STUDIES OF EFFECTS INFLUENCING THE TRANSMISSION OF UDP DATAGRAMS BETWEEN RAILWAY VEHICLES AND INFRASTRUCTURE

Zuzana Kleprlíková[1], David Žák[1]

[1]University of Pardubice, Faculty of Electrical Engineering and Informatics,
Studentská 95, Pardubice 2, 53210, Czech Republic

zuzana.kleprlikova@upce.cz (Zuzana Kleprlíková)

Abstract

This research applies to the quality of GSM signal cover on railways. On the map of Czech Republic we try to display and color-highlight places where there is reduced probability of delivery of UDP datagrams carrying information on the current train position. These messages are sent periodically from a GPS module placed in a train after satisfying certain conditions. We are monitoring if all sent messages are received or if some are missing and why. We aim to detect some effects which can cause losses of UDP datagrams. The effect of speed and daytime was explored.

Keywords: UDP datagram, GSM network, transmission, railway, current position.

Presenting Author's biography

Zuzana Kleprlíková – Master's in the field of information technology at the University of Pardubice (2004, 2009). Ph.D. student at the University of Pardubice Faculty of Electrical Engineering and Information Technology (since 2009). Main research interests is in software engineering.

David Žák – Master's and Ph.D. in the field of Information Technology and Experimental Physics at the Palacky University in Olomouc (1993, 1998). He has been employed at the University of Pardubice since 2006, lecturing courses in Database Systems, Database Architecture and Techniques of Database Systems. Main research interests are related to intelligent network solutions for mobile communication within railway vehicles and telematics applications.

1 General

The installation of communication terminals and railway cab radios on the selected group of railway vehicles of Czech Railways and CD Cargo has proceeded over the past three years. These devices enable data communication via GPRS in the public GSM network between track vehicles and central information systems. One of the employed applications is the current position of the vehicle. These data is transmitted via a UDP protocol to servers of the Central Communication Gateway of the Railway Wireless Communication Network.
This paper aims to make a proposal for analysing effects which affect the transfer of UDP datagrams in the Railway Wireless Communication Network. Factors influencing the loss of the information have been studied on UDP datagrams used for data transmission of the current position of the railway vehicles.

2 Data communication in RWCN

2.1 RWCN description

RWCN (Railway Wireless Communication Network) is the name of the network used for data communication between mobile objects and infrastructure systems in the railway environment.
This network is comprised of a set of transmission networks, communication devices, interfaces, protocols and rules for wireless communication between users and devices. [1] One of the transmission networks integrated into RWCN is a public GSM network. Detail descriptions of the concepts and principles of data communication through RWCN will be stated in the article.

2.2 Application Current Position of the railway vehicle

The application which we will refer to as "Current Position" of the railway vehicle is installed in the communication terminals of railway vehicles. Position data is sent in fixed time intervals or immediately after change of the vehicle status (like stopping or starting) to stationary information systems. To transmit these reports, the UDP protocol was chosen because of its minimal data transmission bandwidth.

Any message about Current Position contains following mandatory information:

a) message number (cyclically from 0 to 255),
b) number of vehicle in UIC number format,
c) current status of the vehicle,
d) current date and time recorded from a GPS receiver (in UTC format),
e) latitude and longitude,
f) speed,
g) azimuth,
h) network used by car radio (150 MHz TRS, TRS 450 MHz, GSM-R, GSM-P). [2]

2.3 Transmission of UDP datagrams

When the UDP datagram is transmitted through RWCN, these following cases may occur:

a) The datagram is properly sent and delivered to a Central Communication Gateway, where it is subsequently processed.

b) The datagram cannot be posted because of the unavailability of public GSM network. As soon as the communication terminal connects to the GSM network, datagrams carrying the new current position is sent immediately. This message is complemented by a status which signifies, that message was not properly sent in time.

   In the analysis, we are monitoring the main causes of undelivered UDP datagrams. One of the so-called global influences is the area coverage by the GSM signal.

c) Datagrams are sent but not delivered to the server of The Central Communication Gateway.

