CORRELATION AND REGRESSION ANALYSIS OF SPRING STATISTICAL DATA OF MARITIME FERRY OPERATION PROCESS

M.S. Habibullah, Fu Xiuju

•

Institute of High Performance Computing, Singapore

e-mail: mohdsh@ihpc.a-star.edu.sg

K. Kolowrocki, J. Soszynska

•

Gdynia Maritime University, Gdynia, Poland

e-mail: katmatkk@am.gdynia.pl, joannas@am.gdynia.pl

ABSTRACT

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 linear and multiple regression equations, expressing the influence of the operation process conditional sojourn times in particular operation states on the ferry operation process total conditional sojourn 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 they are built and with their operating complexities. Modeling of these complicated system operation processes is, first of all, difficult because of the large number of the operation states, 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 components reliability functions. Therefore, the system operation process and its operation states 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 (Blokus et al. 2008) with distinguishes their operation states is proposed in (Kolowrocki & Soszynska 2008). The semi-markov process (Grabski 2002) 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, needs to be estimated on the basis of the statistical data. The methods of these unknown parameters evaluation are developed and presented in details in (Kolowrocki & Soszynska 2009A-B). 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 report is to use these techniques in studying the patterns that can be derived, from realizations of the conditional sojourn times, obtained from the Stena Baltica ferry operation process, for the early spring data (Kolowrocki et al. 2009A-B).
The report is organized is the following way. In Section 1, some general comments on complex technical systems operation processes modeling are given and the problem 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 for the total conditional sojourn time its mean and standard deviation are presented and applied to the spring statistical data of the Stena Baltica ferry operation process. This is then followed by determining the correlation coefficient, linear and multiple regression and root mean square error for the ferry operation process spring data. In Section 5, the report summary is given.

2 SYSTEM OPERATION PROCESS

We assume, similarly as in (Blokus et al. 2008, Kolowrocki & Soszynska 2008), 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 (Blokus et al. 2008, Grabski 2002, Kolowrocki & Soszynska 2008) 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 $[p_b(0)]_{1\times v}$ of probabilities of the system operation process staying in particular operations states at the initial moment t = 0, the matrix $[p_{bl}(t)]_{v\times v}$ of the probabilities of the system operation process transitions between the operation states and the matrix $[H_{bl}(t)]_{v\times v}$ 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 $[h_{bl}(t)]_{v\times v}$ 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. This is performed in the following steps (Kolowrocki et al. 2009A-B):

i) To analyze the system operation process and either to fix or to define the following general parameters:
   - the number of the operation states of the system operation process v;
   - the operation states of the system operation process z_1, z_2, …, z_v;
ii) To fix and collect the following statistical data necessary in evaluating the probabilities of the initial states of the system operations process:
   - the duration time of the experiment Θ;
   - the number of the investigated (observed) realizations of the system operation process $n(0)$;
   - the numbers of staying operation process respectively in the operations states z_1, z_2, …, z_v, at the initial moment $t = 0$ of all $n(0)$ observed realizations of the system operation process $n_1(0)$, $n_2(0)$, …, $n_v(0)$, where $n_1(0)$ + $n_2(0)$+ $n_v(0)$ = $n(0)$;

iii) To fix and collect the following statistical data necessary to evaluating the transient probabilities between the system operation states:
- the numbers n_{bl}, $b, l = 1,2,...,v$, $b \neq l$, of the transitions of the system operation process from the operation state z_b to the operation state z_l during all observed realizations of the system operation process;
- the numbers n_b, $b = 1,2,...,v$, of departures of the system operation process from the operation states z_b, where $n_b = \sum_{l=1}^{v} n_{bl}$;

iv) To fix and collect the following statistical data necessary in evaluating the unknown parameters of the distributions of the conditional sojourn times of the system operation process in the particular operation states:
- the realizations θ_{bl}^{k}, $k = 1,2, \ldots, n_{bl}$, $b, l = 1,2,...,v$, $b \neq l$, 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 during the observation time;

After collecting the above statistical data it is possible to estimate the unknown parameters of the system operation process (Kolowrocki & Soszynska 2009A-B). It is also possible to analyze 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 (Kolowrocki et al. 2009B).

