Mohamed Salahuddin HABIBULLAH, Fu XIUJU
Institute of High Performance Computing, Singapore
Krzysztof KOŁOWROCKI, Joanna SOSZYŃSKA, Beata MILCZEK
Gdynia Maritime University, Poland

CORRELATION AND REGRESSION ANALYSIS OF WINTER EXPERIMENTAL STATISTICAL DATA OF THE OPERATION PROCESS OF THE STENA BALTICA FERRY

Key-words

Operation process, correlation coefficients, single regression, multiple regression.

Summary

These are presented statistical methods of correlation and regression analysis of the operation processes of complex technical systems. The collected statistical data from the Stena Baltica ferry operation process are analysed and used for determining correlation coefficients and single and multiple regression equations, expressing the influence of the operation process conditional sojourn times in particular operation states on the operation process total time.

1. Introduction

Many real transportation systems belong to the class of complex systems. First, and foremost, these systems are concerned with the large numbers of components and subsystems, and they are built and with their operating complexities. Modelling of these complicated system operations processes is primarily difficult because of the large number of the operation states, the

impossibility of their precise definition as well as the impossibility of the exact description of the transitions between these states. Generally, the change of the operation states of the system operations processes causes the changes of these systems reliability structures and their component reliability functions. Therefore, the system operation process and its operation states require proper definition and accurate identification of the interactions between the particular operation states and their influence on the entire system operation process is very important.

The model of the operation processes of the complex technical systems [1] which distinguishes their operation states is proposed in [3]. The semi-Markov process [2] is used to construct a general probabilistic model of the considered complex industrial system operation process. To apply this model in practice, its unknown parameters have to be identified, namely, the vector of the probabilities of the system initial operation states, the matrix of the probabilities of transitions between the operation states, and the matrix of the distribution functions or equivalently, the matrix of the density functions of the conditional sojourn times in the particular operation states. All of which needs to be estimated on the basis of the statistical data. The methods of the evaluation of these unknown parameters are developed and presented in details in [4, 5]. In addition to these methods, the simple data mining techniques, such as correlation coefficient, linear and multiple regression as well as root mean square error can be used on the statistical data samples to perform the analyses. The results of that analysis as well as relevant conclusions that can be reached from the studies may give practically important information in the operation processes of the complex technical systems investigation.

The aim of this paper is to use these techniques in studying the patterns that can be derived from the realisations of the conditional sojourn times, obtained from the Stena Baltica ferry operation process for the winter data [6, 7].

The paper is organised in the following way. In Section 1, the problem that is considered in this report is defined. In Section 2, the general assumptions on the complex system operation process are presented. In Section 3, the Stena Baltica ferry operation process is described. In Section 4, the formulae from the winter data for the total conditional sojourn time and its mean and standard deviation are presented and analysed. This is then followed by the correlation coefficient, linear and multiple regression and root mean square error for analysing the winter data. In Section 5, the paper is concluded.

2. System operation process

We assume, similarly as in [1] and [3], that a system during its operation at the fixed moment t, $t \in <0, +\infty>$, may be in one of v, $v \in N$, different

operations states z_b, b = 1,2, ..., v. Next, we mark by $Z(t)$, $t \in <0, +\infty>$, the system operation process, that is a function of a continuous variable t, taking discrete values in the set $Z = \{ z_1, z_2, ..., z_v \}$ of the operation states. We assume a semi-Markov model [1], [2], [3] of the system operation process $Z(t)$ and we mark by θ_{bl}, its random conditional sojourn times at the operation states z_b, when its next operation state is z_l, $b, l = 1,2,...,v$, $b \neq l$.

Under these assumptions, the operation process may be described by the vector $[p_b(0)]_{1xv}$ of probabilities of the system operation process staying in particular operation states at the initial moment $t = 0$, the matrix $[p_{bl}(t)]_{vxv}$ of the probabilities of the system operation process transitions between the operation states and the matrix $[H_{bl}(t)]_{vxv}$ of the distribution functions of the conditional sojourn times θ_{bl} of the system operation process at the operation states or equivalently by the matrix $[h_{bl}(t)]_{vxv}$ of the density functions of the conditional sojourn times θ_{bl}, $b, l = 1,2,...,v$, $b \neq l$, of the system operation process at the operation states.

To estimate the unknown parameters of the system operations process, the first phase in the experiment is to collect necessary statistical data. After collecting the statistical data, it is possible to estimate the unknown parameters of the system operation process [4], [5]. It is also possible to analyse rather accurately the system operation process sojourn times in the particular operation states and their influence on the entire system operation process total sojourn time [7].

