Detecting ASCs Differing in Performance

Olli Kirsi

Master's thesis
December 2023
Master's Degree Programme in Artificial Intelligence and Data analytics

Jyväskylän ammattikorkeakoulu
University of Applied Sciences

Kirsi, Olli

Detecting ASCs Differing in Performance

Jyväskylä: Jamk University of Applied Sciences. December 2023, 78+22 pages.

Degree Programme in Artificial Intelligence and Data analytics. Master's thesis.

Permission for open access publication: Yes

Language of publication: English

Abstract

Containerized trade has grown steadily for decades. Increasing trade volumes required highly productive and advanced port cranes. The focus shifted towards autonomous ASCs that performed most of the activities without human interference. The objective of the thesis was to detect poorly performing cranes out of the good ones using artificial intelligence. The research method was research-based development.

Most of the KPIs for port cranes measure the performance of the cranes by also including many external factors outside control of the machine suppliers. This presented a challenge to machine suppliers on how to monitor the crane performance. To solve the challenge, a new KPI, move residual, was proposed for the ASCs.

Move residual negated the effects of external factors that are present in other KPIs. The move residual was calculated by predicting the cycle times of the container moves with linear regression. The linear model was trained with only normal and uneventful moves. Domain knowledge and statistical methods were utilized to filter out the bad moves from the training data. The predicted cycle time was then deducted from the actual cycle time. As a result, the predicted cycle time was compared to a predicted cycle time. In conclusion, the performance of ASC, with external factors excluded, can be monitored with the move residual. Move residual was used as a quantitative performance measurement and anomaly score simultaneously.

Keywords/tags (subjects)

anomaly score, artificial intelligence, automated stacking cranes, containers, container terminals, cranes, key performance indicator, machinery, machine learning, ports, regression analysis, rail mounted gantry crane

Miscellaneous (Confidential information)

Appendices 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, and 19 are confidential and removed from the public thesis. The basis for secrecy is section 24(17), section 24(20), and section 24(21) of the Act on the Openness of Government Activities (621/1999), a company's business or trade secret. The period of secrecy is five (5) years, the secrecy will end on 10 December 2028.

Kirsi, Olli

Suorituskyvyltään huonommin toimivien ASC-nostureiden tunnistaminen

Jyväskylä: Jyväskylän ammattikorkeakoulu. Joulukuu 2023, 78+22 sivua.

Data-analytiikan ja tekoälyn koulutusohjelma. Opinnäytetyö ylempi AMK.

Julkaisulupa avoimessa verkossa: Kyllä

Julkaisun kieli: Englanti

Tiivistelmä

Konttikuljetusten määrä on kasvanut vuosikymmenten ajan. Kasvanut kuljetusten määrä tarvitsi tehokkaita ja edistyneitä satamanostureita. Painotus on siirtynyt autonomisiin ASC-nostureihin, jotka voivat suorittaa suurimman osan konttien siirroista ilman ihmisen apua. Opinnäytetyön tavoitteena oli tunnistaa huonommin toimivat nosturit paremmin toimivista tekoälyä hyödyntäen. Tutkimusmenetelmä oli tutkimuksellinen kehitystyö.

Useimmat ASC-nostureiden suorituskykymittarit olivat riippuvaisia monista ulkoisista tekijöistä, joihin konetoimittajat eivät voineet vaikuttaa. Tämä toi haasteita suorituskyvyn mittaamiseen konetoimittajien näkökulmasta. Ratkaisuksi kehitettiin tekoälyn avulla ASC-nostureille uusi suorituskykymittari, siirtoajan jäännösarvo, jonka avulla voitiin jättää huomioimatta konttien siirtoaikaan vaikuttavat ulkoiset tekijät.

Lineaarinen regressiomalli laskettiin lineaarisen regression avulla. Regressiomalli koulutettiin vain normaaleilla ja onnistuneilla siirroilla. Substanssiosaamista ja statistiikkaa käytettiin onnistuneiden siirtojen tunnistamiseen huonoista. Siirtoajan jäännösarvo laskettiin lineaarisella regressiolla ennustamalla siirtoihin tarvittava aika ja vähentämällä sitten ennustettu aika toteutuneesta siirtoajasta. Lopputuloksena ennustettua siirtoaikaa verrattiin toteutuneeseen siirtoaikaan. Johtopäätökseksi tuli, että ASC-nostureiden suorituskykyä voidaan mitata siirtoajan jäännösarvolla, jolloin saadaan ulkoiset siirtoaikaan vaikuttavat tekijät erotettua suorituskykymittauksesta. Siirtoajan jäännösarvoa käytettiin samanaikaisesti sekä kvantitatiivisena suorituskykymittarina, että poikkeamien tunnistamisessa.

Avainsanat (asiasanat)

automaattiset pinoamisnosturit, merikontit, tavaraterminaalit, nosturit, suorituskykymittaristot, tekoäly, koneoppiminen, regressioanalyysi

Muut tiedot (salassa pidettävät liitteet)

Liitteet 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 ja 19 ovat salassa pidettäviä, ja ne on poistettu julkisesta työstä. Salassapidon peruste on Julkisuuslain 621/1999 24§ kohdat 17, 20 ja 21, yrityksen liike- tai ammattisalaisuus. Salassapitoaika on viisi (5) vuotta, salassapito päättyy 10.12.2028

Contents

Figures

Tables

1 Introduction

Port cranes are high technology equipment for moving intermodal containers at ports. This research focuses on the ASCs. The level of automation on the ASCs has increased in the recent years. There is also an increasing interest and demand for fully automated ports. McKinsey & Company conducted a survey to executives and managers of leading practitioners in the top port's companies. This survey clearly showed that the port automation has become a trend (Chu et al., 2018). There are already numerous cranes operating independently. The crane operators merely monitor fully automated cranes. Monitoring takes place frequently on remote operating stations, located at a distance from the crane operating area. Subsequently, increased level of automation requires more instrumentation and software. This makes the cranes more sophisticated but also more prone to malfunctions.

Automation increases the crane production compared to a traditional manually operated crane (Meester, 2022). A good crane operator can reach the production of an automated crane. However, compared to a manual crane, the automated crane does not require breaks. Additionally, the performance is predictable. The automated cranes also require less maintenance. The moves are more fluid, precise and less aggressive. As the ports worldwide are struggling to keep up with demand for greater volumes of cargo, expanding the port capacity with more modern automated equipment is one of the key components to tackle the issue (Meester, 2022). Smart port market is projected to triple in just four years. An example of a smart port is ATI Batangas container terminal at Philippines. According to Meester, equipment uptime has improved and with help of data the average moves per hour has increased by 3,1 (2022).

Global containerized trade has been steadily growing for decades. The only exceptions to the growth have been the financial crisis in 2008 and covid-19 pandemic as can be observed from Figure 1. In total, the containerized trade has quadrupled in three decades. The trade volume is projected to grow further (S&P Global Market Intelligence, 2022). Approximately 80 % of global trade volume is shipped by sea (*Navigating Stormy Waters*, 2022, p. xv). Increase in volume requires more capacity through ports. More capacity requires more equipment and space at ports. The number and capacity has increased and will increase in the future (Stahlbock & Voβ, 2008, p. 552). Port capacity investments, however, are large and expensive (Balliauw, 2021). Modification in layout design and equipment utilization maximization is normally the cornerstone of any capacity expansion project (Golbabaie et al., 2012). Therefore, ports have an increasing interest in having the most production out of their existing and recently purchased equipment.

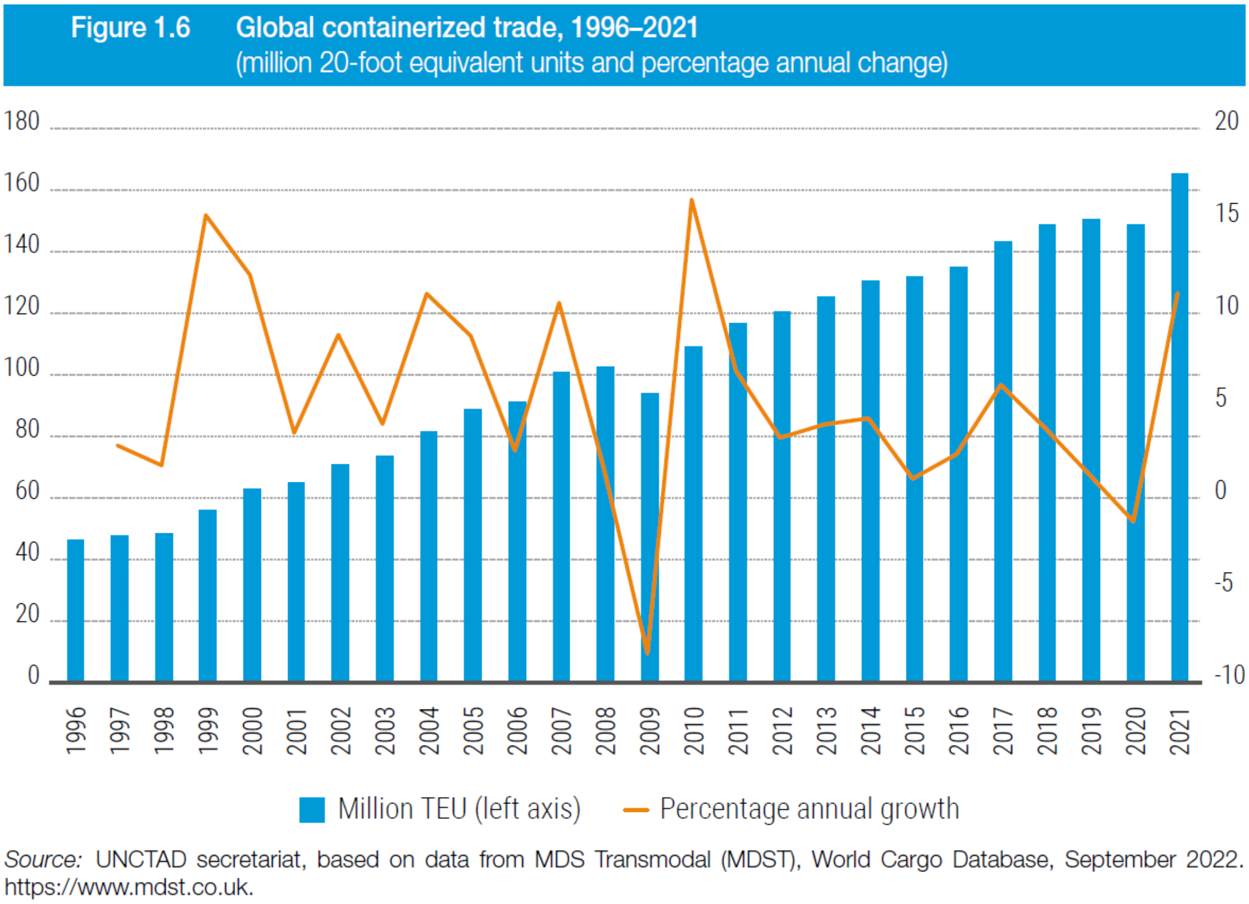

Figure 1. Global containerized trade (*Navigating Stormy Waters*, 2022)

In conclusion, the level of automation and the complexity of the cranes has increased. Consequently, the investments are more expensive. The containerized trade has continued to increase. The overall availability and productivity requirements of the existing cranes have increased. Therefore, the expectations to the cranes, both old and new, are set high. There has been a lot of research done to the optimization of the port operations. However, the optimization relies on productivity of the cranes. The crane productivity is the foundation where the optimization lies upon.