If the message is not delivered to the server, we can locate it using the following method. From the base of knowledge of the last received and next delivered UDP datagrams, we can determine the number of undelivered (lost) messages and locate the area from which the undelivered messages were sent. We know also other information, such as the time interval, railway vehicle code, the speed of vehicle, etc.

3 Factors influencing the transmission of UDP datagrams

The analysis aims to assess which factors have a statistically significant effect in causing the loss of UDP datagrams. The goal is to identify the parts of Czech Republic or individual railway tracks where there is not sufficient coverage of the GSM signal. We also have to look at the possibility that a communication terminal will have higher losses of sent UDP datagrams than other terminals. These issues can appear due to GSM module trouble, or increased attenuation of an antenna cable, or damage of the antenna. The influencing factor may also be a train speed, time of day (and related data network load) and weather.

Self transmission of UDP datagrams in a real environment is obviously influenced by more factors operating simultaneously, since they cannot be isolated from each other. For each analysis it must always be possible to remove data, which would otherwise devalue the result. Among this data we can place, for example, the non-delivered messages that were sent from the standing vehicle under a metal roof, where there is no (or very attenuated) signal of the public GSM network.

Due to the architecture of the used network solution, some of the monitored UDP datagrams may be lost for these reasons:

a) datagrams are lost between the terminal and the BTS (base transceiver station) of GSM network - probably the most common influence,

b) dropping by network elements (e.g. routers) in the GSM network operator (between the BTS and the central MSC (mobile switching center),

c) dropping by network elements providing connection between RWCN and the public GSM network,

d) dropping by network elements in the LAN network connecting servers of The Central Communication Gateway,

e) datagram is delivered to the server of The Central Communication Gateway, but is not stored in the database (stack overflow of output queue may occurs).

We will analyse the data from several tens of railway vehicles, which together send daily more than 50 000 messages about their current position.

Each message contains information about the order called message number. Data type of message number is the byte defined within the range of 0 to 255. Undelivered messages will be missed in the cyclic sequence.

3.1 Adjustment of data for subsequent processing

Before use of the data, an exploratory analysis should be proceeding. [3] The purpose of this analysis is to identify specialties and to verify the expectation for subsequent mathematical and statistical processing.

From the tested data sample, it is necessary to remove first delivered message which came when the communication terminal is switched on. The first message always has the message number 0 and therefore we will ignore this type of messages for following calculations. The wrong interpretation of this data would create the impression, that many UDP datagrams were undelivered.

On the base of real data we can calculate the probability of delivery of UDP datagrams for all monitored vehicles as

$$R = \frac{\sum_{i=1}^{n}\sum_{a=1}^{m} D_{ia}}{\sum_{i=1}^{n}\sum_{a=1}^{m} S_{ia}} \quad ,$$

where index *i* means the specific communication terminal, index *a* means the specific area, index *n* is the number of all communication terminals and *m* is the number of all areas in the Czech Republic from which we received UDP datagrams.

We can say that

$$S_{ia} = D_{ia} + L_{ia} \ ,$$

where

D_{ia} is the number of delivered messages on the server of The Central Communication Gateway from communication terminal *i* and area *a*,

L_{ia} is the number of lost messages sent from communication terminal *i* and area *a*,

S_{ia} is the total number of sent messages from communication terminal *i* and area *a*.

3.2 Effect of the position

As already indicated in the preceding paragraphs, the position has a significant influence on the probability of delivery of UDP datagrams. In the following text we will refer to R_a as the probability of UDP datagrams delivered from the specific area with index *a*.

For R_a apply that

$$R_a = \frac{\sum_{i=1}^{n} D_{ia}}{\sum_{i=1}^{n} S_{ia}} \quad .$$

In the analysis we are interested in places with $R_a < R$. Likely is going about places with weaker signal strength of the GSM network. We can split the Czech Republic into equally sized areas for which we will calculate R_a. For the subsequent processing we will consider only those areas where

$$\sum_{i=1}^{n} D_{ia} \geq 200 \ . \qquad (1)$$

For the calculation of following effects we will use only data from areas where

$$R_a > 0.99 \ . \qquad (2)$$

3.3 Communication terminal failure

The purpose of the implementation of our analysis is also to identify vehicles that are achieving poorer results in terms of R. To prevent degradation of the results we will consider in further processing only data from areas satisfying the conditions (1) and (2).