3. Stena Baltica ferry operation process

The problem considered in this paper is based on real maritime statistical data, obtained from Stena Baltica ferry operation process, whereby the ferry performs continuous journeys from Gdynia in Poland to Karlskrona in Sweden. Table 1 shows the operation states that the Stena Baltica ferry undertakes, beginning with loading at Gdynia then passing through the Traffic Separation Scheme to Karlskrona for unloading/loading and back to Gdynia for unloading/loading. This operation process is repeated continuously, and it is assumed that one voyage from Gdynia to Karlskrona and back to Gdynia is a single realisation of its operation process. For the voyage described, time-series data were collected for the realisation of the conditional sojourn times, θ_{bl} of the system operations process at the operation state, z_b when the next transition

is to the operation state, z_l for winter conditions. These data are shown in the Appendix in Tables A5-A8 in [6].

Table 1. Stena Baltica ferry operation states

Operation state	Description	Operation State	Description
z_1	Gdynia: Loading	z_{10}	Karlskrona: Unmooring
z_2	Gdynia: Unmooring	z_{11}	Karlskrona: Turning
z_3	Gdynia: Navigating to GD buoy	z_{12}	Karlskrona: Navigating to Angoring buoy
z_4	Gdynia: Navigating to TSS	z_{13}	Karlskrona: Navigating to TSS
z_5	Gdynia: Navigating to Angoring buoy	z_{14}	Karlskrona: Navigating to GD buoy
z_6	Karlskrona: Navigating to Verko berth	z_{15}	Karlskrona: Navigating to Turning Area
z_7	Karlskrona: Mooring	z_{16}	Gdynia: Ferry Turning
z_8	Karlskrona: Unloading	z_{17}	Gdynia: Mooring
z_9	Karlskrona: Loading	z_{18}	Gdynia: Unloading

It is also important to note that the operation process is very regular and cyclic, in the sense that the operation states changes from the particular state, z_b, where $b = 1,2....17$ to the neighbouring state, z_{b+1}, where $b = 1,2....17$ only and from z_{18} to z_1. Therefore, based on this definition the winter realisation of the conditional sojourn times, $\theta^k_{b\,b+1}$, where $b = 1,2....17$ and $\theta^k_{18\,1}$ for $k = 1,2,...,n_{bl}$, where $n_{bl} = 40,$ are given in Tables A5-A8. Also included in Tables A5-A8 are the values of the total conditional sojourn time for each realisation, θ^k_T, for $k = 1,2,...,n_{bl}$, where $n_{bl} = 40$. In our analyses, the values of θ^k_T are important in analysing the behaviour of the Stena Baltic ferry operation process.

4. Data analysis on Stena Baltica operation process

In this section, the use of several data mining techniques, on the total conditional sojourn time are described. The techniques adopted are, namely, correlation coefficient, single and multiple regression and root mean square error. These techniques are applied on the winter data from the Stena Baltica ferry operation process.

4.1. Total conditional sojourn times

As discussed above, the Stena Baltica ferry data is shown in the Appendix of [6] in Tables A5-A8 for summer. In analysing the behaviour of the data patterns, this paper examines the total conditional sojourn time (the length of time of one ferry voyage) θ_T by analysing its successive realisations θ_T^k, defined as

$$\theta_T^k = \sum_{b=1}^{17} \theta_{b\,b+1}^k + \theta_{18\ 1}^k \tag{1}$$

for $k = 1,2,...,n_{bl}$, where $n_{bl} = 40$ for winter data. Using equation (1), the total conditional sojourn times were then calculated. These values will form the basis of our conjecture in this paper.

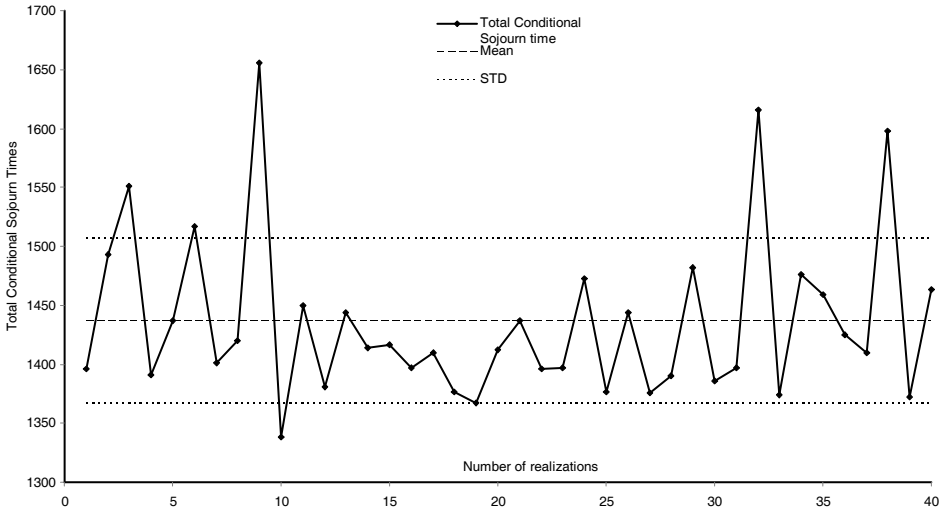

Fig. 1. Plot of realisations θ_T^k of total conditional sojourn time θ_T for winter data