There are several reasons that can degrade crane performance or cause malfunctions. Issues that prevent crane operations completely, are quickly detected by the crane operators. Conversely, problems that degrade crane performance might not be noticed immediately. Additionally, solving the problem can be a challenging task. Abundance of instrumentation and software requires vast amount of expertise to solve the issues that cranes may have. Performance degrading issues vary from adjusting a limit switch to calibrating virtual GPS track that an RTG-crane follows during gantry movement. Regardless of the reason, degradation of crane performance is always a disappointment to both port and the machine supplier.

There is a number of key performance indicators (KPIs) to monitor the crane performance (Rintanen, 2018). However, most of the existing KPIs are affected by TOS optimization and other external factors at port. There is limited amount of research done in the field of pure crane performance. Even though crane performance contributes to these KPIs they are not direct measures of crane performance. This paper focuses on measuring the crane performance using simple artificial intelligence methods and excluding the external factors from crane productivity.

Crane performance is of the utmost importance to the commissioner of this work, Konecranes. Konecranes was founded in 1994 via listing of the KONE corporation (Konecranes, n.d.). The history Konecranes cranes goes back to 1910 when the electrical motor repair shop was founded. The first sizeable electric overheard cranes were delivered in 1933 to pulp and paper and power industries. Konecranes has continued to grow and expand over the decades by growing the existing lifting businesses and expanding to new ones. Konecranes port solutions operates on all continents. It has delivered equipment ranging from AGVs to large goliath gantry cranes. This paper focuses on the ASC of which Konecranes has already delivered hundreds of.

2 Theoretical background

2.1 Automated stacking crane

ASC stands for automated stacking crane. The crane is used to handle containers at container yard. ARMG crane is abbreviation for automated rail mounted gantry crane. Both names are often used interchangeably for the same kind of crane. Most typical use for these cranes is at port for stacking operations between the waterside and landside. An ASC is sometimes also used for intermodal operations with rail cars.

A typical layout for two ASC blocks can be observed from Figure 2. In the figure, there are two blocks, and, in each block, there are two ASCs. The ASC travels on crane tracks, marked on red lines in the figure. ASC always travels on rails. The block has stack area to store the containers. There are usually maintenance areas at the end of blocks for the cranes. During maintenance an ASC block can continue operations with one crane.

When a ship arrives at port, the container is first lifted with quay crane from the ship. The container is then transported by a straddle carrier or AGV to the waterside of ASC block. This is called the horizontal transportation. There are also other ways to implement horizontal transportation, but it has negligible impact to the ASC operations. The container is picked by the ASC from the AGV and transported into the stack. If the container is refrigerated, it is transported to the reefer area. When a truck on landside comes to pick up the container, it is transported by the ASC from the stack onto the truck. Operations from landside to waterside are operated vice versa. The purpose of stack and reefer areas are to store the containers for some time.

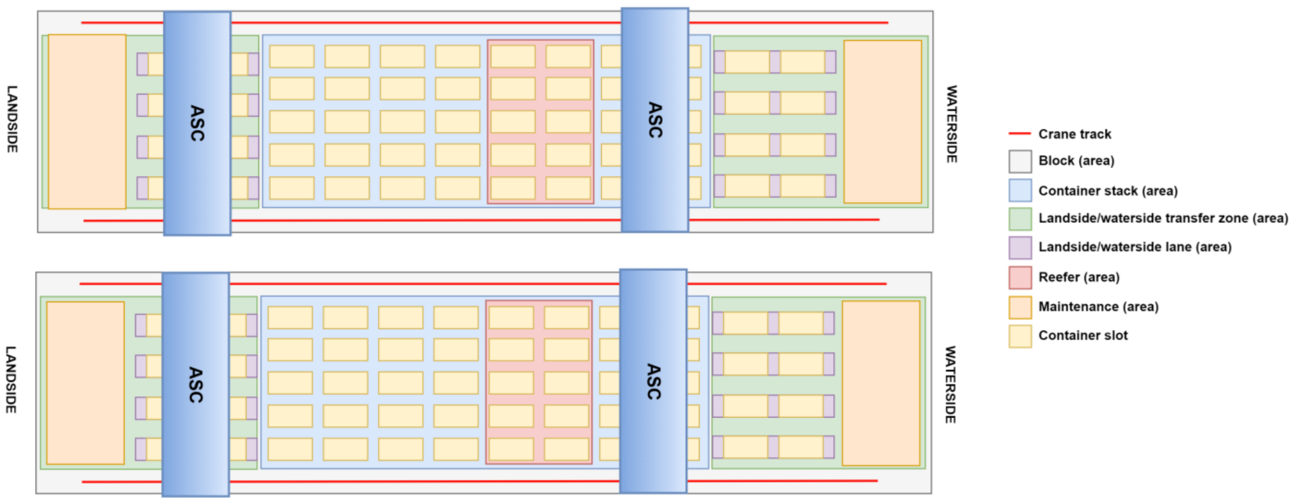

Figure 2. Typical ASC Block layout

The main parts of an ASC are legs, main girder, and trolley. The weight of the crane, and the load, typically rest on 16 wheels. Height of the crane is depending on customer requirements. Typical height for an ASC is 1-over-6. The height is measured as standard containers. 1-over-6 means that the crane can transport one container over stack of six containers. Rail span of the crane is between 6-10 containers. Containers are picked and placed by a spreader that is hanging by ropes from the trolley. A picture of typical ASC port can be observed from Figure 3.

In an ASC block there are coordinate systems for block and the naming convention for the cranes. In Figure 3 above there are numbers 14 and 13 visible on both the cranes and at the end of blocks. These numbers mark the block numbers. Normally there are two ASCs in each block. One ASC is on the landside and another one the waterside. The closest cranes in Figure 3 are waterside-cranes and they are normally marked with W-letter after their block number. The cranes, in Figure 3, that appear to be closer to sea, are landside cranes at this port.

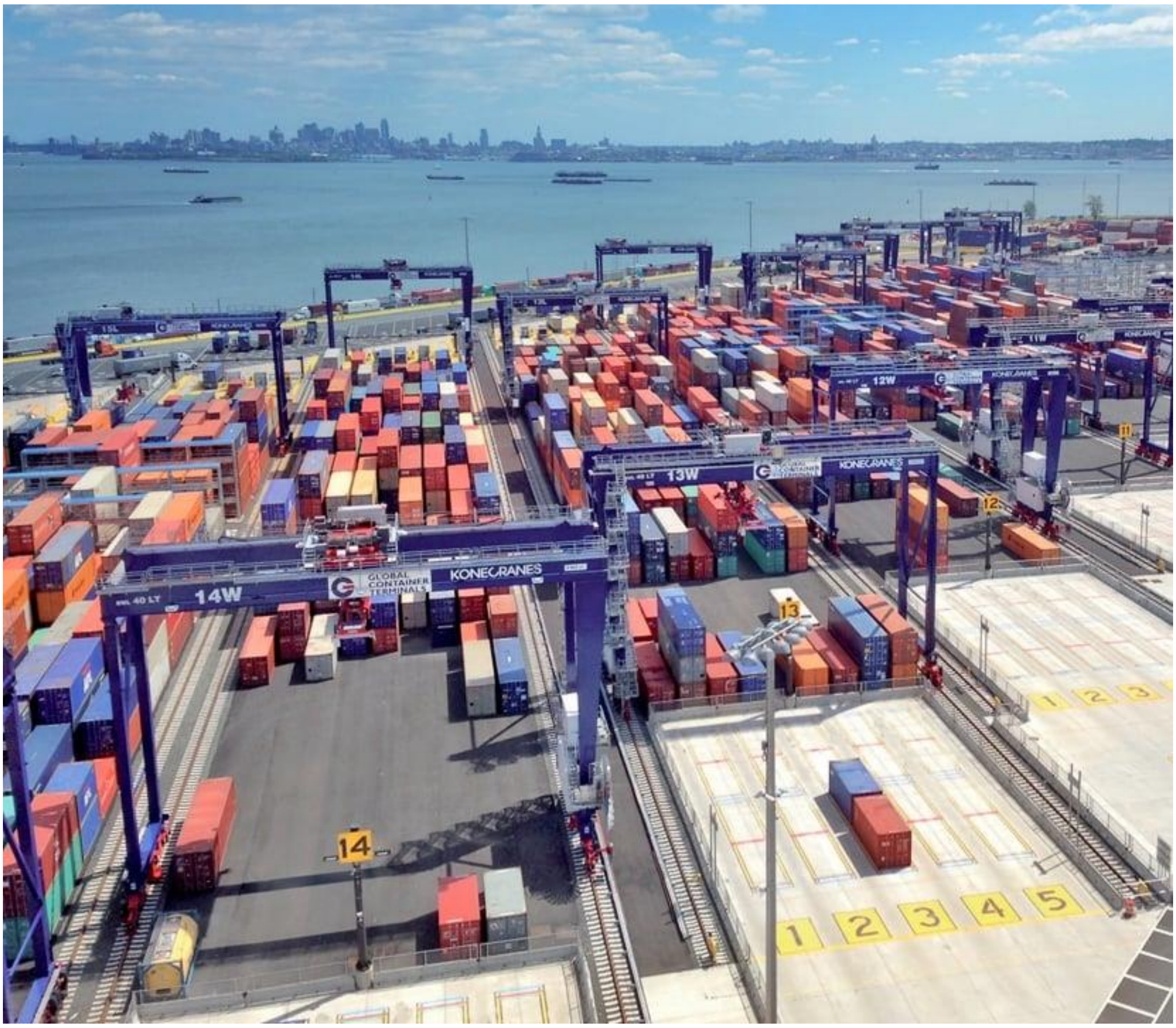

Figure 3. ASCs at port (Konecranes, 2023)

There are three axes on any given block. Coordinate system can vary between different ports. However, normally right-hand coordinate system is used. Origin is on the landside corner where X-axis increases to the waterside. Crane gantry moves along with X-axis. X-axis is called bay. Y-axis, the row, is the direction of trolley movement. Containers are moved vertically by hoist along with Z-axis. Base of the Z-axis is ground level. Z-axis is called tier and it means how many containers are stacked on top of each other.

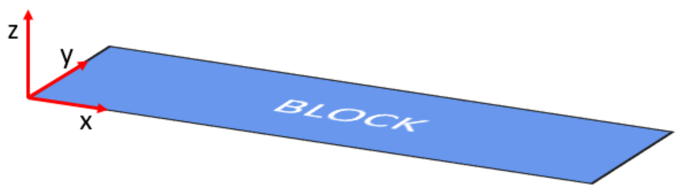

Figure 4. Block coordinate system

2.1.1 ASC KPIs

Previously, it was typical for port cranes to be driven manually by a crane operator (Rintanen, 2018). Traditionally there were three important KPI's for cranes and those were the gantry, trolley, and hoist speeds. Higher speeds enable crane operators to complete the moves promptly. It was important that the crane was operating as fast as the driver could command it. Today, new cranes are frequently run as automated operations. Consequently, automated operations have changed the most important KPI's for cranes. Choosing the best way to measure the crane performance is paramount to accomplish detecting cranes that are performing worse than others.