3 STENA BALTICA FERRY OPERATION PROCESS

The problem considered in this report 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 Kalskrona in Sweden. Table 1 show 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 Kalskrona and back to Gdynia is a single realization of its operation process. For the voyage described, time-series data were collected for the realization 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 spring conditions. These data are shown in the Appendix in Tables A1-A4 coming from (Kolowrocki et al. 2009B).

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

Operation state	Description	Operation State	Description
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 spring realization of the ferry 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 $n_{bl} = 42,$ are given in Tables A1-A4. Also included in Tables A1-A4 are the values of the total conditional sojourn times for each realization, θ^k_T, for $k = 1,2,...,n_{bl}$, where $n_{bl} = 42$. In our analysis the values of θ^k_T are important in analyzing 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 system total conditional sojourn time is described. The techniques adopted are namely, correlation coefficient, linear and multiple regression and root mean square error. These techniques are applied on the early spring data from the Stena Baltica ferry operation process.

4.1 Total conditional sojourn time

As discussed above, the Stena Baltica ferry operation process data for spring is shown in the Appendix in Tables A1-A4 for spring. In analyzing the behavior of the data patterns, this report examines the ferry total conditional sojourn time (the time length of one ferry voyage) θ_T by analyzing its successive realizations θ^k_T, defined as

$$\theta^k_T = \sum_{b=1}^{17} \theta^k_{b\,b+1} + \theta^k_{18\;1} \tag{1}$$

for $k = 1,2,...,n_{bl}$, where $n_{bl} = 42$ for spring data. Using equation (1), the total conditional sojourn times were then calculated for both spring with the values shown in Tables A1-A4. These values form the basis of our conjecture in this paper.

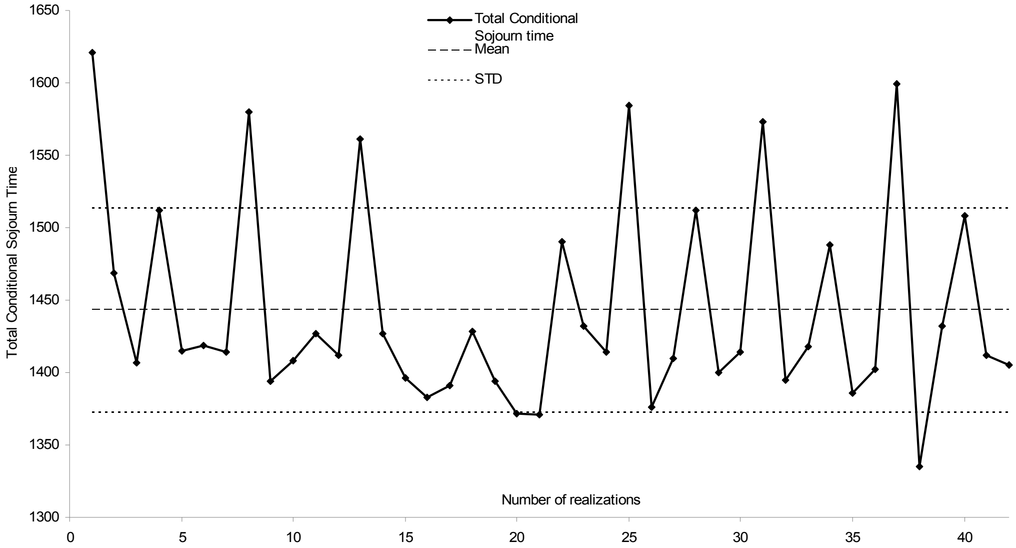

Figure 1. Plot of realizations θ_T^k of total conditional sojourn time θ_T for spring data

Figure 1 shows the plot of the realizations θ_T^k of the ferry total conditional sojourn time θ_T against the realization number k for spring 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 ferry total conditional sojourn time θ_T.

Although the ferry operation process is regular and cyclic, i.e. the operation states follows 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

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

and the standard deviation defined as

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

it was found that nearly 26% 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 realization the contribution of the ferry conditional sojourn time $\theta_{b\,l}^k$ 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 major effect on the ferry total operation process times 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 is presented.

4.2 Correlation

Correlation analysis is a method commonly used to establish, with 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 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 ferry conditional sojourn time θ_{bl} in particular operation states and the ferry 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 $n_{bl} = 42$ is the number of realizations, θ_{bl}^k is the *k*-th realization of the conditional sojourn time θ_{bl}, θ_T^k is the *k*-th realization 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 A1-A4, the correlation coefficient, r_{bl}, were then evaluated using equation (4). Table 2 shows the values of $r_{b\,l}$ for the spring data.

