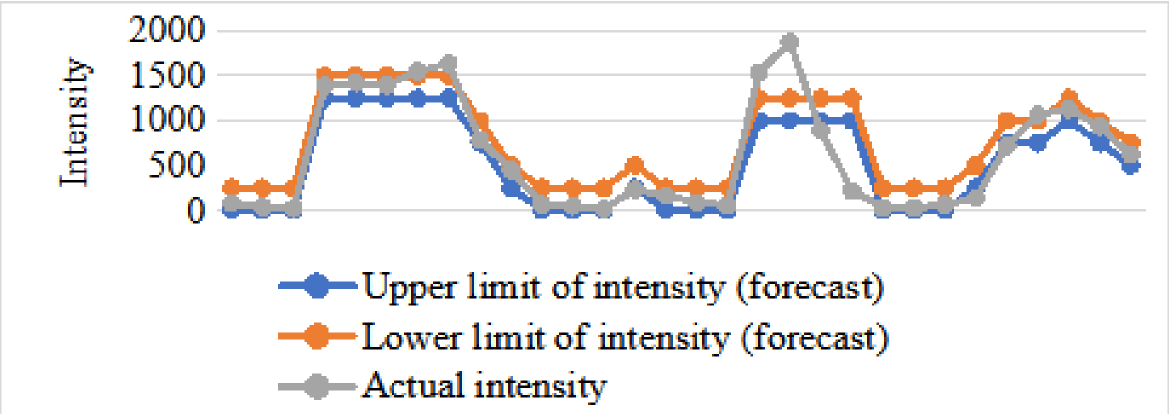

Figure 3: The comparative analysis of forecasted intensity intervals with actual values.

REFERENCES

Khomutov, D.K., 2020. Analysis of statistical data using the Spearman correlation coefficient. In the collection: SNK-2020. Proceedings of the Jubilee LXX open international student scientific conference of the Moscow Polytechnic University. Moscow. Pp. 865-868.

Levit, B.Yu., Salin, V.N. ,2016. Comparative analysis of the exact and approximate formulas for calculating the Spearman rank correlation coefficient. *Questions of Economics and Law.* No. 93. Pp. 104-110.

Zhu, J., Ma, L., Ni, M., Li, Z., 2021. A bootstrap method to calculate the p-value of fisher's combination for a large number of weakly dependent p-values. Communications in Statistics Part B: Simulation and Computation.

Kosheleva, N.N., 2012. Correlation analysis and its application to calculating the Spearman rank correlation. *Actual problems of the humanitarian and natural sciences.* № 5. P. 23-26.

Cherezov, G.A., Leushin, V.B., 2011. Calculating the Spearman rank correlation coefficient in Matlab. In the collection: Education, science, transport in the 21st century: experience, prospects, innovations. II Regional scientific conference dedicated to the 50th anniversary of Russian cosmonautics. Pp. 27-28.

Pugachev, I.N., Sheshera, N.G., 2024. Mathematical modeling in transport studies using data obtained from navigation and telematics systems. *Innovative transport.* No. 4 (54). P. 3-11.

Liquet, B., Riou, J., 2019. Cpmcglm: an r package for p-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology.* Vol. 19. No. 1. P. 1-8.

Shper, V.L., 2021. What is a p-value, and why is there so much noise around it? *Product quality control.* № 3. P. 37-43.

Hill, J.B., 2024. A smoothed p-value test when there is a nuisance parameter under the alternative. *Journal of Statistical Planning and Inference*. Vol. 229. P. 106-096.

Pugachev, I.N., Grigorov, D.Y., Skripko, P.B., Sheshera, N.G., 2024. Improving the Methodology for Determining Injury Rates in Road Accidents using Machine Learning - The Random Tree Method. International scientific and practical conference «Smart cities and sustainable development of regions» (SMARTGREENS 2024): LLC Institute of Digital Economics and Law [OOO «Institut tsifrovoy ekonomiki i prava»]. P. 125-136.

Balasyan, B.G., Matveeva, T.A., Svetlichnaya, V.B., 2019. Calculation of the Spearman rank correlation coefficient. In the collection: Modern problems of science and education. Proceedings of the XI International student scientific conference. P. 133-134.

Pudova, N.V., Nikitin, V.V., 2004. Analysis of the values of the Spearman rank correlation coefficient. *Economic Analysis: Theory and Practice.* № 3 (18). P. 52-56.

Pugachev, I.N., Sheshera, N.G., Grigorov, D.E., 2024. Study of feature interval widths to improve the efficiency of traffic flow intensity forecasting. *Bulletin of SIBADI.* No. 5-(21). P. 726-735.

Pugachev, I.N., Sheshera, N.G., Grigorov, D.E., 2024. Using the tools of modern PYTHON libraries for working with geographic coordinates to solve road safety problems. *Transport of the Asia-Pacific region*. No. 3 (40). P. 60-66.

Tanaseychuk, A.V., Lemeshko, B.Yu., 2009. Distributions of statistics constructed on the basis of the Spearman rank correlation coefficient. In the collection: Computer science and problems of telecommunications. Russian scientific and technical conference. P. 67-69.

Pugachev, I.N., Sheshera, N.G., Grigorov, D.E., 2024. Research of traffic flow intensity using the Deep learning method. *World of transport.* Vol. 22. No. 2 (111). P. 12–24.

Pugachev, I.N., Sheshera, N.G., Grigorov, D.E., 2024. Methodology of intelligent traffic management taking into account changing weather conditions. *Transport of the Urals* No. 4 (83). P. 10-16.

Navoy, D.V., Kapskiy, D.V., Filippova, N.A., Pugachev, I.N., 2024. Analysis of world experience in the application of artificial intelligence in traffic management systems of various levels. *Systems analysis and applied informatics.* (1):26-36.

Zguralskaya, E.N., 2018. Stability of data partitioning into intervals in recognition problems and search for hidden patterns. *Bulletin of the Samara Scientific Center of the Russian Academy of Sciences.* Vol. 20, No. 4-3 (84). P. 451-455.

Pugachev, I.N., Sheshera, N.G., Grigorov, D.E., 2024. Definition of stages of road safety control (design, construction and operation) for an improved methodology of accident rates. *Transport of the Asia-Pacific Region.* No. 2 (39). P. 54–60.

Pugachev, I.N., Grigorov, D.E., Sheshera, N.G., 2024. Study of road safety from the standpoint of driver information load. Techn.

Galimyanov, F.A., 2020. Comparative analysis of algorithms for implementing the backpropagation method for training neural networks. *Scientific and Technical Bulletin of the Volga Region.* No. 2. P. 69-71.

Pugachev, I.N., Evtyukov, S.S., Sheshera, N.G., Grigorov, D.E., 2024. Forecasting the intensity of traffic flow. Learning with a teacher. Random trees method. T-Comm: Telecommunications and transport. Vol. 18. No. 4. P. 36-47.