For specific communication terminal we can calculate the probability of UDP datagrams delivery as

$$R_i = \frac{\sum_{a=1}^{l} D_{ia}}{\sum_{a=1}^{l} S_{ia}} \quad ,$$

where *l* is a number of areas satisfying conditions (1) and (2).

For the calculation of following effects we will use only data from communication terminals where

$$R_i > 0.99 . \quad (3)$$

3.4 Effect of the daytime

In the study we will analyse the probability R in relation to daytime used as R(t). The aim is to see if this probability is dependent on:

a) the number of delivered messages from all communication terminals, that are sending messages containing the current position,

b) the load of the public GSM network during a day.

Data has to satisfy conditions (1,2,3) and we can write

$$R(t) = \frac{\sum_{i=1}^{k}\sum_{a=1}^{l} D_{ia}(t)}{\sum_{i=1}^{k}\sum_{a=1}^{l} S_{ia}(t)} .$$

This formula calculates the probability of delivery of messages in specified time interval. Variable *t* means the daytime (specific time interval, e.g. 1 hour). Variable *k* means the number of communication terminals satisfying the conditions (1,2,3).

3.5 Effect of the speed

UDP datagrams sent from the train in the most of time are delivered. But sometimes we lost some. We will calculate the probability of delivery of sent messages in a dependency on the speed of railway vehicle. We will describe if the speed is the factor which has main effect on delivery of UDP datagrams.

Next formula calculates the probability of delivery messages in a dependency on the speed

$$R(v) = \frac{\sum_{i=1}^{k}\sum_{a=1}^{l} D_{ia}(v)}{\sum_{i=1}^{k}\sum_{a=1}^{l} S_{ia}(v)} .$$

4 Results

Based on the analysis described in Chapter 3, the acquired data describing the position of 56 sets of trains 471 series was processed. The period of the research was from January to April 2010. Totally, from these railway vehicles in that period was delivered nearly 6 million messages of current train position.

Messages of current position of the railway vehicles were generated according the following rules:

- periodically every 30 seconds or after a track distance of 720 m from the position mentioned in a previous message. This depends on whichever occurs first,
- when the railway vehicle stands, messages are generated periodically every 300 s. Or, if the train speed is less than 5 km/h, messages are generated after a track distance of 100 m from the current position mentioned in the previous report. This depends on whichever occurs first,
- always when the railway vehicle starts and stops,
- when the data or information has changed (e.g. train number).

4.1 Adjustment of data for subsequent processing

In the first step was calculated additional information to all delivered messages in the database - train distance, the time interval (counted from the previous message we have received) and eventually the number of lost messages between current and previously message from the same communication terminal of the railway vehicle.

For subsequent processing, so-called exploratory data analysis was carried out. We were looking for outsider data and suspicious data in the selection. Excluded were position data, which was significantly diverted from normal values, i.e. that has failed at least one of the following conditions:

a) track distance between two consecutive messages coming from the same communication terminal is less than or equal to 10 000 meters,

b) the time interval between two consecutive messages coming from the same communication terminal is less than or equal to 3600 s,

c) the speed of the train indicated by the GPS receiver is less than or equal to 140 km/h (maximum speed of the train of 471 series in the common traffic),

d) GPS position is outside the Czech Republic,

e) parameter message number is smaller than the same parameter for the immediately preceding message from the same communication terminal (this rule limits the potential negative effect on the results in a case, when the communication terminal has been restarted and would be generating message numbers again from 0),

f) less than 10 000 messages came from the communication terminal in that period.

In connection with input data review was detected a communication terminal, which had significantly

less probability of message delivery in comparison with other communication terminals. Based on the evaluation of this data, the need for service of the failed terminal was created.