There is plethora of KPIs related to port operations in general. Some of the KPIs are for comparing ports to another, such as CPPI (World Bank, 2023, p. 27). Others are for monitoring port operations internally. For instance, according to Morales-Fusco et al., port of Rotterdam is using 32 different KPI to grade its port operations (2016, p. 376). These KPIs are valuable for measuring port performance in many ways. However, the KPIs in the report are used to measure port performance and not the performance of ASC's. The cranes are contributing to the ports KPI's. However, the performance of an ASC is measured differently. The KPIs for ASCs according to Rintanen are following (2018):

- Moves per hour
  - Number of container moves crane can perform in one hour.
- Cycle time
  - Time it takes for a crane to perform the job order directed by ECS or TOS.
- Truck service time
  - Time the truck must be present on the landside interchange area.
- Truck turn-around time
  - Time between truck entering and exiting port area.
- MMBF
  - Mean moves between failure.
- Availability
  - The degree to which crane is in operational state.

There is abundant amount of research related to KPI selection. The KPIs listed by Rintanen are useful for port operators in many ways to measure the ASCs. However, they are not necessarily as important to a machine supplier as they are to the port operator. It does not mean that a machine supplier would not care for their customer. The way many of the KPIs are calculated are only partially applicable to a machine supplier. The amount of important KPIs to a machine supplier will be narrowed down by elimination.

Moves per hour is quite commonly used KPI for measuring a crane workload. Most terminal are able to accomplish 25-35 moves per hour per crane (Visser, 2017, p. 16). It would be problematic to use this performance metric to detect cranes differing in performance. The workload of the cranes is highly dependent on the traffic at port. Ports operating model is also rarely around the clock, throughout the year, on all cranes. Moves per hour is also highly dependent on TOS optimization. TOS will prioritize the work orders depending on the situation at port (Cobo, 2016, p. 14). This can lead to a situation where one crane in the block is getting more work orders than the other. It means that moves per hour on the crane getting less work orders will decrease. Additionally, different blocks can get more moves depending on the situation at port. Performance comparison between the cranes would be hard to make and we can exclude moves per hour from further evaluation.

Truck service time and truck turn-around time are excellent indicators of port truck handling performance. Reducing the truck turn-around time is increasingly important for ports. High turn-around times lead to congestions, unbalanced distribution of workload and reduced utilization of terminal equipment (Abdelmagid et al., 2022, p. 22). Ports are monitoring their emissions as part of their KPI's and are actively working on reducing them (Port of Rotterdam Authority, 2022, p. 10). Truck turn-around time is part of the emissions monitoring (Interreg Europe, 2020, p. 62). ASC crane itself is affecting the turn-around time only for the duration of truck service time. Therefore, we can discard the turn-around time as it is not a good indicator of ASC performance. Additionally, for the ASC, the landside moves are only representing one thirds of the job cycles. Truck service time is acceptable performance indicator but there are other indicators that represent the crane performance better as a whole.

What is lacking in Rintanen presentation of ASC KPIs is the hit rate; specifically, the hit rate that is calculated by dividing the number of successful moves by total number of moves. Unsuccessful moves are defined as moves that require unplanned operator assistance. Hit rate is a valid and important metric to both port and machine supplier. Ports assume that the automated crane is performing most of the jobs without manual intervention. Consequently, machine supplier is keen on having the hit rate high. There is abundance of variations in ASC operational environment that can affect the hit rate. Calibration of the sensors can be off, containers can have deformities, high gusts of wind can move the container during lift and ground can be misaligned, to name a few examples that can cause hit rate to decline.

Cycle time is an important variable for consideration. Cycle time has several benefits over other the other variables. It represents the whole duty cycle of the crane. It can also be assumed that many of the problems on the cranes affect the cycle times. It is also impervious to intermittent port operations. When the port operations are halted, the cycle time remains unaffected. The recorded cycle times in addition to waiting times available for each duty cycle are available for this research. However, the waiting time does not account for the deceleration and acceleration time of the crane.

The cycle time has also some disadvantages. Overall equipment effect can be observed from Figure 5. The cycle time can be compared to the productive time in the figure. Therefore, the downtime and idle time are not affecting the cycle time. Downtime and idle time are not caused by the machine supplier or indicative of ASC performance. However, TOS optimization is partially realized as a plan loss in Figure 5. Two cranes working on the same area can lead to one crane waiting for the other crane for area clearance which is realized as a plan loss. Waiting of the other crane is sometimes unavoidable. The waiting time can be mitigated using the data available. The crucial factors to a machine supplier, which are included in the cycle time, are equipment performance, equipment quality, automation quality.

However, what cannot be accounted for in cycle times is the situation when both cranes inside the block require the other crane to move in the reverse direction of its work order. This is called the dead-lock avoidance. It means that one crane needs to reverse and to make way for the other crane. The reverse move is not recorded as a waiting time. There also is not a way to mitigate the effect of reverse move in the realized long cycle time. Therefore, usage of cycle time directly is not representing an explicit measure of ASC performance. Again, this KPI is relevant for the port operator but not necessarily to the machine supplier supplying the ASCs.

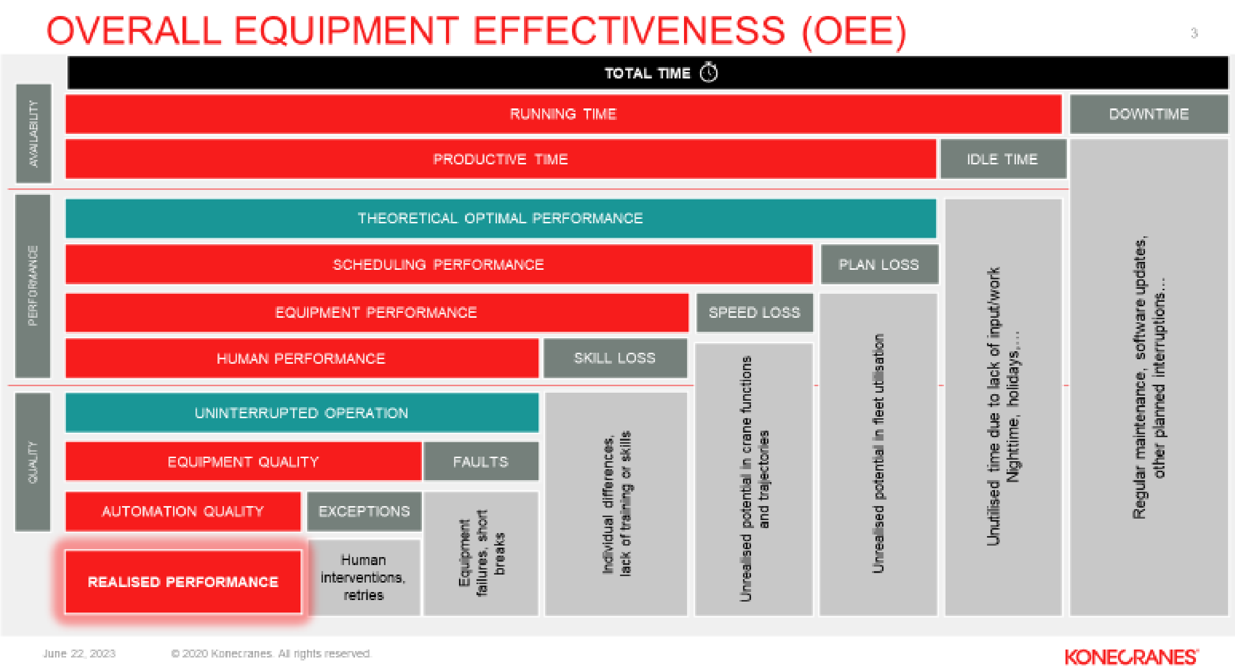

Figure 5. Overall equipment effectiveness (Konecranes, 2023)

Availability is also a common and important metric to all port cranes. However, the availability of ASCs cannot be calculated from the data that is available for this research. Therefore, it will not be discussed any further.

2.2 Machine learning

Most machine learning can be split into two categories. The categories are supervised and unsupervised learning (Goodfellow et al., 2016, p. 96). In supervised learning, the output variable is predicted based on the input measures. In contrast, the unsupervised learning is not given any outcome to output (Hastie et al., 2009, p. 2). Unsupervised learning is typically used for clustering, pattern recognition and relationships. Supervised learning is typically used for classification and regression.

The problem setting, detecting cranes that perform worse than others, can be approached as both supervised and unsupervised learning. The challenge with supervised learning is that there must be something to measure. Something to measure is also positive for the problem setting. There is also possibility to use unsupervised learning, that doesn't require pre-defined output. However, the interpretability can be hindered without the pre-defined output of unsupervised method. It can be challenging to estimate why one crane is marked as differing in performance without any clear indication for the reason. The target for this research is to provide understandable metric to monitor the crane performance. The interpretability is in the core of this metric. An engineer should be able to understand what crane is performing better or worse based on pre-defined performance metrics. Therefore, unsupervised machine learning will be excluded from further consideration. The work will proceed with supervised learning methods.

The supervised learning method must be further narrowed down. Most of the supervised learning can be further categorized as either regression or classification. There is multitude of both classification and regression that can be done with cranes.

With classification, it is possible to classify many concepts with cranes as good and bad performing. Items that could be classified, as an example, are crane performance, successful moves, and successful auto-landings. It is also possible to classify whether some crane is having more alarms or faults than others. However, the number of alarms or their duration is not directly indicative of crane performance. What can be assumed is that the faults in cranes lead to longer cycle times. Moves can also be classified as successful and unsuccessful and the ratio could be counted for each crane individually. However, the success ratio can be too inexplicit metric for the engineers to monitor. Labeling a crane to a bad performer would raise question regarding the reason why it’s labeled as bad performing. Another issue with classification is that it doesn’t tell how badly a crane is performing.

ASC moves are actively classified by the hit rate. Hit rate is an excellent metric that is monitored by both port operators and the machine suppliers. However, it does not explain the overall crane performance. Nevertheless, moves with manual intervention can be expected to have longer cycle times than moves without manual intervention.

Classification can also be used to detect problems in the cranes by utilizing the faults and alarms. It would be possible to help solving a problem in a crane by using the alarm and fault data. However, the alarms that are generated by the cranes can be a chain-reaction. One alarm can cause multiple consecutive alarms. It is also challenging to define the root cause of the problem from the alarms alone. Additionally, if the crane is completely down and unusable it is quickly noticed by the operators. There is little need or benefit to indicate that a crane is down using classification. Additionally, there is no information available in the research data how the faults and alarms in cranes were resolved.

With regression, the cycle times and various other performance metrics can be predicted. When thinking of the crane operations from machine supplier’s perspective, it is often the move cycle time that explains most of the crane performance. However, as previously mentioned, the cycle time is not explicitly an important metric for an ASC. The cycle times can vary a lot depending on TOS. Additionally, the ongoing port operations can affect the cycles times. For instance, cranes perform housekeeping moves to prepare for a ship that is approaching the port. This way, the containers that are loaded to the ship are in the optimal location when the ship arrives. Additionally, slots near the waterside can be emptied to make space for the upcoming containers. Housekeeping moves can have long cycle times, but they cannot directly be classified as good or bad moves. Short moves near the waterside to waterside can have small cycle times but they are not necessarily successful or unsuccessful based on the cycle time.

The most viable way to proceed is with supervised machine learning and more specifically regression. The focus will be on predicting the cycle time. However, external factors will be accounted for. The exact methodology for mitigating the external factors will be described in chapter 3.

2.3 Regression

Montgomery et al. describes regression model building process as an iterative and repetitive process (2012, p. 10). The process starts by using the theoretical knowledge of the process and the data that is available. The data is often visualized to help in specifying the model. The result is initial regression model. After the initial model is built, the parameters are estimated, and it is done typically using least squares or maximum likelihood. The adequacy of the model is then checked. Adequacy is first checked by assessing the variables that were inputted to the model. The model can contain unnecessary input variables or lack the important ones. Also, the data that is inputted to the model can be faulty and it can affect the model. The adequacy of the model is checked each time after changing the input variables and assessing the input data. After the model is adequate, it is validated. The model must produce acceptable results for the final application. The process is yet again repeated until the results are viable for the final application. The process is visualized in Figure 6.