Figure 1 shows the plot of the realisations θ_T^k of the total conditional sojourn time θ_T against the realisation number k for winter data. In the picture, by STD, there are marked 1-sigma lower $\overline{\theta}_T - \overline{\sigma}_T$ and upper $\overline{\theta}_T + \overline{\sigma}_T$ bounds for the total conditional sojourn time θ_T. From the figure, it can be seen that although the ferry operation process is regular and cyclic, *i.e.* the operation states follow the process in Table 1, it can be observed that the values of θ_T are

not constant. Furthermore, by using the mean total conditional sojourn time $\overline{\theta}_T$, evaluated from the following equation

$$\overline{\theta}_T = \frac{1}{n_{bl}} \sum_{k=1}^{n_{bl}} \theta_T^k \tag{2}$$

and the standard deviation defined as

$$\overline{\sigma}_T = \sqrt{\frac{1}{n_{bl}} \sum_{k=1}^{n_{bl}} (\overline{\theta}_T^k - \overline{\theta}_T)^2} \tag{3}$$

it was found that nearly 16% of the θ_T^k values fall outside of the interval $< \overline{\theta}_T - \overline{\sigma}_T, \overline{\theta}_T + \overline{\sigma}_T >$.

The results in Figures 1 seem to indicate a pattern, whereby in each realisation the contribution of the conditional sojourn time θ_{bl}^k for some operation states towards θ_T^k is more for some than that for others. Thus, identifying the conditional sojourn time for such operation states, which has a major effect on the total ferry operation process times, would enable the total conditional sojourn time for the operation process to be studied, analysed, and predicted. These are discussed in the following sections where the use of data mining techniques, to understand the behaviour of θ_T^k, are presented.

4.2. Correlation

Correlation analysis is a method commonly used to establish, with a certain degree of probability, whether a linear relationship exists between two measured quantities. This means that, when there is correlation, it implies that there is a tendency for the values of the two quantities to effect one another. Vice-versa also holds true, if there is no correlation, which implies no effect on each other. Furthermore, using the values of the correlation coefficient, a positive or negative relationship can also be identified. If the coefficient values are close to 1, it implies a positive linear relationship, whilst values close to 0 imply no linear relationship. Thus, based on the values of the correlation coefficient, the relationship between two measured quantities can be determined. The adopted formula for evaluating the correlation coefficient r_{bl} between the conditional sojourn time θ_{bl} and the total conditional sojourn time θ_T is given by

$$r_{bl} = \frac{\sum_{k=1}^{n_{bl}} (\theta_{bl}^k - \bar{\theta}_{bl})(\theta_T^k - \bar{\theta}_T)}{\sqrt{\sum_{k=1}^{n_{bl}} (\theta_{bl}^k - \bar{\theta}_{bl})^2} \sqrt{\sum_{k=1}^{n_{bl}} (\theta_T^k - \bar{\theta}_T)^2}} \tag{4}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, where $n_{bl} = 40$ is the number of realisations, θ_{bl}^k is the *k*-th realization of the conditional sojourn time θ_{bl}, θ_T^k is the *k*-th realisation of the total conditional sojourn time θ_T evaluated from (1), $\bar{\theta}_T$ is the mean total conditional sojourn time evaluated from the equation (2) and $\bar{\theta}_{bl}$ is the mean conditional sojourn time obtained from

$$\bar{\theta}_{bl} = \frac{1}{n_{bl}} \sum_{k=1}^{n_{bl}} \theta_{bl}^k \tag{5}$$

Thus, using the values from Tables A5-A8, the correlation coefficient, r_{bl}, were then evaluated using equation (4). Table 2 shows the values of r_{bl} for the winter data.