Table 2. Correlation coefficient $r_{b\,l}$ values for spring data

Operation State	Correlation coefficient	Operation state	Correlation coefficient
z_1	0.221169	z_{10}	0.401463
z_2	0.298071	z_{11}	0.324054
z_3	-0.13934	z_{12}	0.306238
z_4	0.642635	z_{13}	0.640848
z_5	0.738339	z_{14}	0.365648
z_6	0.020627	z_{15}	0.099242
z_7	-0.04948	z_{16}	0.142937
z_8	0.2035	z_{17}	0.149159
z_9	0.1559	z_{18}	0.057029

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.

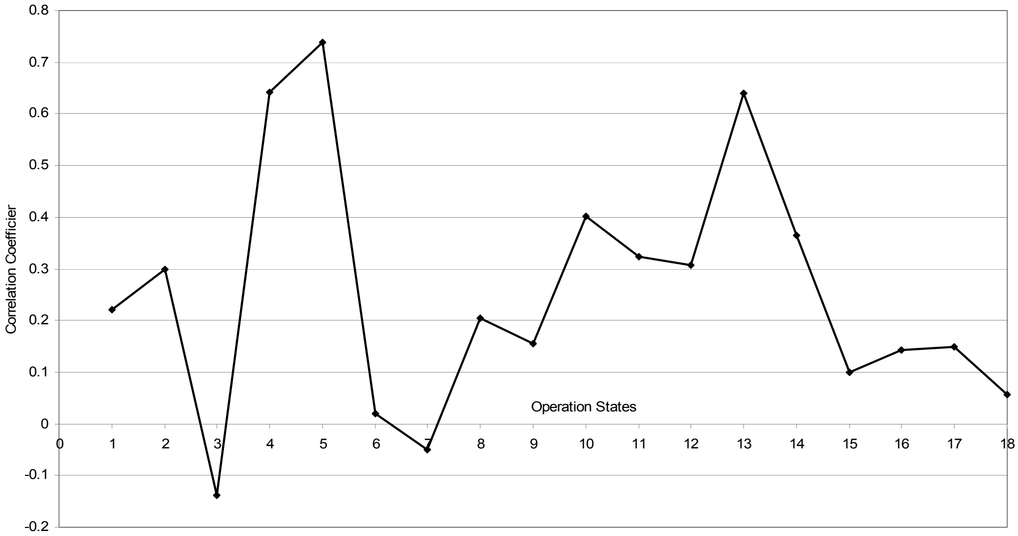

Figure 2. Plot of correlation coefficient r_{bl} between conditional sojourn time and total conditional sojourn time for spring data

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 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 is 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.

4.3 Regression

Regression analysis is a data mining technique used in modeling, analyzing and predicting numerical data. In linear regression, input statistical data are necessary, whereby the data is modeled 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 regression model. It is similar to linear regression but in this model the interest is on examining more than one predictor variables. In this technique the aim is to determine whether the inclusion of additional predictor variables leads to increased prediction of the outcome. Here, the use of both linear and multiple regression models on the spring data are described.

From the above discussions, it can be seen that the aim of using the linear regression technique is to use initial sample data of the conditional sojourn times θ_b to predict subsequent behavior of the total conditional sojourn time θ_T. In the paper 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 $\alpha_b,\ \beta_b$ are the unknown regression coefficients and ε_b is the random noise.

Before predicting the subsequent behavior, the values of α_b and β_b based on varying realizations of the operation process need to be evaluated. Here, the unknown regression coefficients α_b and β_b are evaluated by minimizing 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 realizations θ_T^k of the total conditional sojourn time θ_T and the realizations θ_{bl}^k of the conditional sojourn times $\theta_{b\,l}$ defined as follows

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

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 realizations 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 A1-A4 into the system of equations (6), 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 represents among 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 stabilize for larger N. In our analyses, it was discovered that the value of β_b stabilizes at $N = 30$. Although not shown in the paper this behavior also holds true for all the other operation states.