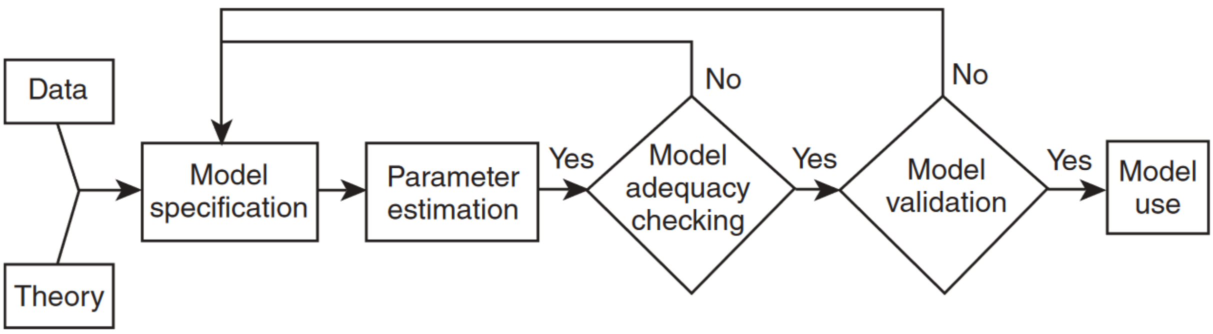

Figure 6. Regression model-building process (Montgomery et al., 2012, p. 11)

2.3.1 Linear regression

Linear regression is one of the easiest and popular ways of doing machine learning (Javatpoint, n.d.). Montgomery et al. describes the regression analysis as a statistical technique for investigating and modeling the relationship between variables (2012, p. 1). The data points are called independent variables. The dependent variable is the one that is being predicted by the linear regression with the independent variable. An example of simple linear regression can be observed from

Figure 7. Dependent variable y is explained by the independent variable x. The data points are in purple color. The resulting regression line is in green color.

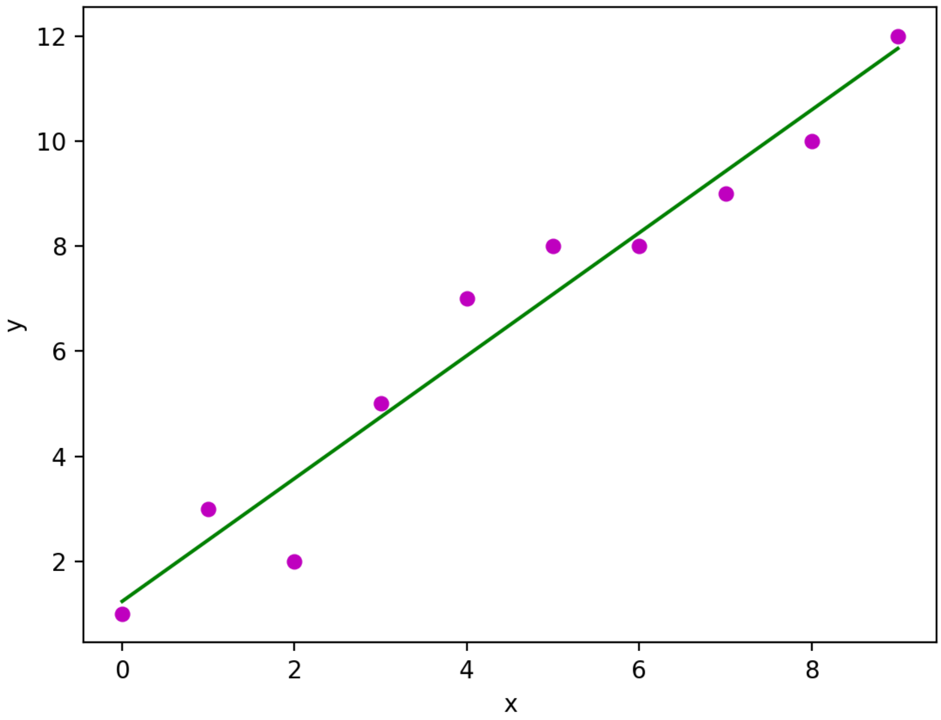

Figure 7. Simple linear regression

Linear regression is commonly a multiple linear regression. Multiple linear regression was one of the most widely used statistical methods decades ago and remains popular today (Hoerl & Kennard, 1970a, p. 69). Instead of one independent variable, there are multiple independent variables. Formula for multiple linear regression is following.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_\kappa X_\kappa$$

Where:

- Y is dependent variable,
- β_0 is constant/intercept
- $\beta_{1\ldots\kappa}$ is coefficient and
- $X_{1\ldots\kappa}$ is independent variable.

After fitting the values there is a residual error between the fitted value and actual value. Residuals are the deviation between the data and the fitted value (Montgomery et al., 2012, p. 130). The residuals will be closely analyzed as they explain a lot how the model is fitting. Residuals are also often used in anomaly detection. For instance, magnitude of residual can be used as an anomaly score (Chandola et al., 2009, sec. 15:32). Residuals are calculated by following formula, where e is the error for observation i. $\hat{Y}_i$ is the predicted value of i_{th} observation and Y_i is the actual value of the i_{th} observation.

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_\kappa X_{\kappa i})$$

One of the fundamentals of machine learning is the performance measurement. One of the measurements is the sum of squares error. The term has different names depending on the source. We will use the term sum of squares errors (SSE). The most common way to do linear regression is with ordinary least squares (OLS). OLS is usually the starting point to proceed with other methods of machine learning. The principle of OLS is that it minimizes the sum of squared errors between the data points and the regression line. Analysts use this type of linear regression most frequently (Frost, 2019, p. 25). Formula for calculating the SSE is shown in the equation below:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The SSE measures the overall variability of the distance between the data points and the fitted values. There are also other types of sums of squares in OLS. Regression sum of squares (RSS) measures the amount of additional variability the model explains to a model using only mean to predict the dependent variable. Total sum of squares (TSS) measures the overall variability of the dependent variable around its mean (Frost, 2019, p. 33).

$$RSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2$$

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

These three formulas tell of the total variability of the model. OLS minimizes the SSE, which is the unexplained variability of the model. Consequently, when the unexplained variability of the model decreases, explained variability increases. Therefore, model improves as the SSE decreases and the RSS increases. These three formulas are also tied to another.

$$TSS = SSE + RSS$$

$$Total\ variation = Unexplained\ variation + Explained\ variation$$

Coefficient of determination, or R-squared, (R^2) is a measurement that assesses the model's ability to predict or explain the outcome in linear regression. R^2 indicates the proportion of variance of the variance in the dependent variable that is predicted by the model with independent variable X (Enders, n.d.). R^2 is always a value between 0 and 1. A good R^2 result is depending on the context.

Studies that try to explain human behavior have an R^2 values typically less than 50 % (Frost, 2019, p. 123). In physical sciences, R^2 can be expected to be close to 100 % (Enders, n.d.).

The ASC-cranes can be defined as physical sciences rather than human behavior. However, there is also strong human factor involved with the ASC-cranes. Port operations are not always straightforward, truck drivers can be erratic, and the ROS-operator can be unpredictable. There are also inadequacies in the data that will negatively affect the R^2 results. However, a well-defined model in this domain can be expected to reach high R^2 -values over 90 %. R^2 can be calculated from RSS and TSS or SSE and RSS with following formulas.

$$\mathrm{R}^2 = \frac{RSS}{TSS} = 1 - (\frac{SSE}{TSS})$$

Neither SSE nor R-squared is good measurement for linear model on its own. Smaller SSE with nested models is always associated with increasingly complex model (Yan & Su, 2009, pp. 165–166). R-squared has similar problems as the SSE. More variables can be added to the model that will result in higher SSE. However, TSS will remain the same. Therefore, the model will always get a better R-squared score when adding more variables (Montgomery et al., 2012, p. 87).

Model complexity must always be taken into consideration when creating the model. More complex model that achieves slightly better SSE than its simpler counterpart is not necessarily better. Simple model is usually more easily interpreted than a complex model. Interpretability is also a measurement of a good model. However, interpretability is hard to measure. Some methods, such as centered input variables, have been proposed to improve to interpretability of a complex model (Schielzeth, 2010, p. 103). The emphasis will be put on making the model as simple and interpretable as possible. Simpler model with slightly worse regression results is better than a complex one.

Some model builders prefer to use adjusted R-squared (R_a^2) over the R-squared (Montgomery et al., 2012, p. 87). The benefit for using the adjusted R-squared is that variables that don’t add explanatory value to the model can cause adjusted R-squared to decrease (Frost, 2019, p. 147). However, Yan & Su argue that the adjusted R-squared can be inadequate for model selection since it lacks the ease of interpretably of R-squared (2009, p. 166). Burnham et al. describes the adjusted

R-squared useful as a descriptive statistic but not useful for model selection (2002, p. 37). Adjusted R-squared is nevertheless useful for evaluating and comparing the candidate regression models (Montgomery et al., 2012, p. 88). In this paper, adjusted R-squared will be included as one of the metrics when comparing the regression models. Adjusted R-squared can be calculated with following formula:

$$R_a^2 = 1 - (\frac{MSE}{TSS})$$

2.3.2 Other regressions methods and multicollinearity

Ridge regression is one of the proposed solutions to solve the problems of multicollinearity with OLS (Hoerl & Kennard, 1970b, p. 85). Multicollinearity is the existence of near-linear relationships among regressors, predictors or input variables. Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false and nonsignificant p-values, and degrade the predictability of the model (Saleh et al., 2019, pp. 3–4). According to Montgomery et al. there are four primary sources of multicollinearity: data collection method, constraints on the model, model specification and over defined model (2012, p. 286). While the OLS is unbiased estimator, ridge regression is biased.

There is a measure of overall multicollinearity of the variables that is called the condition number (Young, 2017, p. 113). Condition number between 10-30 indicates presence of multicollinearity and when the value is over 30, there is a high level of multicollinearity (J. H. Kim, 2019, p. 559; Young, 2017, p. 113). Condition number r is calculated by correlation matrix, where λ_1 is the minimum, and λ_{p-1} is the maximum eigenvalue:

$$r = \sqrt{\frac{\lambda_{p-1}}{\lambda_1}}$$

There is also a method of regression through origin (RTO). It resembles OLS. However, RTO deletes the β_0 constant from the formula. Therefore, the pivot point for the regression line is always at

0,0. Using this type of regression is generally not recommended (Young, 2017, pp. 13–14). As a result, RTO will be excluded from further examination.

2.3.3 Model and sample selection

Sample selection

Typically, part of the research data is withheld from the model during tuning of the hyperparameters. The data is split into separate data sets that are called the training and test set. The main purpose of this is to achieve optimal complexity of the model without overfitting. Overfitting cannot be diagnosed with same data set that has been overfit (Burnham et al., 2002, p. 248). Overfitting occurs when the models adapts itself too closely to the training data and will not generalize well (Hastie et al., 2009, p. 30). Typical split is 80 percent training data and 20 percent test data (Goodfellow et al., 2016, p. 119).

Figure 8 is an example of bias-variance tradeoff that is also called as principle of parsimony (Burnham et al., 2002, p. 31). As the model complexity increases, lower bias and higher variance will be achieved. Additionally, as the model complexity increases the training error decreases and eventually, the test error starts to increase. This is a sign of overfitting. In addition to overfitting, high variance can be a sign of collinearity issues (Kuhn & Johnson, 2013, p. 98) Ideally, a model whose test error is as small as possible would be chosen.