Table 2. Correlation coefficient r_{bl} values for winter data

Operation State	**Correlation coefficient**	**Operation state**	**Correlation coefficient**
z_1	0.155212	z_{10}	0.129131
z_2	-0.03004	z_{11}	0.205063
z_3	0.317888	z_{12}	0.279811
z_4	0.294948	z_{13}	0.524968
z_5	0.656862	z_{14}	0.253331
z_6	0.319895	z_{15}	-0.01054
z_7	0.10485	z_{16}	-0.05677
z_8	0.166899	z_{17}	-0.02651
z_9	0.461141	z_{18}	0.406053

Figure 2 shows the plot of the correlation coefficient r_{bl} against the number b of the operation state z_b. It can be seen that $\theta_{4\,5}$, $\theta_{5\,6}$ and $\theta_{13\,14}$ has the strongest positive linear relationship, as compared to the conditional sojourn times in the remaining operation states, where $\theta_{5\,6}$ and $\theta_{13\,14}$ coincides with the longest parts of the voyage. This implies that any variations in the conditional sojourn times $\theta_{b\,b+1}$ associated with these 3 operation states, namely z_4, z_5 and z_{13}, will significantly effect the total conditional sojourn time θ_T.

The plots given in Figure 2 also shows that most of the r_{bl} values are more than 0, which seems to indicate a positive linear relationship, albeit a weak linear relationship for some. Thus, from the correlation coefficient values, it can be deduced that the values of the total conditional sojourn time θ_T are strongly dependent on the conditional sojourn times $\theta_{b\,l}$ for some operation states. In the following section, this understanding of the data behaviour will be used in the regression model to predict the values of the total conditional sojourn time θ_T.

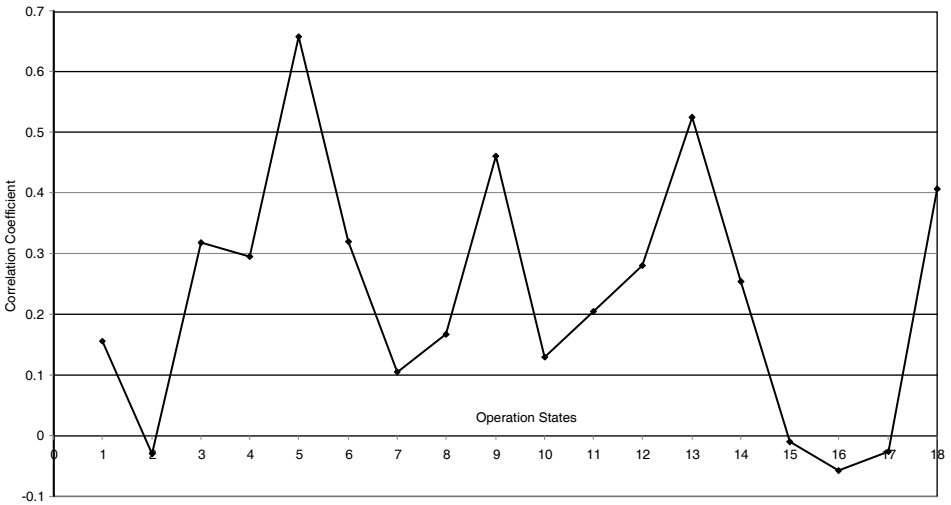

Fig. 2. Plot of correlation coefficient r_{bl} between conditional sojourn time and total conditional sojourn time for winter data

4.3. Regression

Regression analysis is a data mining technique used in modelling, analysing and predicting numerical data. In a single regression, input statistical data are necessary, whereby the data is modelled as a function, in coming out with the

model parameters. These parameters are then estimated so as to give a "best fit" of the data, which are then used to predict future data behaviour. Multiple regression is another type of a single regression model. It is similar to a single regression, but in this model the interest is on examining more than one predictor variable. In this technique, the aim is to determine whether the inclusion of additional predictor variables leads to an increased prediction of the outcome. Here, the use of both single and multiple regression models on the winter data are described.

From the above discussions, it can be seen that the aim of using the single regression technique is to use initial sample data of the conditional sojourn times θ_b to predict the subsequent behaviour of the total conditional sojourn time θ_T. In this report, the equation adopted is given by

$$\theta_T = \alpha_b + \beta_b \theta_{b\,l} + \varepsilon_b \tag{6}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, $l = 1$, where α_b, β_b are the unknown regression coefficients and ε_b is the random noise.

Before predicting the subsequent behaviour, the values of α_b and β_b, based on varying realisations of the operation process, need to be evaluated. The unknown regression coefficients α_b and β_b are evaluated by minimising the functions

$$\Delta(\alpha_b, \beta_b) = \sum_{k=1}^{N} [\theta_T^k - (\alpha_b + \beta_b \theta_{b\,l}^k)]^2 \tag{7}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, defined as the measure of divergences between the empirical values θ_T^k and defined by (6) the predicted values $\theta_T(\theta_{bl}^k) = \alpha_b + \beta_b \theta_{b\,l}^k$ of the total conditional sojourn time θ_T.