Parameter	Number of msg.	Msg. ratio
Sum of delivered messages	5 877 913	
Number of vehicles	51	
Sum of losted messages	52 263	0,9%
Count of dropouts	41 744	0,7%
Vehicle status		
- moving	4 486 619	76,3%
- standing	1 391 294	23,7%
Reason of sending message		
- Time exceeting	4 090 897	69,6%
- Distance exceeting	503 021	8,6%
- Start and stop	1 272 837	21,7%

Tab. 1

The table 1 describes data that is entered into a further process after implementation of the above-mentioned conditions. We have large group of data – number of messages is close to 6 millions. Dropout in this concept means one or more consecutive missing UDP datagrams for the same communication terminal.

4.2 Effects of the position

To determine R_a, we have decided to divide the whole country into square areas. The size of elementary area is 100x100 m.

We decided to present results on maps of the Czech Republic. When displaying the map we adjust the size and transparency of the highlighted color area according to the current map scale. In a smaller map scale we use a bigger area such as 200x200, 500x500 or 1000x1000 meters which ensues from a more elementary area of 100x100 m.

Those larger areas can be clearly visualized on the map of the whole Czech Republic, but these large areas are too big for finding specific positions. Therefore, we are progressively reducing the elementary areas according to a zoom level. From areas is evaluated following information:

- the number of messages delivered from the area,
- the number of undelivered messages, following or previous messages delivered from this area,
- the average time period at which delivered messages were sent,
- the number of delivered messages according to statuses (standing, start, stop, motion, other).

For subsequent evaluations only data from areas satisfying the condition (1) were processed. This value is chosen as a base value. It cannot happen that one randomly undelivered message causes the insertion of the area into another (worse) category. Significance level was set at a value of 0.5 (i.e. the probability of message delivery has been not gone down below the specified boundary 99.5% in a case of one lost message).

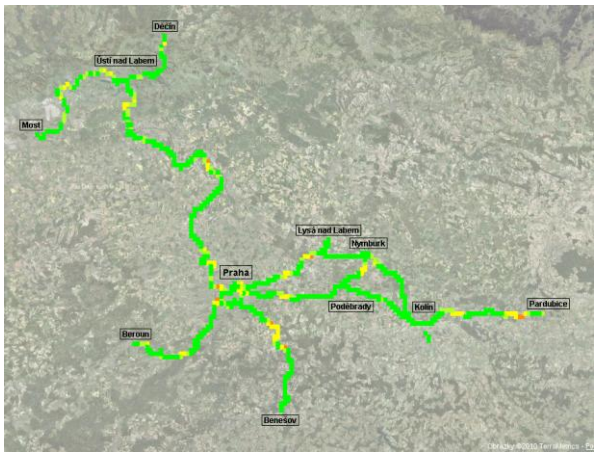

Fig. 1 Map

Figure 1 shows the Czech part of the rail network, on which are moving monitored railway vehicles 471 series. Places highlighted with the yellow colors and red colors can be characterized as an area which has a reduced R_a.

Color	R_a	Ratio areas	Ratio messages
Green	More than 99%	89,6 %	86, 0 %
Light green	Between 99% and 97%	3,9 %	7,2 %
Yellow	Between 97% and 90%	3,6 %	4,9 %
Orange	Between 90% and 80%	1,4 %	1,4 %
Red	Less than 80%	1,4 %	0,5 %

Tab. 2

Table 2 describes colors used in figure 1. The table also shows that from almost 90% of all areas are messages delivered without problems.

4.3 Equipment failure

UDP datagram delivery probability R_i was investigated for railway vehicles, specifically, the communication terminal. On the x-axis of graph Figure 2 is probability of delivery of UDP datagrams R_i, the y-axis shows the number of communication terminals. From the graph is clear that the most of railway vehicles has R_i 99.5%.