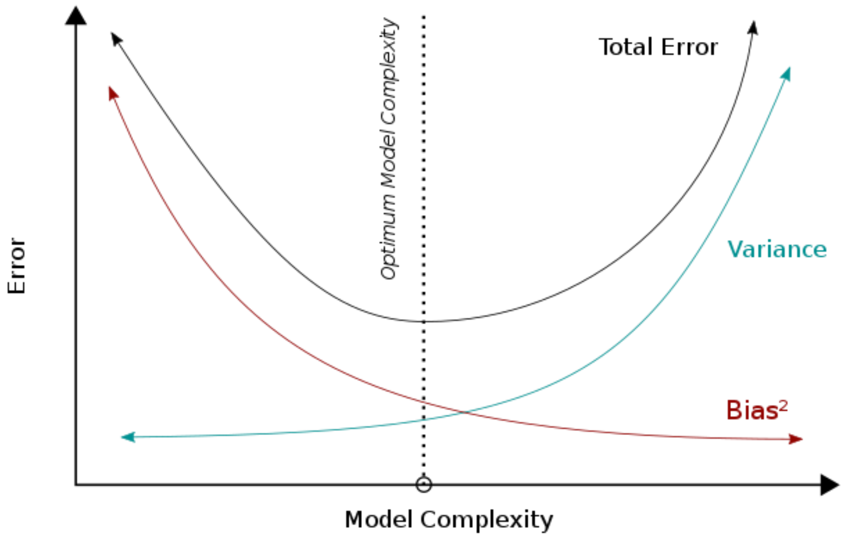

Figure 8. Error as a function of model complexity

However, with the problem setting of this paper, overfitting is not big consideration. The model will be simple rather than complex. The training sample compared to the model complexity will be exceptionally large. Overfitting is generally problem when the number of input variables is too high compared to number of observations (Frost, 2019, pp. 182–183). Therefore, separate training and test sets will not be used. Emphasis is put on the model simplicity which will, consequently, avoid problems of overfitting.

Model selection

There are multiple methods for selecting the optimal model. There is not common consensus on which selection method is the optimal one. Most common mathematical method used is sequential testing with either stepup or stepdown (Burnham et al., 2002, pp. 35–36).

Stepdown, or backward elimination, is a method where the model starts with all the predictors. F-test statistic is calculated for the model that compares model with all predictors to a model with each predictor removed individually. Least significant predictor is then removed from the model. Process is repeated until the largest p-value is smaller than the threshold significance level. Common choices for significance levels are 0,10 and 0,05. Problem with stepdown is that the variable has no chance to re-enter after it has been eliminated from the model. Variable that is excluded from the model at early stage can become significant later on when other variables are excluded from the model (Yan & Su, 2009, pp. 171–172).

Stepup, or forward addition, is a method where the model starts with just the intercept. Stepup works inversely to the stepdown method. The model is then tested with each variable individually. The variable that has the highest significance level based on F-test statistic is selected to the model. The process is then repeated until threshold significance level is reached. Typically, significance level 0,05 is used (Freedman, 2009, p. 70; Yan & Su, 2009, p. 171). Conversely to stepdown model, variable that was significant in the early model can turn out to be insignificant in the later model. Stepup method is rarely used in applications (Yan & Su, 2009, p. 172).

Stepwise search is a method, which tries to avoid the problems in both stepup- and stepdown-methods. It works like the stepup-method where variables that have the highest significance level based on F-test statistic are selected to the model. However, stepwise search checks the previously added variables of the model before adding a new one. Previously selected variables, that have p-value greater than pre-selected threshold-value are deleted from the model. This helps the avoid the problem like in the stepup-method where variable can turn out to be insignificant when more variables are added to the model. Stepwise search continues to add variables until there are no variables available, that have F-test statistic more significant than the selected threshold-value. Stepwise search also continues to remove insignificant variables from the model until model only has significant variables left. Out of these three stepwise selection procedures, the stepwise search algorithm performs best (Yan & Su, 2009, pp. 172–173).

There are also numerous other variable selection-methods, such as the Bonferroni-method. It is however rather conservative in its variable selection. Type II error is a common problem for the Bonferroni-method. More liberal method such as Fisher's LSD or previously mentioned stepwise search should rather be used (Lee & Lee, 2018, p. 357). De and Baron also concluded that the stepwise methods have advantages over the Bonferroni-method (2012, p. 2067).

Yan & Su advice that one should be careful when using the automatic variable selection procedures (2009, p. 173). They argue that there is a high probability of including unimportant predictors and excluding the important predictors. The model building process will be conducted as advised by Yan & Su, in iterative manner by alternating between model selection and diagnostics. Therefore, both domain knowledge and automatic variable selection procedures will be utilized. Automatic variable selection will be carefully inspected once performed.

Sample size

Austin and Steyerberg studied the number of subjects required per variable in linear regression. The study concluded that the minimum number of samples required per independent variable is merely two (2015, p. 636). Green gathered some support for a rule-of-thumb that N > 50 + 8 m for the multiple correlation and N > 104 + m for partial correlation. N represents the number of subjects and m the number of predictors (1991).

As an example, with 10 independent variables, a minimum of 20 samples would be required according to Austin and Steyerberg. By utilizing Greens rule-of-thumb, with 10 independent variables, at least 130 samples are required for multiple correlation and 114 samples for partial correlation.

It must be noted that both studies have been conducted in the field of psychology. Most of the studies in the field of linear regression sample sizes are done in the fields of medical and psychological, where the sample sizes are often limited. For this paper there is abundance of data samples available for linear regression. Additionally, the number of predictors will be low. Therefore, it can be concluded that the sample size for making the regression model will be adequate if previous minimum recommendations are followed.

2.4 Statistics

2.4.1 RMSE and MSE

In addition to previously presented metrics, other methods will be used to measure the regression model. Both R_a^2 and R^2 are common metrics for regression model builder. However, they both suffer from same problem. They are not so easily interpreted by people not familiar with statistics. The results of this research are also presented to people not familiar with statistics. Additionally, high R^2 can be achieved and still have insufficient model because of high RMSE (Kuhn & Johnson, 2013, pp. 96–97).

Two common methods will be used to communicate the error of our model. The first one is RMSE (Root-mean-squared-error) and the second one is MAE (Mean-Absolute-Error). RMSE can be derived from MSE (Mean-Squared-Error) (Freedman, 2009, p. 21). However, RMSE is more useful than MSE as it uses the same unit scale as the original dependent variable (Kuhn & Johnson, 2013, p. 95).

MAE is also useful metric to convey the regression results forward. It is probably the easiest metric to comprehend. RMSE will always be larger or equal to MAE. This is due to squaring of the errors in RMSE. Large difference between the two is indicative of some high errors. RMSE and MAE are calculated by using following formulas:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

2.4.2 Pearson and Spearman correlation

Pearson correlation coefficient is a bounded index that provides unitless measure for the strength and direction of the association between two variables (Young, 2017, p. 14). Correlation varies between -1 to 1. Zero implies that there is no correlation. Both Pearson and Spearman correlations are good tools to measure correlation between the dependent variable and the independent variables. Independent variables are the candidates for linear regression. Pearson correlation coefficient is calculated by following formula:

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman is more suitable for data that is not normally distributed. However, it is always recommended to calculate both Pearson and Spearman-correlation (Rovetta, 2020, p. 6). Spearman correlation is calculated by:

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2}}$$

Thus, both Pearson and Spearman-correlation will be calculated for the independent variables. However, correlation is not necessarily causation. With the ASCs and the input variables candidates, it can be assumed that relatively high levels of causation exist for the main moves. A simple reason for this is because the move cycle is completed by using the main moves.

2.4.3 Normal distribution

There are different kind of tests to test whether the data is normally distributed, such as the Shapiro-Wilk test. Many statistical tests assume that the sampled data comes from normal distribution (Thode, 2002, p. 2). One of the assumptions of OLS regression is normality of the data. Small departures from the normality assumption do not affect the model greatly, but gross non-normality is potentially more serious as it affects dependability of multiple tests (Belsley, 2004, pp. 18–19; Montgomery et al., 2012, p. 136). Despite the different shapes and forms of the normal distribution all have following characteristics (Frost, 2020, p. 129):

- They are all symmetric. The normal distribution cannot model skewed distributions.
- The mean, median and mode are all equal.
- Half of the population is less than the mean and half is greater than the mean.

Emphasis is put on the visual appearance of the normal distribution. Data is assumed to follow normal distribution. The data will be prepared so that only small departures from normal distribution can be observed. Large departures of the dependant variable from normal distribution are assumed to be bad moves.

2.4.4 Kernel Density estimate

Kernel density estimation (KDE) is a way to estimate the probability density function (PDF) of a random variable in a non-parametric way (*Scipy.Stats.Gaussian_kde*, n.d.). KDE will be used to plot probability estimates of the cycle times. It will provide a visual presentation of the data distribution. Existing methods of Seaborn and Scipy will be utilized. KDE could also be utilized to generate data points that appear to come from a certain dataset.

2.4.5 Influental data points

There is a lot of unexplained variance in normal ASC operation due to various circumstances. This will affect the recorded cycle times of the moves. This in turn will lead to influential points in our data. Influential observations are data points, where data point x value is usually moderately unusual and the y coordinate is unusual as well (Montgomery et al., 2012, p. 211). There are no expectations for unusual x values in the data. However, the y-values can be highly unusual. It means that the actual cycle time can be odd in terms of regression independent variables. Therefore, statistical methods are required to detect these points from the data. Otherwise, they will skew the regression model. Target is to have normally distributed cycle times for the OLS.

There are numerous methods to handle the outliers. Yan & Su lists the most common methods as HAT diagonal elements, DFFITS, DEBETAS, Cook's D, Generalized variance and covariance ratio (2009, pp. 149–150). Alternatively, Arimie et al. recommends using Jackknife residuals and Atkinson's measure to detect outliers (2020, p. 1). Each have their benefits and downsides. Out of these, the Cook's distance and DFFITS shall be utilized mostly due to their well-researched and documented usage. They are also readily available in Python external modules.

Both Cook's distance and DFFITS will be utilized to detect the influential data points from the data. DFFITS is abbreviation of Difference in fits. The benefits of both methods are, that they give more insights on the observations than using the residual of regression. The problem with using residual is that the high number of influential observations is already skewing the regression. There are general guidelines mentioned for cutoff-values for both methods. These guidelines are often presented as fixed values when there are not so many observations in the regression model. However, the results of both methods are highly dependent on the number of observations. The data includes large number of observations. Therefore, cutoff-values that are dependent on number of observations shall be utilized. There are also guidelines available on using observation-dependant cutoff-values. However, it is a common advice in literature to use application-specific cutoff-values (Montgomery et al., 2012, p. 218). Yan & Su also advice, that caution should be taken to remove the outliers (2009, p. 150). This application, however, requires deliberate use of removing the outliers.