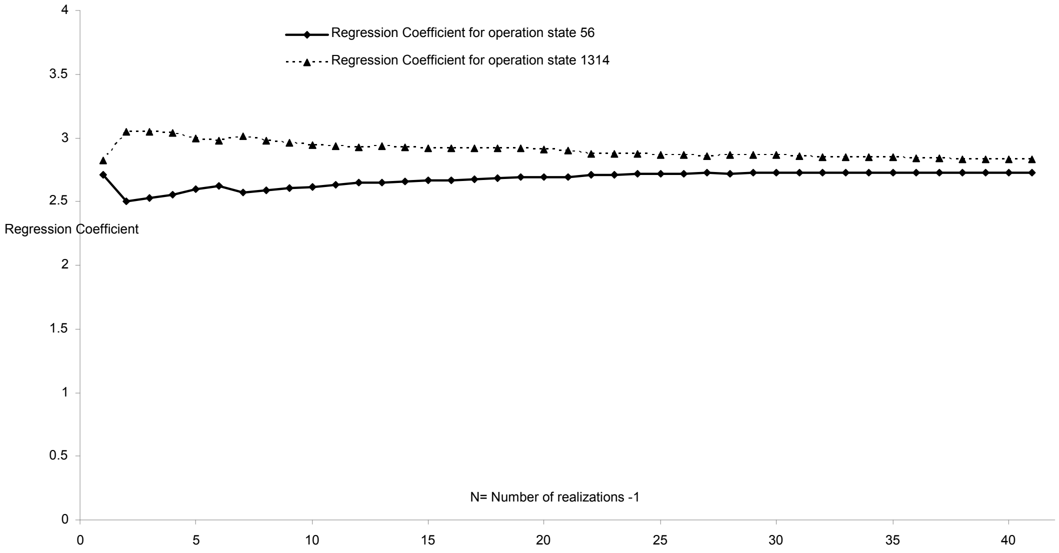

Figure 3. Plot of regression coefficient β_b for spring data

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

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

(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 $N = 30$.

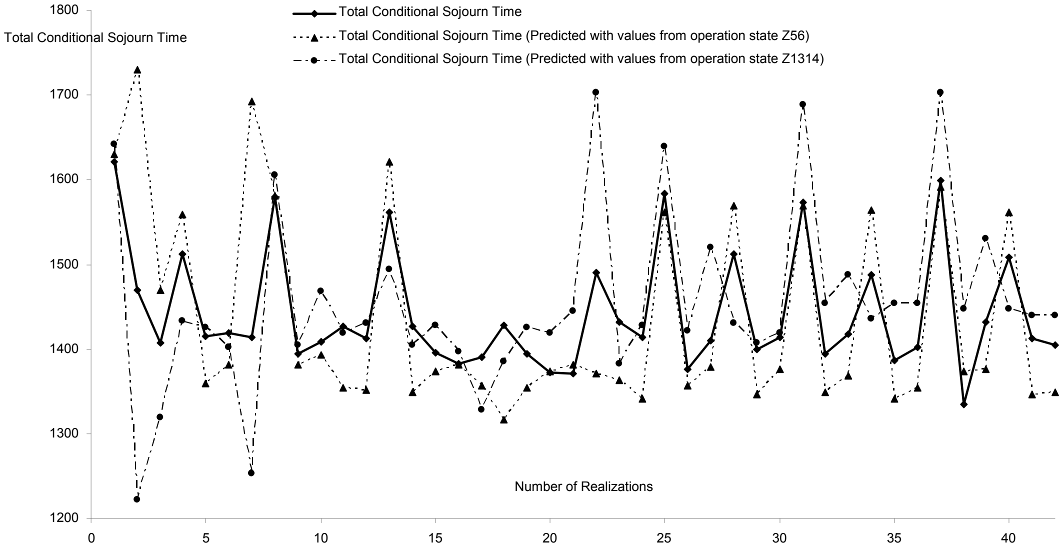

Figure 4. Plots of empirical realizations and predicted from linear regression values of total conditional sojourn time for spring data

Figure 4 shows the comparison plots of the values of the empirical realizations 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 realizations k for summer 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 pattern of behaviour were also seen 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 operation process.