From the necessary condition, *i.e.* after finding the first partial derivatives of $\Delta(\alpha_b, \beta_b)$ with respect to α_b and β_b and putting them equal to zero, we get the system of equalities involving the realisations θ_T^k of the total conditional sojourn time θ_T and the realisations θ_{bl}^k of the conditional sojourn times $\theta_{b\,l}$ defined as follows

$$N\alpha_b + \sum_{k=1}^{N} \theta_{b\,l}^{k} \beta_b = \sum_{k=1}^{N} \theta_T^k \tag{8}$$
$$\sum_{k=1}^{N} \theta_{bl}^{k} \alpha_b + \sum_{k=1}^{N} (\theta_{b\,l}^{k})^2 \beta_b = \sum_{k=1}^{N} \theta_{bl}^{k} \theta_T^k$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, $l = 1$ and $N = 1,2,..., n_{bl}$.

The remaining question that needs to be addressed is how many realisations marked by N does it take to obtain a reasonable representation of α_b and β_b. By using Matlab and putting the values from Tables A5-A8 [6] into the system of equations (8), the varying α_b and β_b values were calculated for $N = 1,2,..., n_{bl}$.

Figure 3 shows the plot of the regression coefficient β_b against N, for the operation states of z_5 and z_{13}. From the discussions in Section 4.2, these 2 operation states represent the longest part of the voyage and has major influence on the total conditional sojourn time. From the plot, it can be observed that other than the initial instability for low values of N, the values of β_b seems to stabilise for larger N. In our analyses, it was discovered that the value of β_b stabilises at $N = 30$. Although not shown in this paper, this behaviour also holds true for all the other operation states.

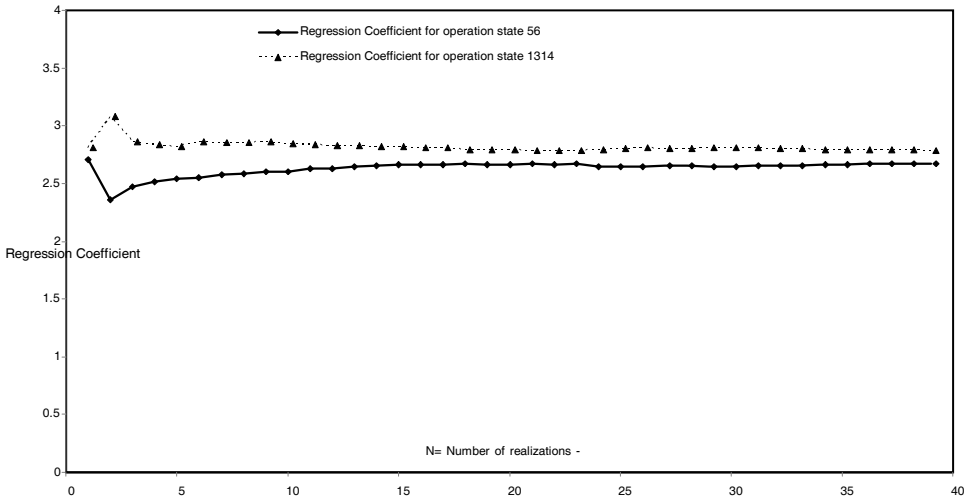

Fig. 3. Plot of regression coefficient β_b for winter data

Thus, based on the above observations, the predicted total conditional sojourn times, θ_T^*, can then be evaluated using β_b values at $N = 30$. In evaluating θ_T, the formulation in the system of equations (8), leading to

$$\theta_T^* = \alpha_b^* + \beta_b^* \theta_{b\,l} \tag{9}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, $l = 1$, where α_b^* and β_b^* are, respectively, the value of α_b and β_b at $N = 30$.

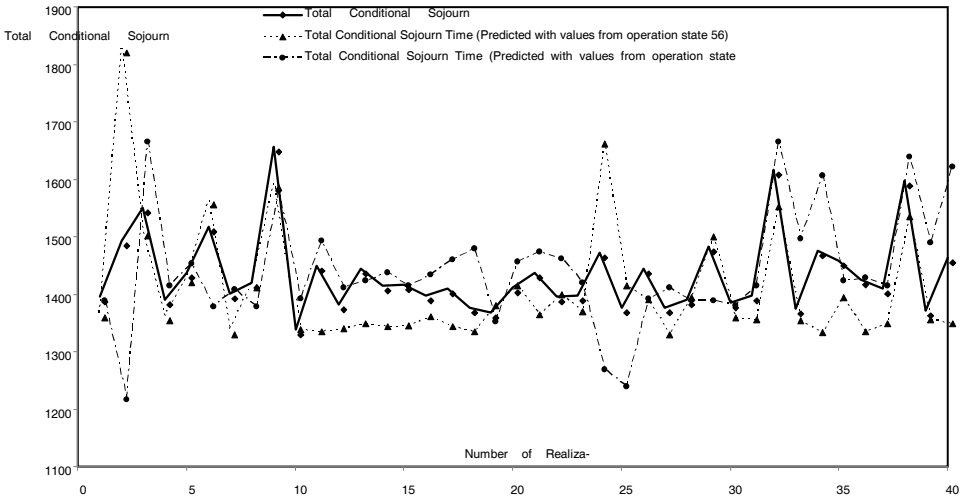

Fig. 4. Plots of empirical realisations and predicted from single regression values of total conditional sojourn time for winter data