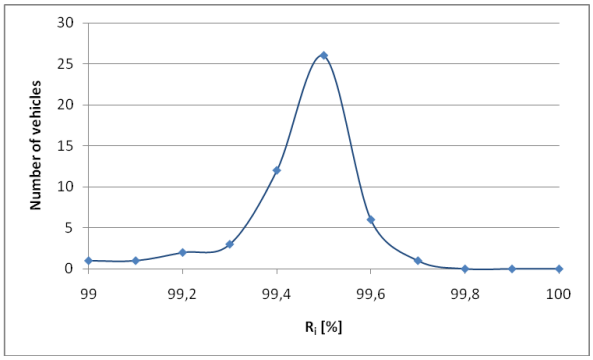

Fig. 2 Dependency of number of communication terminals on R_i

4.4 Effect of the daytime

Effect of daytime on the probability of delivery of UDP datagrams R(t) is monitored on the sample of data from areas that satisfy conditions (1,2,3). Two graphs were generated based on this data - one for weekdays and one for weekends. This was done because these days have different sets of rail traffic and also the usage of the public GSM network operator. See Figure 3 and 4. Data is calculated in 1h intervals. Each time of message delivery is truncated to the closest less whole hour.

In both graphs we can see a relatively balanced R(t) between 3h-21h. R(t) visibly increased during the night.

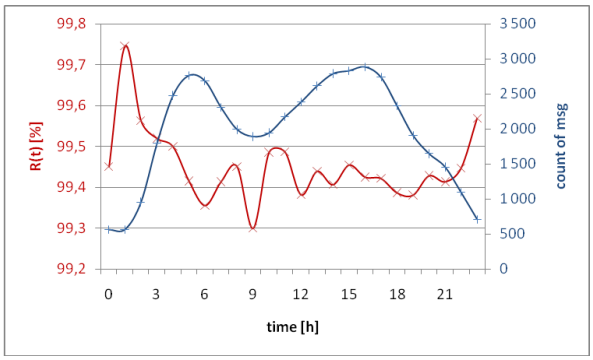

Fig. 3 Effect of the daytime – the average number of delivered messages per hour (blue) and R(t) (red) depending on time of day during the week

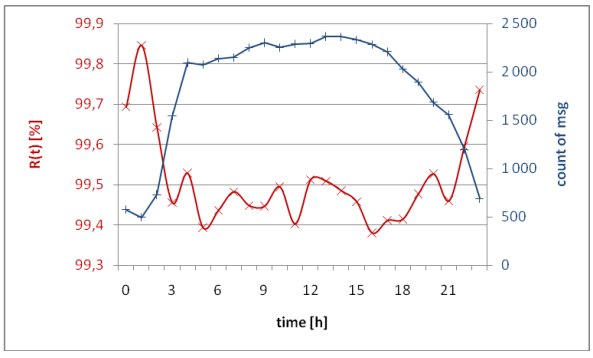

Fig. 4 Effect of the daytime – the average number of delivered messages per hour (blue) and R(t) (red) depending on time of day during the weekend

4.5 Effect of the speed

Over the same data set as in Chapter 4.4, we did the research R(v) based on the current speed of the railway vehicle. The speed is for the purposes of this analysis rounded up to tens of km/h. The probability of delivery of the UDP datagram R(v) is shown in Figure 5 where the average number of daily received messages for specific speed interval is also shown.

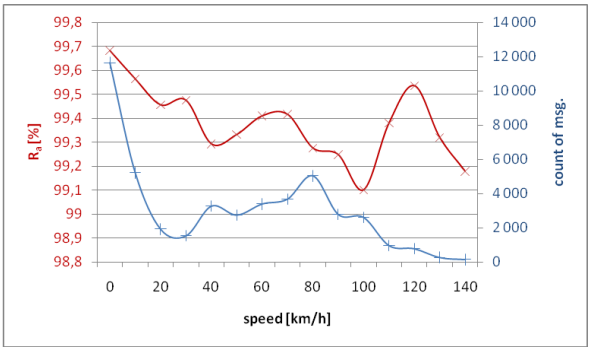

Fig. 5 Effect of the speed - the average number of daily delivered messages (blue) and the R(v) (red) depending on the speed of the train.

From this graph is clear that R(v) does not behave linearly according to the speed. Very interesting is the fact that from the speed of around 100 km/h is a higher probability R(v) than before, and from the speed of around 120 km/h decreases again.