Since linear regression does not provide an accurate prediction of the total conditional sojourn time the multiple regression technique is explored instead. As described earlier, the difference in the multiple regressions technique is that in this method, more than one predictor variables are considered. It is envisaged that the inclusion of additional predictor variables will lead to 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 realizations of the operation process need to be evaluated. The unknown regression coefficients α_B, β_1, β_2, …, β_B are obtained by minimizing 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 $\theta_T(\theta_{1\,l}^k, \theta_{2\,l}^k, ..., \theta_{B\,l}^k) = \alpha_B + \sum_{b=1}^{B} \beta_b \theta_{b\,l}^k$ of the total conditional sojourn time θ_T defined by (8).

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 realizations θ_T^k of the total conditional sojourn time θ_T and the realizations θ_{bl}^k of the conditional sojourn times $\theta_{b\,l}$ defined as follows,

$$\begin{gathered} N\alpha_B + \sum_{b=1}^{B}\sum_{k=1}^{N} \theta_{b\,l}^k \beta_b = \sum_{k=1}^{N} \theta_T^k \\ \sum_{k=1}^{N} \theta_{1\,l}^k \alpha_B + \sum_{b=1}^{B}\sum_{k=1}^{N} \theta_{1\,l}^k \theta_{b\,l}^k \beta_b = \sum_{k=1}^{N} \theta_{1\,l}^k \theta_T^k \\ \text{……..} \end{gathered} \tag{12}$$

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

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 realizations marked by N in (12) does it take to obtain a reasonable representation of α_B, β_1, β_2, …, β_B. Thus, by using Matlab and putting the values from Tables A1-A4 into the system of equations (10) 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 stabilizes 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 variables, better results will be obtained. The aim is also to use as minimal number of predictor variables to generate accurate results, within as 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 formulations in the system of equations (10) lead to

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

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 $N = 30$.

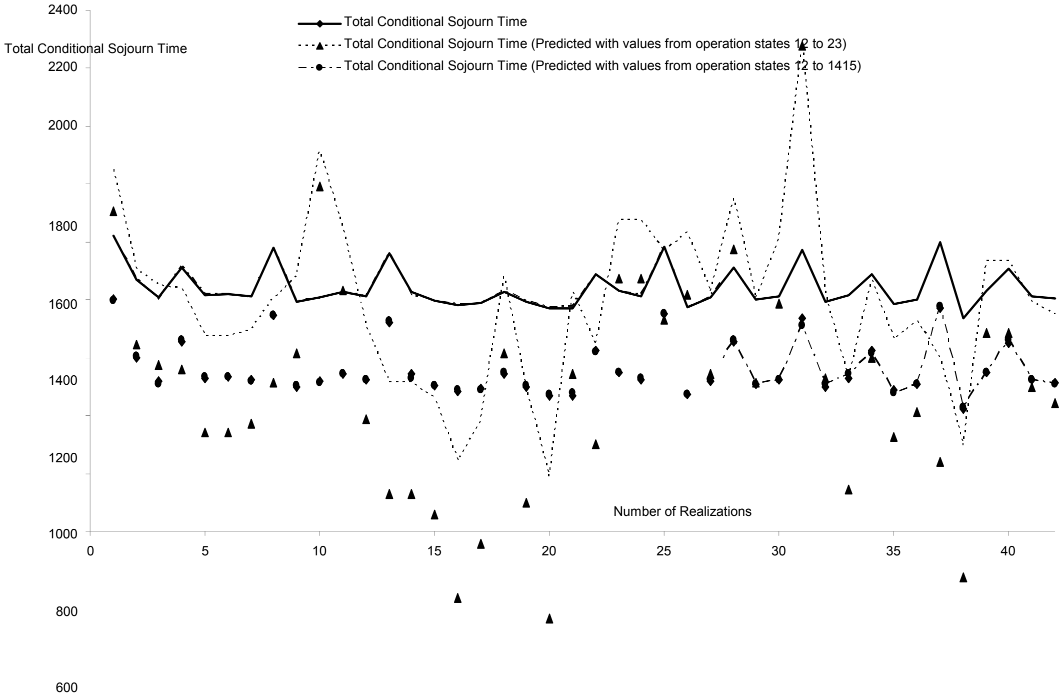

Figure 5. Plots of empirical realizations and predicted from multiple regression values of total conditional sojourn times θ_T^k and θ_{T*}^k for spring data