Figure 4 shows the comparison plots of the values of the empirical realisations θ_T^k of the total conditional sojourn time θ_T and the predicted values θ_{T*}^k of the total conditional sojourn time θ_T^* defined by the equation (9) against the number of realisations k for winter data. It can be observed that for both the operation states of z_5 and z_{13}, the predicted θ_{T*}^k values are not close to the empirical θ_T^k values. Similar patterns of behaviour were also observed when the values of θ_{T*}^k for other operation states, were considered. These results seem to indicate that linear regression does not provide an accurate means of predicting the behaviour of the Stena Baltica ferry process.

Since the single regression does not provide an accurate prediction of the total conditional sojourn time, the multiple regression technique will be explored instead. As described earlier, the difference in the multiple regression technique is that in this method, more than one predictor variable is considered. It is envisaged that the inclusion of additional predictor variables will lead to an increased prediction of the total conditional sojourn time. Thus, for multiple regressions, the equation adopted is given by

$$\theta_T = \alpha_b + \sum_{b=1}^{B} \beta_b \theta_{b\,l} + \varepsilon_b \tag{10}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, $l = 1$ and $B = 1,2,...,\nu$, $\nu = 18$, where α_b, β_1, β_2, …, β_B are the unknown regression coefficients and ε_b is the random noise.

Before predicting the subsequent behaviour of α_b, β_1, β_2, …, β_B values based on varying realisations of the operation process need to be evaluated. The unknown regression coefficients α_b, β_1, β_2, …, β_B is obtained by minimising the functions,

$$\Delta(\alpha_B, \beta_1, \beta_2, ..., \beta_B) = \sum_{k=1}^{N} [\theta_T^k - (\alpha_B + \sum_{b=1}^{B} \beta_b \theta_{b\,l}^k)]^2 \tag{11}$$

for $b = 1,2,...,17$, $l = b+1$ and $b = 18$, $l = 1$ and $B = 1,2,...,\nu$, $\nu = 18$, that is the measure of divergences between the empirical values θ_T^k and predicted values $\theta_T(\theta_{1l}^k, \theta_{2l}^k, ..., \theta_{Bl}^k) = \alpha_B + \sum_{b=1}^{B} \beta_b \theta_{b\,l}^k$ of the total conditional sojourn time θ_T defined by (10).

From the necessary condition, *i.e.* after finding the first partial derivatives of $\Delta(\alpha_B, \beta_1, \beta_2, ..., \beta_B)$ with respect to α_b, β_1, β_2, …, β_B and putting them equal to zero, we get the system of equalities involving the realisations θ_T^k of the total conditional sojourn time θ_T and the realisations θ_{bl}^k of the conditional sojourn times $\theta_{b\,l}$ defined as follows:

$$N\alpha_B + \sum_{b=1}^{B} \sum_{k=1}^{N} \theta_{b\,l}^k \beta_b = \sum_{k=1}^{N} \theta_T^k \tag{12}$$

$$\sum_{k=1}^{N}\theta_{1\,l}^{k}\alpha_B+\sum_{b=1}^{B}\sum_{k=1}^{N}\theta_{1\,l}^{k}\theta_{b\,l}^{k}\beta_b=\sum_{k=1}^{N}\theta_{1\,l}^{k}\theta_T^{k}$$

$$\ldots\ldots\ldots$$

$$\sum_{k=1}^{N}\theta_{B\,l}^{k}\alpha_B+\sum_{b=1}^{B}\sum_{k=1}^{N}\theta_{B\,l}^{k}\theta_{b\,l}^{k}\beta_b=\sum_{k=1}^{N}\theta_{B\,l}^{k}\theta_T^{k}$$

for $b=1,2,...,17$, $l=b+1$ and $b=18$, $l=1$ and $B=1,2,...,\nu$, $\nu=18$ and $N=1,2,...,n_{bl}$.