We cannot see from this graph the direct correlation between the speed and delivered messages. Delivering of messages depends on many factors. But very interesting could be find why messages are not delivered and if the main reason could be speed of the train. This problem was investigated with the use of multidimensional statistical methods. Software STATISTICA was selected for multivariate methods for data processing and software solution.

We calculated the correlation matrix of all variables to determine whether there are groups of variables, or if variables are interdependent. Subsequently, the main component analysis and factor analysis will be discussed.

In correlation matrix can appear positive or negative numbers. The relationship between the characters or variables x and y can be positive if the (approximately) relation is y = kx (direct correlation), or a negative relation y = -kx (indirect correlation). In the correlation matrix we found out the coefficient ratio of lost messages to be equal to 0.611. It means that the relationship of speed to the ratio of lost messages indicates a direct dependency. This assumption should be confirmed by following additional analysis.

4.6 PCA Analysis

The aim of this method is the transformation data from the original characters, or variables

$x_j, j = 1,\ldots,m$ into a smaller number of latent variables y_j. These variables are more suitable properties. They are significantly less, describe almost all the variability of characters, and are mutually non-correlated. Correlation coefficients between latent variables $y_i,\ldots,y_m$ are equal to 0. Latent components are therefore key components that describe most of the variability of the data. The main component has a shape as

$$y_1 = \sum_{j=1}^{m} v_{1j} x_j = v_1^T x \quad [3]$$

Object x contains sign $x_1,\ldots,x_m$. For the vector of coefficients, $v_1^T = (v_{11}, v_{12},\ldots, v_{1m})^T$ applies, that the variability in terms of scattering $D(y_1) = v_1^T S v_1$ is the maximum, while S denotes the covariance matrix of the original data.

If the combined share of the variability is sufficiently close to unity, respectively 100% is sufficient to consider only the first components. In our case first 2 components that have eigenvalues greater than 1 and give the total sum of 93.1% are sufficient. Examination was carried out by Cattel graph of eigenvalue, which shows the relative size of each eigenvalue.

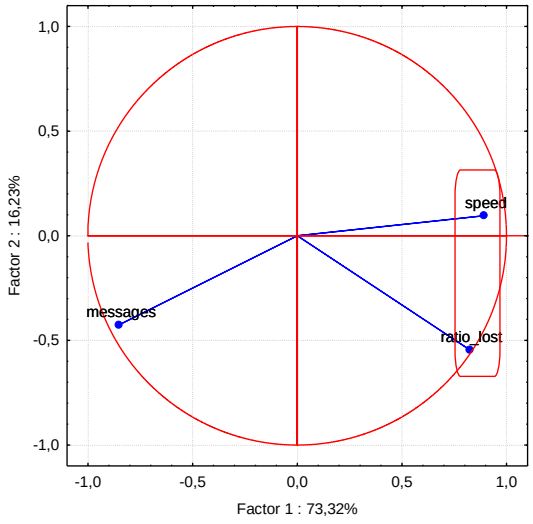

Fig. 6 The graph of main component

The graph shows the characteristics of the component weights and the distance between them. Variable messages means the number of delivered messages for certain speed divided by number of all delivered messages. The variable ratio_lost means the numbers of lost messages for certain speed divided by number of all delivered messages and variable speed is speed of the train. Variables placed close together are in correlation (speed, ratio_lost). The chart shows that the speed and variable ratio_lost are in close correlation. In the case of messages (as shown in Figure 6), a negative correlation exists, because of the angle between the vector is almost 180 degrees.

4.7 Factor analysis

There is also a method that reduces the number of characters. It is called Factor Analysis and this method can sufficiently describe the variability of characters in the data. FA in each character can be expressed as a linear combination of a small number of common factors and one hidden error factor. PCA analysis helped us to find the principal components to determine their dispersion. Factor analysis will try to prove that there is a covariance and correlation between these characters.