Figure 5 shows the comparison plots of the values of the empirical realizations 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) against the number of realizations k for summer 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 doesn't change much, indicating that 14 predictors variables provides 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 linear 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 mean that the more inaccurate is the prediction. Here, the values of the root mean square errors for both the linear and multiple regressions are calculated. The adopted for the root mean square error equation is given by

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

where $\theta_T^{*k} = \theta_T^*(\theta_{b\,l}^k)$ for linear regression, $\theta_T^{*k} = \theta_T^*(\theta_{1\,l}^k, \theta_{2\,l}^k, ..., \theta_{B\,l}^k)$ for multiple regression and $n_{bl} = 42$ in the case of spring data. By using the predicted values θ_T^{*k} for both linear and multiple regressions and the empirical value of θ_T^k from the spring data the values of ε were calculated. It was found for spring data that for instance for linear regression with one predictor variable θ_{56} that $\varepsilon \cong 77.9$ and for multiple regression with 14 predictor variables θ_{12}, θ_{23}, …, θ_{1415} this value was $\varepsilon \cong 5.3$. These values of the the root mean square errors validate the results obtained from the regression analyses, indicating the accuracy of multiple regressions as compared to linear regression.

5 SUMMARY

This report has described the use of simple data mining techniques on the Stena Baltica ferry operation process statistical data given in Tables A1-A4. The aim is to observe the behaviour of the ferry operation process total conditional sojourn time and use it to predict future behaviours. In our analyses, we applied the correlation coefficient, linear and multiple regressions and root mean square error on spring data. From the results, it can be concluded that multiple regressions technique provides an accurate of predicting the ferry total conditional sojourn time.

8 REFERENCES

Blokus-Roszkowska, A., Guze, S., Kołowrocki, K., Kwiatuszewska-Sarnecka, B., Soszyńska, J. 2008. *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.

Grabski, F. 2002. *Semi-Markov Models of Systems Reliability and Operations*. Monograph. Analysis. Monograph. System Research Institute, Polish Academy of Science, (in Polish).

Kołowrocki, K., Soszyńska, J. 2008. *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.

Kołowrocki, K., Soszyńska, J. 2009A. *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.

Kołowrocki, K., Soszyńska, J. 2009B. *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.

Kołowrocki, K., Soszyńska, J., Kamiński, P., Jurdziński, M., Guze, S., Milczek, B., Golik, P. 2009A. 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.

Kołowrocki, K., Soszyńska, J., Salahuddin Habibullah, M., Xiuju, F. 2009B. *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.

Appendix

Statistical summer data collection of the Stena Baltica ferry operation process

In the *Tables A1-A4* there are given realizations of the conditional sojourn times in particular operation states on the basis of a sample composed of $n = 42$ realizations of the Stena Baltica ferry operation process. It is assumed that one voyage from Gdynia to Kalskrone and back to Gdynia of the ferry is a single realization of its operation process. The conditional sojourn times in particular operation states of each single realization of the ferry operation process are given in separate columns. The operation process is very regular in the sense that the operation state changes are from the particular state z_b, $b = 1,2,...,17,$ to the neighboring state z_{b+1}, $b = 1,2,...,17,$ only and from z_{18} to z_1. Therefore the realizations of the conditional sojourn times θ^j_{bb+1}, $b = 1,2,...,17,$ $j = 1,2,...,42,$ are given in the Tables b-th row and the realizations of the conditional sojourn time $\theta^j_{18\,1}$, $b = 1,2,...,17,$ are given in the Tables 18-th row.

Appendix 5A

5A. 1. Statistical summer data collection of the Stena Baltica ferry operation process

Date/2008	24/25 Jan	26/27 Jan	27/28 Jan	11/12 Feb	12/13 Feb	26/27 Feb	27/28 Feb	28/01 Mar	01/02 Mar	02/03 Mar	11/12 Mar	12/13 Mar

In the *Tables A1-A4* there are given realizations of the conditional sojourn times in particular operation states on the basis of a sample composed of $n = 42$ realizations of the Stena Baltica ferry operation process. It is assumed that one voyage from Gdynia to Kalskrone and back to Gdynia of the ferry is a single realization of its operation process. The conditional sojourn times in particular operation states of each single realization of the ferry operation process are given in separate columns. The operation process is very regular in the sense that the operation state changes are from the particular state z_b, $b = 1,2,...,17,$ to the neighboring state z_{b+1}, $b = 1,2,...,17,$ only and from z_{18} to z_1. Therefore the realizations of the conditional sojourn times θ^j_{bb+1}, $b = 1,2,...,17,$ $j = 1,2,...,42,$ are given in the Tables b-th row and the realizations of the conditional sojourn time $\theta^j_{18\,1}$, $b = 1,2,...,17,$ are given in the Tables 18-th row.