Cook's distance

Cook's distance is the distance between the least squares estimates of regression coefficients with an observation included and excluded (Yan & Su, 2009, p. 149). Therefore, the Cook's distance is considered a deletion diagnostic (Montgomery et al., 2012, p. 217). Cook's distance considers both the leverage and residual of the observation. Disadvantage of Cook's distance is that an observation can have high influence on the estimated coefficients but only a minor influence on the predicted values of the dependent variable. Therefore, an observation can have a low Cook's Distance despite of its large influence on the estimated coefficients (M. G. Kim, 2017, p. 317). Cook's distance is calculated by following formula (Montgomery et al., 2012, pp. 215–216; Yan & Su, 2009, p. 149):

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{p * MSE}\left(\frac{h_{ii}}{(1 - h_{ii})^2}\right)$$

The distance, D_i, summarizes how much all of the fitted values change when the i^{th} observation is deleted. (*Identifying Influential Data Points*, n.d.). p is the number of independent variables. h_{ii} is the leverage. Cook's distance is highly dependent on the residual in the first term and the leverage. Thus, Di is made up of a component that reflects how well the model fits to the i^{th} observation. Another way to interpret Cook's distance is that it is the squared Euclidean distance that the vector of fitted values moves when the i^{th} observation is deleted (Montgomery et al., 2012, p. 216).

DFFITS

DFFITS is a statistical method for identifying the influential data points. It is defined as the difference between the standardized fitted values with and without the i^{th} observation (Yan & Su, 2009,

p. 148). Cook's distance measured how much all the fitted values change when i^{th} observation is removed. DFFITS measures how much the fitted value changes when i^{th} observation is removed. The formula for calculating DFFITS is below:

$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

DFFITS resembles Cook's distance quite a lot. It is also utilizing the leverage and residual values. The principle is also the same: deleting one observation and seeing how it affects the fit.

3 Research question, method and methodology

This research is research-based development with focus on working life development. The purpose is to detect cranes that are performing worse than others. Therefore, a way to measure cranes that are not performing well is required.

The current ASC KPIs are in different ways insufficient for monitoring the ASC performance from machine supplier's perspective. Additionally, many of the KPI's are not available or applicable to a machine supplier. Currently, one of the most important KPI to a machine supplier is the hit rate. However, it only measures the container handling accuracy of the crane. Hit rate is not directly indicative of high productivity of the crane. Although it is contributing to the productivity. Crane with a good hit rate can still have low productivity. However, a crane with a bad hit rate most likely has low productivity.

There can be dozens of cranes at one site. The machine supplier might not have constant local presence at port. Thus, the situational awareness decreases. It can become unclear which cranes are performing well and which ones badly. Therefore, the conclusion is that the current KPIs are measuring a small part of the operation or measuring the overall productivity depending on the prevailing situation at port. An additional KPI is required that is measurement of crane productivity excluding the external factors.

The solution to measure crane productivity is a machine learning model that combines reactive and predictive maintenance. The model identifies issues that require immediate attention that cannot be predicted beforehand. The model also detects problems that degrade the crane performance over time. Models' principle is to identify the cranes whose performance is deviating from others. Ideally, cranes whose performance is deviating from others is detected rapidly. The scope also includes data analysis of the available data.

Due to limitations of the current KPIs a new KPI is devised for ASCs. The KPI is move residual. The concept is to predict the cycle time for each container move individually with linear regression. The independent variables for the regression are the travel distances of the main moves and the type of move. Some additional input variables can be utilized if deemed necessary. The exact independent variables will be examined during the analysis of the regression model. This predicted cycle time is then deducted from the actual cycle time. The result is residual of the regression. In this case it is called the move residual.

$$\mathrm{Y}_{residual} = \mathrm{Y} - \widehat{\mathrm{Y}}$$

$$\text{Move residual} = \text{Actual cycle time} - \text{Predicted cycle time}$$

The move residual is a new way to measure crane productivity quantitatively. It tells the observer how many additional seconds the move took. As the travel distances of the main moves are used as input variables, heavy emphasis is put on the travel speeds. A crane with slower movement speeds will be observed from the move residuals. Additionally, crane with constant bad overall performance will be observed from the move residual.

The usage of main move travel distance brings a benefit to the performance monitoring. It completely negates the effects of TOS from crane performance monitoring. The alternative would be to calculate distance required to move a container from pickup-point to drop off-point. However, the research data clearly indicates that it would be impossible to get reliable results that way. For instance, cranes must sometimes give way to another crane working in the same block. As a result, cranes can record extreme travel distances to a seemingly short move. Thus, recorded travel distances are more usable than the calculated distances.

Another benefit of using the travel distances comes from operational status of the port. Many KPIs for ASCs correlate to operational situation at port. There are numerous external factors that affect these KPIs. The external factors include, but are not limited to, ship port calls, shift changes and holidays. The way ASCs operate can change when there is a port call. Additionally, if there is not staff monitoring the cranes, the cranes do not operate. Additionally, the TOS in use will affect many KPIs. All of the previously mentioned external factors do effect on the KPIs such as moves per hour. The move residual is impervious to these external factors.

Quantitative measurement brings a benefit to the performance monitoring. The observer is given a chance to make his own decision based on the graphs of the move residual. Quantitative measurement tells how well the crane is performing. Therefore, the observer is not constrained to any binary information of which crane is not performing well.

Additionally, the observer can observe how the move residual develops. If the move residual is high but becoming smaller there is less cause for concern. In contrast, should the move residual increase after some changes in the crane, the change can be determined to decrease the crane productivity.

The move residual is best utilized by visualizing it. When the move residual is visualized the observer can immediately assess the productivity of the cranes. High move residual is realized as deviation from other cranes. Thus, indicating to the observer the crane to inspect further for possible issues. This way the loss of situational awareness can be mitigated. The visualization will be altered with rolling average for the purposes of better visual representation. This will inevitably lead to some loss of data in the visualization. However, there can be a lot of hysteresis in the move residual between the moves. Therefore, for a general overview, it is best practice to filter the data.

Occasionally, cranes accumulate waiting times due to others cranes or other external factors. Waiting time accumulate for instance when a crane is waiting for the truck driver to acknowledge that the trailer of the truck is in the exchange-area. The waiting times that the cranes encounter during moves are not indicative of crane performance in any way. Thus, the effects of waiting times will be deducted from the move cycle times as it is available in the research data.

The linear regression model will only be trained on normal and uneventful moves. Otherwise, the bad moves cause skewing to the linear model. The regression building process in Figure 6 will be followed to create the linear model. Additionally, visualizations of the model fit, particularly residual plots are used to understand whether the model is fit for purpose (Kuhn & Johnson, 2013, p. 95). First, domain knowledge will be used to filter most obvious bad moves. During the process, more moves will be filtered as required. Applicable statistical methods will be used for further filtering the unexplained variance from the training data. The target is that completely normal move residual is zero. Thus, the move residual is tracking how many seconds the move is lacking behind than it was supposed to take. It also means that the better the bad moves are removed from training data, more accurately the model can perform. Therefore, all the outliers will be removed.

The approach of removing the outliers is not completely unforeseen. However, it is not common to aggressively remove outliers from the data. In fact, there are numerous sources that advise to be very cautious of removing the outliers (Belsley, 2004, p. 3; Frost, 2019, p. 235, 2020, p. 37; Yan & Su, 2009, p. 150). Nevertheless, this application benefits from this approach. The effects of bad moves with unexplained variance must be negated for the move residual to perform adequately.

The outliers are later fitted to the trained model. These moves will have a greater move residuals. It means that the model could not predict the cycle times of these moves accurately. It is an indication of unexplained variance in the move. This is a more statistical way to describe the move residual, measurement of unexplained variance in the move.

Move residual has some limitations. One of the limitations of this method is that there must be well performing cranes for training the model. Although it can be argued that if none of the cranes are performing satisfactorily, the machine supplier will know about. Thus, any model for indicating bad cranes would not be required. Such situation could only be encountered during early stages commissioning of the cranes. The situation changes as the commissioning is completed and the cranes are in production. As the cranes have been in production for a few months there is already enough material for training the model. Thus, the move residual is best trained when the cranes have been in production for some time. Additionally, after the cranes have been in production for some time it is a good time to check the productivity. The productivity will inevitably be checked when a linear model is trained.

Move residual cannot indicate the reason for the longer cycle times. However, it indicates the cranes which require further examination. If the residual is high, move cycles are taking longer than on other cranes. The move residual is most useful when visualized. Later, trends and changes in the move residuals of the cranes will be visualized. These changes are noticeable and rapid. Move residual makes it easy to visually observe the cranes lacking behind in production. Additionally, when a problem on a crane is fixed it is quickly noticed from the move residual visualization. Additionally, if the residual trend line is straight, it is an indication that the crane has not completed any moves. Therefore, if the residual is not changing it can be an indication of crane down-situation. However, crane down-situations are observed by other methods than the move residual. It is more likely that a straight line is indicative of no work orders rather than crane down-situation.

Move residual is also impervious to crane path optimization. Therefore, if crane paths are optimized in a better way, it would not show in move residual. Cranes would travel shorter distances during the same moves and the predicted cycle time would be lower. Thus, the move residual remains the same. Nevertheless, it is better that the move residual is impervious to path optimization. If the move residual would include path optimization The path optimizations for the cranes can be monitored and calculated differently.

Focus of the paper is to predict the cycle times using the ordinary least squares-regression. The cycle times could also be predicted using other machine learning methods. Even more accurate prediction results might be gotten using the other methods. In this application however, more weight is put on the ease of interpretability. Slightly less accurate model is more credible if it is easier to understand. Indubitably, ordinary least squares-model must provide accurate and dependable results. All things considered; move residual is a completely new concept to port machinery industry. Thus, the most viable approach is to convey the concept with methods that are easiest to understand and maintain.

The focus will be on the most common and widely known methods when choosing the most relevant ways to measure the performance of ordinary least squares. Statistics and machine learning are relatively low utilised concepts in the domain of port cranes. Therefore, the most valid methods to convey the results are the ones that are the most common and understandable. The most import metrics, RMSE, MAE and R^2, are for measuring the performance of the linear model.

The data has some limitations. There are some pieces information which are not present in the data. The missing information is regarding the TOS-commands. The cranes are sometimes commanded by TOS to opposite direction of their pick- or place-destination. This is to avoid deadlocks. The deadlocks can be caused by two cranes having work order which requires other crane to give way. There is no possibility to extrapolate this information afterwards. However, the deadlock avoidance and other limitations do not seem to have a major impact on the model or the end-results.

The data-analysis and machine learning model will be completed by utilizing Python programming language. Applicable Python packages are utilized. The main packages for data analysis are Pandas, Numpy, Scipy, Statsmodels, Matplotlib and Seaborn.

4 Implementation

4.1 Description of the data

The research data is quantitative operational data from ASC-cranes that are in production at a port. The data has been collected for several months. The sample is as large as possible for a good overview. The sampling method is non-probability. However, due to large amount of data we can argue, that the sampling method is probability sampling.

The research data consists of several tables of data:

- Job cycles
- ROS cycles
- MROS request cycles
- CMS alarms
- Landside approach cycles
- Autolanding cycles
- Software version of CCS

Job cycles includes the most important data for examination. The job cycle data consists of total 47 columns. There is total of 323 673 rows of job cycle data. The data is indexed by the job id, which comes from TOS or ECS. The job id, however, is not necessarily unique due to constraints in the TOS or ECS. The job cycle describes the time it takes from start to finish of a job cycle. The job cycle data also includes various other operational data such as the number of times the crane has requested for ROS and MROS during job cycle. ROS cycle is a planned or intentional manual intervention. A sample of the job cycles data can be observed from 0. The data table has been transposed for publication.

There are some functions that the crane cannot manage automatically and that's when the ROS is requested by the crane. The crane also occasionally encounters unexpected situations where the crane is unable proceed on its own. On these occasions, the MROS is called by the crane. Both the ROS and MROS operations are performed by an operator at the control room. Both operations are recorded on separate databases with similar information. Both operations normally take place during a job cycle. However, the crane operator can also initiate both operations manually. The ROS and MROS-data will be excluded from further examination. The job cycles data already includes the most relevant information of the ROS and MROS-operations.