The remaining question that needs to be addressed here is that how many realisations marked by N in (12) does it take to obtain a reasonable representation of α_b, β_1, β_2, …, β_B. By using Matlab and putting the values from Tables A5-A8 [6] into the system of equations (12) for $N=1,2,...,n_{bl}$, the varying α_b, β_1, β_2, …, β_B values were calculated.

In our analyses on the values of α_b, β_1, β_2, …, β_B, the observation is that the values of α_b, β_1, β_2, …, β_B stabilises at $N=30$. It was also observed that α_b, β_1, β_2, …, β_B vary with respect to the number B, $B=1,2,...,\nu$, $\nu=18$, of predictor variables considered changing 1 to 18. The argument for this method is that, by using more than one predictor variable, better results will be obtained. The aim is also to use a minimal number of predictor variables to generate accurate results, within a short period of time. Thus, based on the above observations, the predicted total conditional sojourn time, θ_T, can then be evaluated using α_b, β_1, β_2, …, β_B values at $N=30$. In evaluating θ_T, the formulation in the system of equations (10), leading to

$$\theta_T^{*}=\alpha_B^{*}+\sum_{b=1}^{B}\beta_b^{*}\theta_{b\,l} \tag{13}$$

for $b=1,2,...,17$, $l=b+1$ and $b=18$, $l=1$ and $B=1,2,...,\nu$, $\nu=18$, where α_B^{*}, β_1^{*}, β_2^{*}, ..., β_B^{*} are respectively the value of α_b, β_1, β_2, …, β_B at $N=30$.

Figure 5 shows the comparison plots of the values of the empirical realisations θ_T^{k} of the total conditional sojourn time θ_T and the predicted values θ_{T*}^{k} of the total conditional sojourn time θ_T^{*} defined by the equation (11)

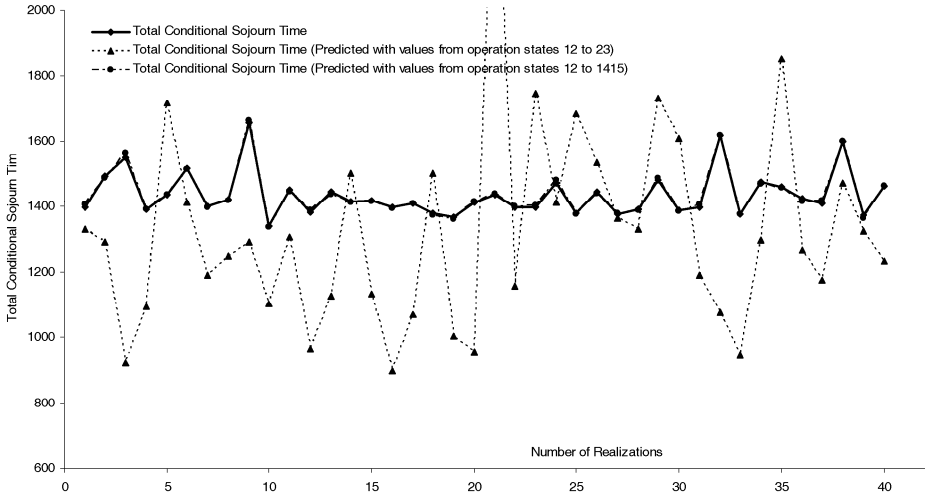

Fig. 5. Plots of empirical realisations and predicted from multiple regression values of total conditional sojourn times θ_T^k and θ_{T*}^k for winter data

against the number of realisations k for winter data. It can be seen that, if only 2 predictor variables, θ_{12} and θ_{23} $(B=2)$ are used in the equation (11), then the predicted values differ much from the empirical values θ_{T*}^k and are not accurate at all. It was discovered that, as we increased the number of predictor variables, the accuracy improves, leading to the best accuracy at $B=14$ predictor variables θ_{12}, θ_{23}, …, θ_{1415}. It was also observed that, if more than 14 predictor variables were used, the results do not change much, indicating that 14 predictor variables provide a good representation of the prediction. The analyses also show that multiple regression is a better method of predicting the behaviour of the Stena Baltica ferry data than the single regression.