Table A1: Realization of conditional sojourn times in operations states (early spring)

Realization number k	1	2	3	4	5	6	7	8	9	10	11	12
	Realization of conditional sojourn times in operations states (in minutes)											
Operation state z_b	θ_{bb+1}^{1}	θ_{bb+1}^{2}	θ_{bb+1}^{3}	θ_{bb+1}^{4}	θ_{bb+1}^{5}	θ_{bb+1}^{6}	θ_{bb+1}^{7}	θ_{bb+1}^{8}	θ_{bb+1}^{9}	θ_{bb+1}^{10}	θ_{bb+1}^{11}	θ_{bb+1}^{12}
z_1	55	52	47	75	60	60	62	43	50	61	65	63
z_2	4	3	3	2	2	2	2	3	3	4	3	2
z_3	28	31	32	35	37	48	33	38	39	43	40	42
z_4	52	46	48	65	53	47	49	62	45	46	51	47
z_5	598	635	539	572	499	507	621	580	507	511	497	496
z_6	35	42	42	44	35	37	34	40	36	33	38	38
z_7	7	9	8	7	7	5	5	5	5	5	8	7
z_8	25	20	23	27	20	31	15	17	16	21	33	34
z_9	75	59	56	40	66	47	26	60	65	25	55	40
z_{10}	5	3	2	3	2	3	5	6	3	4	4	2
z_{11}	6	5	4	5	4	5	4	4	4	6	4	5
z_{12}	25	22	25	25	23	25	20	33	24	24	22	22
z_{13}	574	427	461	501	498	490	438	561	491	513	496	500
z_{14}	61	43	43	46	49	52	42	63	46	60	50	50
z_{15}	33	32	33	36	35	33	35	34	31	33	34	36
z_{16}	4	4	5	4	4	4	3	4	4	4	4	4
z_{17}	8	10	6	5	5	6	4	5	8	7	6	7
z_{18}	26	26	30	20	16	17	16	22	17	8	17	17
Total θ_T^k	1621	1469	1407	1512	1415	1419	1414	1580	1394	1408	1427	1412

Table A2: Realization of conditional sojourn times in operations states (early spring)

Date/2008	13/15 Mar	15/16 Mar	16/17 Mar	17/18 Mar	18/19 Mar	19/20 Mar	20/21 Mar	21/22 Mar	22/23 Mar	23/24 Mar	08/09 Apr	09/10 Apr
Realization number k	13	14	15	16	17	18	19	20	21	22	23	24
	Realization of conditional sojourn times in operations states (in minutes)											
Operation state z_b	θ_{bb+1}^{13}	θ_{bb+1}^{14}	θ_{bb+1}^{15}	θ_{bb+1}^{16}	θ_{bb+1}^{17}	θ_{bb+1}^{18}	θ_{bb+1}^{19}	θ_{bb+1}^{20}	θ_{bb+1}^{21}	θ_{bb+1}^{22}	θ_{bb+1}^{23}	θ_{bb+1}^{24}
z_1	45	45	40	20	33	50	43	15	45	57	97	68
z_2	2	2	2	2	2	3	2	2	3	2	2	3
z_3	35	36	36	36	37	35	34	34	36	36	39	36
z_4	51	51	51	49	53	44	51	52	50	53	53	54
z_5	595	495	504	507	498	483	497	504	507	503	500	492
z_6	34	39	38	39	38	35	37	36	37	34	38	40
z_7	7	8	7	10	8	8	7	8	8	8	7	9
z_8	18	16	13	3	15	6	9	25	19	31	30	35
z_9	75	77	60	73	82	118	71	55	30	24	34	41
z_{10}	5	2	2	2	3	4	2	2	3	3	2	5
z_{11}	4	4	4	4	4	4	4	4	4	4	4	4
z_{12}	24	24	25	24	23	22	23	22	22	22	26	22
z_{13}	522	491	499	488	464	484	498	496	505	595	483	499
z_{14}	72	50	48	50	48	52	47	53	51	61	61	48
z_{15}	34	35	35	34	35	34	31	32	33	46	34	34
z_{16}	5	5	5	4	4	4	5	5	3	4	6	6
z_{17}	7	7	6	4	4	7	5	5	7	5	4	5
z_{18}	26	40	21	34	40	35	28	22	8	2	12	13
Total θ_T^k	1561	1427	1396	1383	1391	1428	1394	1372	1371	1490	1432	1414