CMS stands for crane monitoring system. CMS alarm data is a combination of all the alarms from all the cranes in the fleet. There are approximately 300 different alarms in the data sample. The alarm types consist of following three types: alarm, fault, and bypass. Alarm-type of alarm is a warning that does not necessarily stop the crane operations. Fault-type alarm indicates that the crane is not operational. Bypass-type alarm indicates that the crane or human protection features are bypassed. Such features are for example restricted area or circuit breaker. The data also includes the timestamp of the alarm and the duration of alarm, when applicable. There is no indication in the data how the alarms were resolved.

Autolanding and approach data contains various information of the autolanding- and approach-cycles. Both cycles are part of a job cycle and are meant for gathering more detailed data during the job cycle. Approach cycle means the phase when the crane is approaching a truck at landside using gantry. Truck location is scanned constantly during the gantry movement using lidars. Approach cycle data contains all the necessary destination information to approach the truck at a right row and bay. Approach cycle is used for both pick- and place-jobs at landside. After the approach cycle is completed, autolanding cycle begins and the spreader is lowered. Both autolanding- and approach cycles are executed during a job cycle. Both cycles are used only at landside. Move jobs between stack-to-stack or waterside-stack do not have a separate approach or autolanding cycles. We will exclude both autolanding- and approach data from further examination as it describes only the landside-operations. The focus of this paper is the crane performance overall. Neither autolanding nor approach has information that could further enhance the job cycles data.

4.2 Data preparation

To proceed, the data must be prepared and filtered for the regression. First, enriching of the data will be done. Categorical dummy variables of different move types will be prepared for the regression. The job cycles-data will be used as a basis for the research. This data table represents the best overall picture of the daily crane operations. Information from the other tables will be joined into job cycles-data as required.

Filtering will also be applied to the data. There are rows that have inconsistencies in the data. All the row amounts given below are based on comparison to the original data shape. Some of the rows have multiple inconsistencies. Therefore, the row amounts cannot be summed together as number of rows dropped from the original dataframe.

4.2.1 Enriching the data

The job cycles consist of different types of moves. There are five different types of moves. The moves-types are:

- Stack to stack
- Waterside to stack
- Stack to waterside
- Trucklane to stack
- Stack to trucklane

It is assumed, that the move type affects the job cycle length. Stack to stack-movement is completely automatic. In a different type of move from stack to trucklane, human intervention is required to confirm that the move completed successfully. It can also be presumed that the stack-to-stack movement takes shorter time than lifting a container from truck to stack. However, the differences between the move categories must be examined further. The move type exists in the job cycles data, but it is not directly usable for linear regression. Therefore, dummy-variables will be created from the data to distinguish different kind of moves.

The travel distances of each main move are separated as empty and loaded cumulative distance during a job cycle. The loaded distance means that the crane travels with a container. Empty means that the crane travels without a container. These cumulative distances will be summed together as three new columns. The benefit of utilizing the combined travel distances would be that the regression formula would be slightly simpler. There shouldn't be differences in the move speeds of gantry or trolley whether it is loaded or empty. The hoist, however, has different speeds when loaded and empty. Later, it is verified whether it is viable to make regression out of the summed distances or if the split distances must be used.

Cranes sometimes must stop the movement during a job cycle. For example, they must wait for the area to clear of another crane. There are different kinds waiting times available in our data to subtract the waiting times from the job time. However, these waiting times don't include the acceleration and deceleration times of the moves. In these cases, it can be assumed that waiting degrades the job cycle time to some degree. Therefore, a new binary column will be created whether the crane has waited during a job cycle. This column is an input candidate for the linear regression model.

4.2.2 Job cycles filtering

Preparation of the job cycles is continued by filtering the data further. First, the rows that are missing some of the important data points are filtered. Job cycles include information regarding pick target, pick actual, place target and place actual fields. Some of the rows are missing data from at least one of these data fields. The data contains approximately 3 % of rows that are missing one of these four fields. The reason for missing information is ambiguous. For instance, it is known that sometimes moves are completed without an input from TOS. Such moves are completed by the ROS-operator and not by the crane automation.

It is good to examine the filtered data points more precisely. From the rows it can be examined that sometimes a container is lifted from truck onto trucklane. Containers are normally not lifted from truck onto trucklane. These kinds of moves have also been prepared differently in TOS since the place-target is not the trucklane. The consecutive move on the same crane is the picking the same containers from the trucklane. Judging by the change in the container weight, the container is then emptied on the trucklane. Currently empty container is finally lifted from the trucklane into the stack. These movements were recorded in two rows of data in the job cycles. They are good examples of moves that do not represent normal operations of the crane.

Many of the job cycles that are missing job information also have a job time very low compared to distribution of cycles times in rest of the data. In many cases the cycle time is so low, few seconds. The job cycle simply cannot be completed in that time. Therefore, they cannot be considered feasible. It can be concluded that all the placement data must be available for further utilization of the job cycle data as training data.

The count and the distribution of the job cycles with one of the placement fields empty is be plotted in Figure 9. The number of cycles with missing job information is not high. However, they can significantly affect the results with residuals. Therefore, these rows will be filtered out and they will not be re-introduced back into the data when the model is fitted with whole dataset.

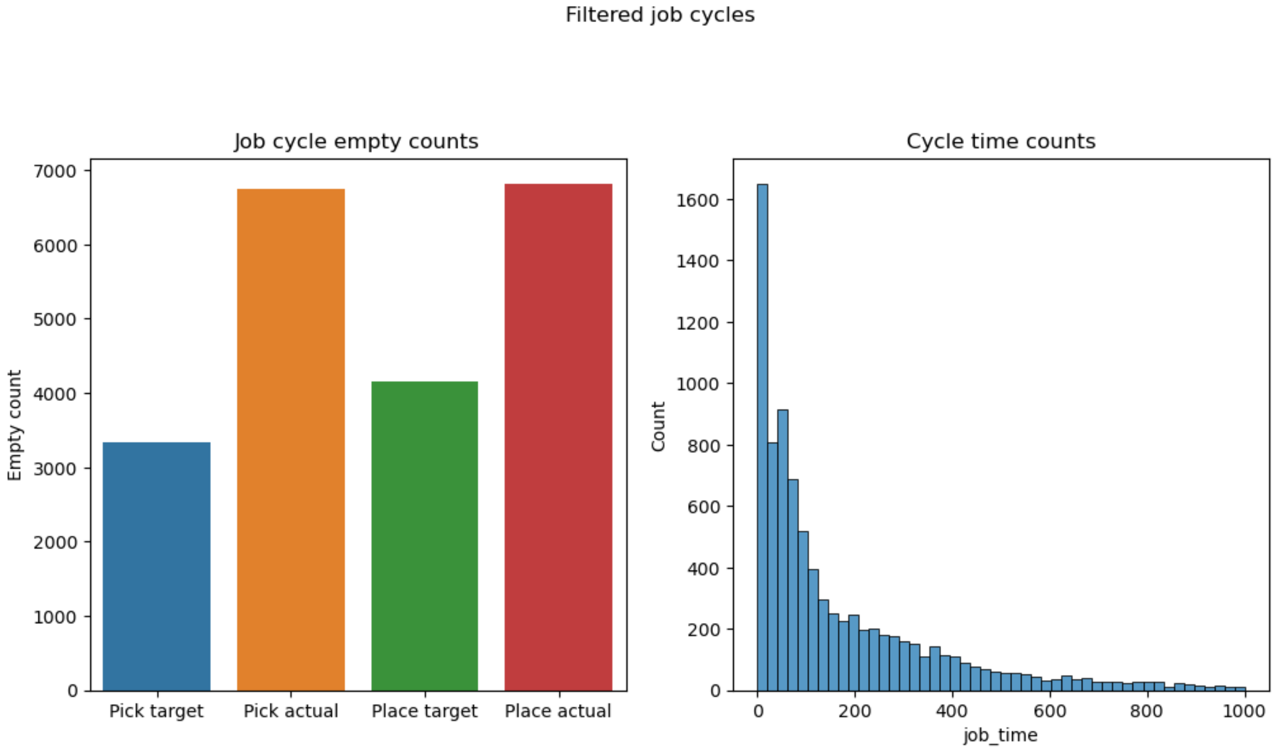

Figure 9. Job cycle empty counts

In addition to missing information, the job cycles data also has rows where the pick target is not matching to the pick actual field. Same can be observed from place target and place actual fields. It might be an indication that TOS inventory is not completely consistent, or the crane has encountered a problem executing the work order. There are approximately 5 000 rows of data with inconsistent placement fields. These rows seem to have high levels of manual interventions. The job cycles also have exceedingly high average cycle times, 1 176 seconds. Many of the rows also have an invalid container id. This could be a sign that the handled containers are worn, and the cranes have problems handling them. However, it cannot be confirmed for sure what has gone wrong with these moves. However, it is sufficient to conclude, that something has gone wrong. Therefore, these rows will now be removed from the training data.

These rows have different characteristics compared to the previous ones with missing row information. The rows with missing information had signs indicating inconsistent port operations. However, the rows with different planned and actual pick and place targets can possibly be correlated with bad crane operations. Thus, these rows will be reintroduced back into the data when calculating the residuals.

4.2.3 CMS cycles

Occasionally, alarms happen during crane operations. The alarms can degrade the performance of an ASC. However, it can also be an information to the operator and not affect the crane performance.

First, the cumulative time of each alarm will be inspected. The cumulative time of each alarm will be calculated in Appendix 2. The two largest times are the gantry leg discrepancy faults. These are false alarms. The alarm generation method has since been fixed but they exist in the research data. Even though they are faults they do not degrade crane performance in the research data.

Largest cumulative alarm time after the leg discrepancy faults is the yard light. The light curtain's purpose is to protect personnel and stop the crane if it has tripped. It does not represent the performance of the crane in any way. However, if the light curtain is tripped, and the crane is at landside, the crane will stop, and it will affect the performance. However, there is no direct correlation between the crane performance. The amount of time light curtain alarm has been on is not the time it has influenced the crane.

Second largest cumulative alarm time is the mros request. Mros will be inspected more closely in chapter 4.2.4. The remaining alarms will now be examined. There are different kinds of alarms remaining. Some of the alarms will affect the crane performance and others do not. There is no possibility to make definite conclusions of the alarm cumulative times.

Alarms will be further inspected by counting the occurrence of individual alarms. The counts will be calculated in Appendix 3. The data is from all the cranes. The discrepancy alarm is more prevalent with the alarm counts. The rest of the alarms will now be observed. It can be concluded that there is no possibility to make definite conclusions of alarm counts either. The number of alarms is not directly correlating with crane performance.

Gantry leg discrepancy fault alarm count represents over half of the CMS data. Additionally, there is no way of detecting false leg discrepancy faults from the real ones. Therefore, this alarm is best filtered out since it is not degrading the crane performance. It will be removed from the data completely.

It can be concluded that neither alarm cumulative time nor alarm count is giving large insights of the crane performance in general. However, the faults will be utilized for the purposes of filtering out bad job cycles. It will be assumed that faults effect the crane performance. The pre-filtered CMS alarms data will now be joined into the job cycles dataframe. From the CMS alarms, only fault-type of alarms will be joined. Bypass and alarm type of alarm are presumed not to influence the performance of the cranes.