Factor analysis model is as follows:

$$x_1 = l_{11} f_1 + l_{12} f_2 + \ldots + l_{1p} f_p + \varepsilon_1$$
$$x_2 = l_{21} f_1 + l_{22} f_2 + \ldots + l_{2p} f_p + \varepsilon_2$$

$$x_m = l_{m1} f_1 + l_{m2} f_2 + \ldots + l_{mp} f_p + \varepsilon_m \quad ,$$

where $f_1, f_2,\ldots,f_p$ are factors that cause correlations between the characters and $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m$ are error vectors, which contribute to the variance of individual characters. Coefficients l_{ik} called factor loadings of the i object in the k common factors f_k, represent matrix elements factorial loads. The model then can be rewritten in matrix form as $x = \mathrm{L}f + \varepsilon$. [4]

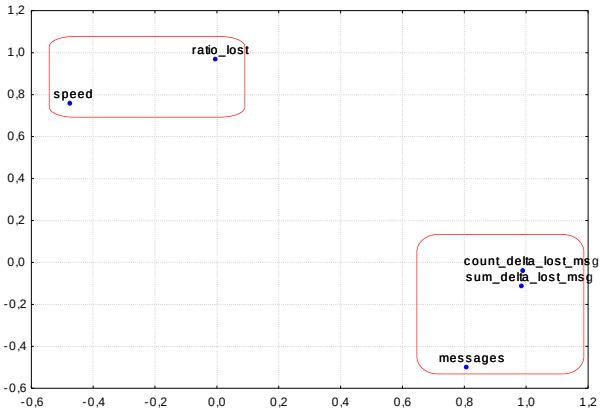

Fig. 7 The graph of Factor Analysis

The graph shows the same variable ratio_lost and speed compared to other variables. Again confirms that there is a correlation between this data.

4.8 Analysis conclusion

From all executed analysis, it can be seen that there is a strong correlation between the variable speed and ratio_lost.

5 Simulation model

The presented research is a part of a wider project aimed at the alternative software support related to the dispatching control of the real-time railway traffic.

The mentioned support is based on the data reflecting the positions of the railway vehicles (obtained from the various systems – e.g. GPS). The special simulation model (reflecting the traffic within the specified part of a railway network) is currently under construction at our workplace. It will be utilised for testing employability of positional data for the dispatching control (e.g. for contra-rides identification of the trains on the single-track lines etc.). The model needs the results presented in this paper in order to be able to simulate transmissions of messages (considering relevant non-deliveries) from the trains to the corresponding information system.

6 Conclusion

The article introduces to readers the factors affecting the quality of transmission of UDP datagrams in RWCN. The results can be used to improve the signal coverage of GSM network in selected areas where there are greater losses of UDP datagrams. It is possible to monitor the operational features of each device and when there occurs the deterioration of these facilities repair can be provided before a failure suspends their operation. Similarly, it is possible to use long-term monitoring of other factors to ensure reliable operation of the entire RWCN in the future, as it will be able to respond to any negative changes in a timely manner.

7 References

[1] ŠÍDLO, M., ŽÁK, D. *Železniční bezdrátová přenosová síť (koncepce, komunikační jednotka, GW).* Sborník 3. konference Moderní zabezpečovací, řídící a telekomunikační technika na tratích ČR jako součásti evropského železničního systému, České Budějovice, 2007, 221-224.

[2] ŽÁK, D., ČEGAN, L. *Možnosti využití aplikace aktuální poloha kolejových vozidel v dopravních systémech*. In Dopravní systémy 2008. Vyd. 1. *Pardubice*: Dopravní fakulta Jana Pernera, Univerzita *Pardubice*, *2008.* Perner´s Contacts, ročník třetí, číslo V., vyšlo 30.12.2008, 308 – 313. ISSN 1801-674

[3] MELOUN, M., MILITKÝ, J. *Statistická analýza experimentálních dat.* ACADEMIA, 3. VYDÁNÍ, 2004. ISBN 80-200-1254-0

[4] MELOUN, M., MILITKÝ, J. *Kompendium statistického zpracování dat.* ACADEMIA, 2. VYDÁNÍ, 2006. ISBN 80-200-1396-2