Table A3: Realization of conditional sojourn times in operations states (early spring)

Date/2008	10/12 Apr	12/13 Apr	13/14 Apr	14/15 Apr	15/16 Apr	16/17 Apr	18/19 Apr	19/20 Apr	20/21 Apr	05/06 May	06/07 May	07/08 May
Realization numbr k	25	26	27	28	29	30	31	32	33	34	35	36
	Realization of conditional sojourn times in operations states (in minutes)											
Operation state z_b	θ_{bb+1}^{25}	θ_{bb+1}^{26}	θ_{bb+1}^{27}	θ_{bb+1}^{28}	θ_{bb+1}^{29}	θ_{bb+1}^{30}	θ_{bb+1}^{31}	θ_{bb+1}^{32}	θ_{bb+1}^{33}	θ_{bb+1}^{34}	θ_{bb+1}^{35}	θ_{bb+1}^{36}
z_1	58	35	45	75	72	62	37	44	46	78	59	65
z_2	3	4	3	3	2	3	6	3	2	2	2	2
z_3	37	36	35	39	37	36	37	36	36	37	36	36
z_4	67	51	50	62	49	48	64	51	53	63	55	53
z_5	573	498	506	576	494	505	576	495	502	574	492	497
z_6	36	37	35	38	38	36	35	39	37	36	38	37
z_7	8	7	5	7	10	9	10	6	7	7	6	6
z_8	25	11	17	31	23	25	23	15	18	19	18	24
z_9	55	55	43	45	52	48	50	58	53	30	30	45
z_{10}	3	3	3	3	2	3	2	2	3	3	2	2
z_{11}	4	4	5	5	4	5	4	5	4	5	4	4
z_{12}	23	22	23	26	23	23	24	23	24	23	28	24
z_{13}	573	497	531	500	492	496	590	508	520	502	508	508
z_{14}	58	51	54	47	40	51	47	47	56	47	46	42
z_{15}	34	35	33	35	35	34	33	34	35	36	35	35
z_{16}	5	5	6	5	4	6	5	5	4	4	5	4
z_{17}	4	5	5	5	7	6	5	6	6	10	5	4
z_{18}	18	20	11	10	16	18	25	18	12	12	17	14
Total θ_T^k	1584	1376	1410	1512	1400	1414	1573	1395	1418	1488	1386	1402

Table A4: Realization of conditional sojourn times in operations states (early spring)

Date/2008	08/09 May	10/11 May	11/12 May	12/13 May	13/14 May	14/15 May						
Realization number k	37	38	39	40	41	42						
	Realization of conditional sojourn times in operations states (in minutes)											
Operation state z_b	θ_{bb+1}^{37}	θ_{bb+1}^{38}	θ_{bb+1}^{39}	θ_{bb+1}^{40}	θ_{bb+1}^{41}	θ_{bb+1}^{42}						
z_1	53	25	55	84	71	67						
z_2	2	2	3	2	2	2						
z_3	38	37	40	36	37	34						
z_4	60	49	46	57	53	51						
z_5	584	504	505	573	494	495						
z_6	38	35	36	39	36	36						
z_7	5	7	5	5	6	6						
z_8	15	6	40	28	32	28						
z_9	70	35	35	47	40	50						
z_{10}	2	2	3	3	3	2						
z_{11}	5	4	5	5	4	4						
z_{12}	25	25	24	23	26	24						
z_{13}	595	506	535	506	503	503						
z_{14}	42	45	47	46	51	43						
z_{15}	34	35	34	34	33	33						
z_{16}	6	4	4	5	5	4						
z_{17}	5	3	4	5	3	5						
z_{18}	20	11	11	10	13	18						
Total θ_T^k	1599	1335	1432	1508	1412	1405						