4.4. Accuracy

To further access the accuracy of the predicted data, the root mean square error ε is applied. The root mean square error is commonly used to calculate the error and is often used to measure the success of numerical prediction. If the value of ε is 0, it simply means that there is no error to the prediction and the prediction is accurate. The greater values of ε imply that the more inaccurate is the prediction. In the paper, the values of the root mean square errors for both the single and multiple regressions are calculated. The root mean square error equation adopted is given by

$$\varepsilon = \sqrt{\frac{1}{n_{bl}} \sum_{k=1}^{n_{bl}} (\theta_T^{*k} - \theta_T^k)^2} \tag{14}$$

where $\theta_T^{*k} = \theta_T^*(\theta_{bl}^k)$ for single regression, $\theta_T^{*k} = \theta_T^*(\theta_{1l}^k, \theta_{2l}^k, ..., \theta_{Bl}^k)$ for multiple regression and $n_{bl} = 40$ in the case of winter data. By using the predicted values θ_T^{*k} for both single and multiple regressions and the empirical value of θ_T^k from the winter data, the values of ε were calculated. It was found for winter data that, for instance, for a single regression with one predictor variable θ_{56} that $\varepsilon \cong 80.5$ and for multiple regression with 14 predictor variables θ_{12}, θ_{23}, …, θ_{1415} this value was $\varepsilon \cong 5.5$. These values of the root mean square errors validate the results obtained from the regression analyses, indicating the accuracy of multiple regressions as compared to single regression.

Acknowledgements

The paper describes part of the work in the Poland-Singapore Joint Research Project titled "Safety and Reliability of Complex Industrial Systems and Processes" supported by grants from the Poland's Ministry of Science and Higher Education (MSHE grant No. 63/N-Singapore/2007/0) and the Agency for Science, Technology and Research of Singapore (A*STAR SERC grant No. 072 1340050).

Conclusions

This paper has described the use of simple data mining techniques on the Stena Baltica ferry operation process statistical data. The aim was to observe the behaviour of the ferry's total conditional sojourn time and use it to predict future behaviours. In our analyses, we applied the correlation coefficient, single and multiple regressions, and root mean square error on the winter data. From the results, it can be concluded that the use of multiple regression technique on the data provides an accurate way of predicting the ferry's total conditional sojourn time.

References

1. Blokus-Roszkowska A., Guze S., Kołowrocki K., Kwiatuszewska-Sarnecka B., Soszyńska J.: Models of safety, reliability, and availability evaluation of complex technical systems related to their operation processes. WP 4 – Task 4.1 – English – 31.05.2008. Poland-Singapore Joint Project, Gdynia, 2008.

2. Grabski F.: Semi-Markov Models of Systems Reliability and Operations. Monograph. Analysis. Monograph. System Research Institute, Polish Academy of Science, (*in Polish*), Warsaw, 2002.
3. Kołowrocki K., Soszyńska J.: A general model of technical systems operation processes related to their environment and infrastructure. WP 2 – Task 2.1 – English – 31.05.2008. Poland-Singapore Joint Project, Gdynia, 2008.
4. Kołowrocki K., Soszyńska J.: Methods and algorithms for evaluating unknown parameters of operation processes of complex systems. Proc. Summer Safety and Reliability Seminars -SSARS 2009, Vol. 2, 211–221.
5. Kołowrocki K., Soszyńska J.: Data mining for identification and prediction of safety and reliability characteristics of complex systems and processes. Proc. European Safety and Reliability Conference – ESREL 2009, Vol. 2, 853–863.
6. Kołowrocki K., Soszyńska J., Kamiński P., Jurdziński M., Guze S., Milczek B., Golik P.: Data mining for identification and prediction of safety and reliability characteristics of complex industrial systems and processes.WP6 – Task 6.2. Preliminary statistical data collection of the Stena Baltica ferry operation process and its preliminary statistical identification. WP6 – Sub-Task 6.2.5 – Appendix 5A – English – 31.10.2009. Poland-Singapore Joint Project, Gdynia, 2009.
7. Kołowrocki K., Soszyńska J., Salahuddin Habibullah M., Xiuju F.: Data mining for identification and prediction of safety and reliability characteristics of complex industrial systems and processes.WP6 – Task 6.1.3. Experimental statistical data correlation and regression analysis – Correlation and regression analysis of experimental statistical data of the operation process of the Stena Baltica ferry. Task 6.1.3 – Section 4 and Section 5.5.4 – English – 31.08.2009. Poland-Singapore Joint Project, Gdynia-Singapore, 2009.

Reviewer:
Janusz SZPYTKO

Analiza korelaji i regresji zimowych danych statystycznych procesu eksploatacji promu Stena Baltica

Słowa kluczowe

Proces eksploatacji, współczynnik korelacji, regresja jednowymiarowa, regresja wielowymiarowa.

Streszczenie

Przedstawione są statystyczne metody analizy regresji i korelacji procesów eksploatacji złożonych systemów technicznych. Zebrane dane statystyczne z procesu eksploatacji promu Stena Baltica zostały zanalizowane i użyte do wyznaczenia współczynnika korelacji oraz do wyznaczenia równań jednokrotnej i wielokrotnej regresji, wyrażającej wpływ warunkowych czasów przebywania procesu eksploatacji w poszczególnych stanach eksploatacyjnych na całkowity czas procesu eksploatacji.