Job cycles dataframe has now been joined with the CMS faults. Thus, the effect of fault can by inspected by calculating the mean cycle time with and without a fault. The results can be observed from Appendix 7. The mean cycle time is close to twice as much when there is a fault during the job cycle compared to a normal job cycle. Therefore, it is now confirmed that a fault correlates with a bad job cycle. The linear regression model will only be trained with good job cycles. As a result, the rows with faults will be removed from the dataframe. The row amount removed from the dataframe is approximately 7 000 rows. These rows will be reintroduced back into the dataframe when calculating the move residuals with trained model.

4.2.4 MROS

There are additional rows in the data that can be left out from the training data. The job hasn't completed without problems when an mros call has been made. It can be assumed that the job cycle takes longer time if the mros is called. By the time the crane asks for mros, it has already tried to complete the job cycle on its own. After the mros request has been made, additional time has already been spent. Additionally, it takes some time for the operator to orient himself to the situation crane is in.

The data supports the assumption that the cycles with mros-call take longer time. The mean cycle time with and without mros request will be calculated. The results can be observed from Appendix 8. Job cycles with mros-call take approximately twice the amount of time than without mros-call. The mean cycle time with mros call and fault is nearly the same. The reason being that an mros call is also considered as a fault. The rows with mros call will now be removed from the training data. The rows with mros calls will be later used when calculating the move residuals.

4.2.5 Crane retries

The crane will retry placing or picking up the container a few times. If unsuccessful, the crane will request for mros. The rows that have pick or retry count more than one will be filtered out. This will only consist of 1 000 rows of data. It's a small number of rows that doesn't add much explanatory value to our model. This data will be later used for the whole model when calculating the move residuals.

4.2.6 Cycle time filtering

There are also moves with high cycle time. Highest cycle time in the dataframe is 75 883 seconds which equals to approximately 21 hours. It is known from domain experience that this kind of cycle time is completely unreasonable. By observing the CMS alarms of this cycle, it can be confirmed that there is an alarm type of alarm of the ACM during the cycle. Therefore, it can be assumed that there was a fault with ACM during the job cycle. The repair job must have been done during the cycle.

Second highest cycle time is 18 357 seconds which equals to approximately 5 hours. In contrast to the previous one, this move doesn't have any alarms, mros-calls or any other variable that would explain this high cycle time. This move is a good example what kind of moves the move residual is useful for. There must be a valid explanation for this highly irregular cycle time. However, the explanation does not exist in the data. It can be assumed that there are numerous other moves in the data with same kind of irregular cycle time.

Continuing with the cycle time, unfiltered job time histogram plot can be observed from Appendix 4. The plot was done using the histplot method of Seaborn (*Seaborn.Histplot*, n.d.). The data seems to be skewed right and not normally distributed. The data has characteristics of lognormal distribution. Right-tail of the distribution plot was limited to 1 000 seconds to compensate for the high cycle times of some cycles. The data has one main peak, prominent secondary peak and one minor peak. It is likely that these peaks represent different types of job cycles. There seems to be invalid data points on the lower end of cycle times. In contrast to what can be observed from the plot, it can be assumed that the normal job cycles follow normal distribution. It seems that the long right tail of the data is caused by the non-normal job cycles of the crane.

Job cycle times must be investigated further. It is confirmed that a fixed filter for the job cycles times must be implemented to account for the extreme outliers in the data. First, the distributions of different kind of job cycles are visually confirmed. To verify the distributions, kernel density plot was made in Appendix 5. The plot was made using the kdeplot method of Seaborn library (*Seaborn.Kdeplot*, n.d.). The colors of the plot represent different types of moves.

There aren't any major differences in the distributions of different kinds of moves. There are, however, some differences at the highest densities of the move types. The most frequent type of move is clearly stack-to-stack -type of move. However, there aren't so major differences in the cycle times in different types of moves. Therefore, same fixed filter value will be utilized to cycle time for all move types. The fixed filter will not filter the highest densities of any move type. Thus, no valuable training data will be lost during filtering for any of the move type.

To proceed, correlation between travelled distance of the main moves and cycle time will be visually verified. This can be observed from histogram heatmap in Appendix 6. The plot was prepared by using Seaborn histplot method (*Seaborn.Histplot*, n.d.). Stronger color represents higher number of moves at that point. In contrast, faded color means that their frequency of correlating move distance and cycle time is not so prominent. The plot has been prepared so that the previously discussed filters have already been applied. White area means there isn't correlating move distance and cycle times.

From Appendix 6, especially high correlation between travelled gantry distance and the cycle time can be observed. There is also seems to be fair correlation between hoist travel distance to cycle time and trolley travel distance to cycle time. It can be concluded that the previously mentioned remarkably high cycle times are extreme outliers. The graph was limited to a cycle time of 500 seconds for better visual representation. It is also important to note, that even after applying the previous filtering, outliers seem to be prominent when inspecting the gantry distance-subplot.

There is now enough information gathered to decide a fixed limit for filtering with cycle times. It can be concluded that the cycle times over 500 seconds are not feasible. This fixed limit is conservative and keeps plethora of bad moves in the dataframe. The limit could be adjusted lower. However, there are better methods to further refine the data. It is best practice to remove these now using the domain knowledge, so they do not interfere later methods. Job cycles with cycle time over 500 seconds will be removed from dataframe which is used to train the model. These rows will be reintroduced back when using the model to calculate move residuals.

4.2.7 Variable correlation

Before beginning with regression, the correlation of independent variables to dependent variable will be verified. A pairwise Pearson and Spearman correlation will be calculated using pandas corr method (*Pandas.DataFrame.Corr*, n.d.). The results of Pearson correlation can be observed from Appendix 9 and Spearman-correlation from Figure 10. As there are still outliers in the data, Pearson correlation coefficients result values are slightly lower than the Spearman-correlation. Spearman-correlation is performing slightly better as it is not giving so much weight to the outliers.

It can be observed that the highest correlation with the job time is gantry time. This is in line with domain-knowledge of the ASC. There are also modest differences of correlation for both loaded and empty movement distances. The correlation-plot already includes the most important independent variables for our regression. The initial variables are main moves and the move-type. The Spearman correlation alone is not adequate metric to decide whether the main moves should be separated as empty and loaded or combined as total distance travelled. Both Pearson and Spearman correlation is showing relatively high correlation between the main moves and cycle time. Clearly weakest correlation out of three is the hoist movement. However, hoist movement is important to include as a regression variable. Thus, we can confirm that since both correlation-calculations have adequate correlations, the correlation holds (Rovetta, 2020, p. 6).

The container weight does not seem to have any correlation to the job time. It can therefore be excluded from candidates of input variables. All the time-variables can also be removed from the input candidates. The use of time would be problematic as the predicted variable is time on its own. The move times can accumulate simultaneously since the main moves can all move at the same time. The crane can encounter situations where the speed is limited but the timer is still running.

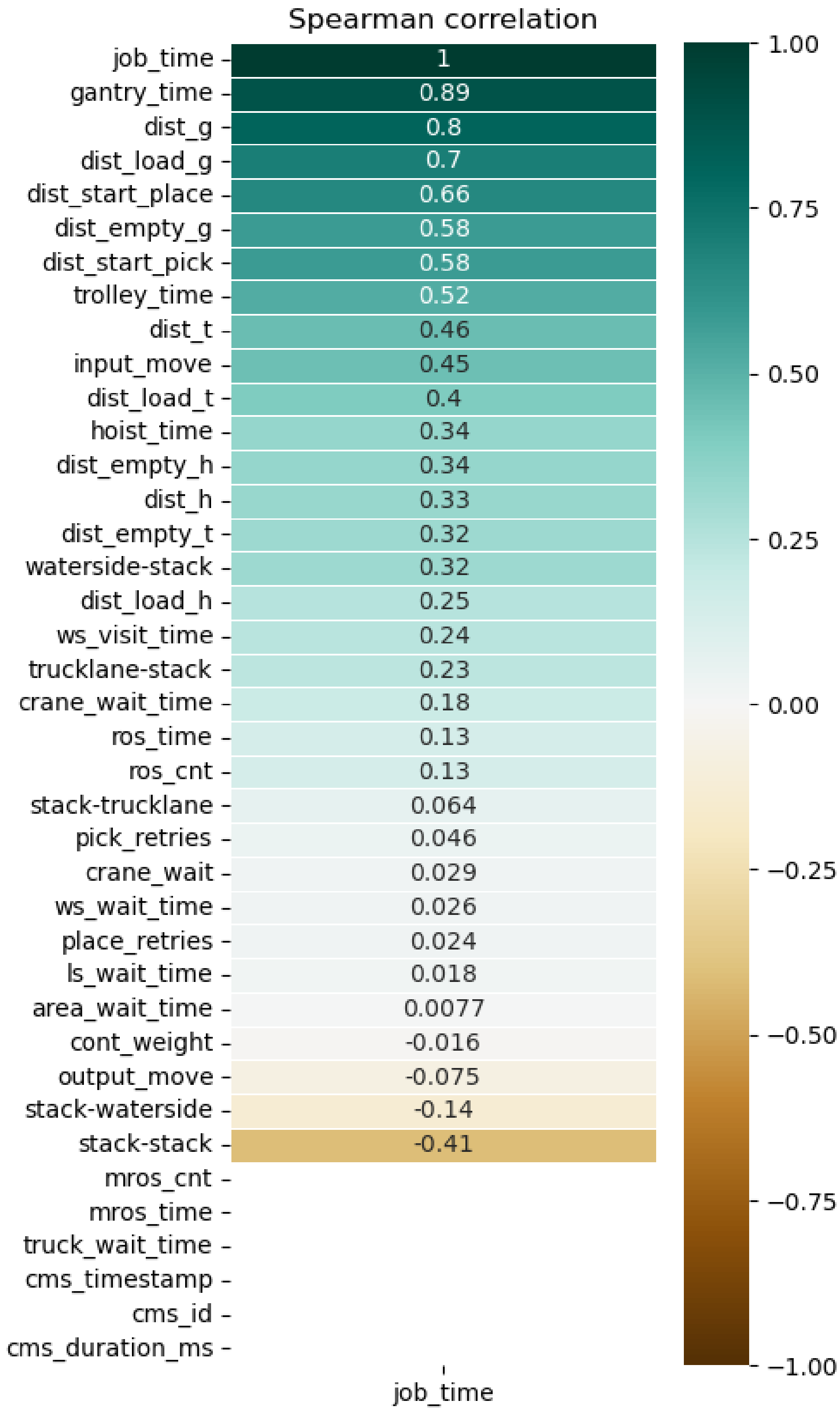

Figure 10. Spearman correlation of job time

The previous assumption of negative-correlation with stack-stack-moves is also confirmed from Figure 10. In addition, there is also fair correlation with the stack-to-waterside-moves. Strongest positive correlation of the move-times is with waterside-stack -type of moves. Based on the knowledge of ASC-cranes, there is no reason for the waterside-stack -type of move to be affecting cycle time so heavily.

From Figure 10, it was observed that the most significant variable for the cycle time, excluding the time variables, is the gantry distance. Therefore, it is good to verify the average distances the gantry is moving during different types of moves. The travel distances will be calculated in Figure 11. From the figure it can be concluded that there is a valid reason why the waterside-to-stack -type of moves are having such effect on the cycle times. The gantry is by far moving the greatest distance in this type of moves.

From Figure 11, a cause for the stack-to-stack negative correlation with the cycle time can be concluded. The plot has been made with Seaborn barplot method (*Seaborn.Barplot*, n.d.). Stack-to-stack type of movements have, on average, shortest gantry travel distances. The most significant correlating variable with cycle time is the gantry travel distance. Thus, it can be concluded that even though there is a negative correlation between stack-to-stack -type of movements with the cycle time, there is not necessarily causation.

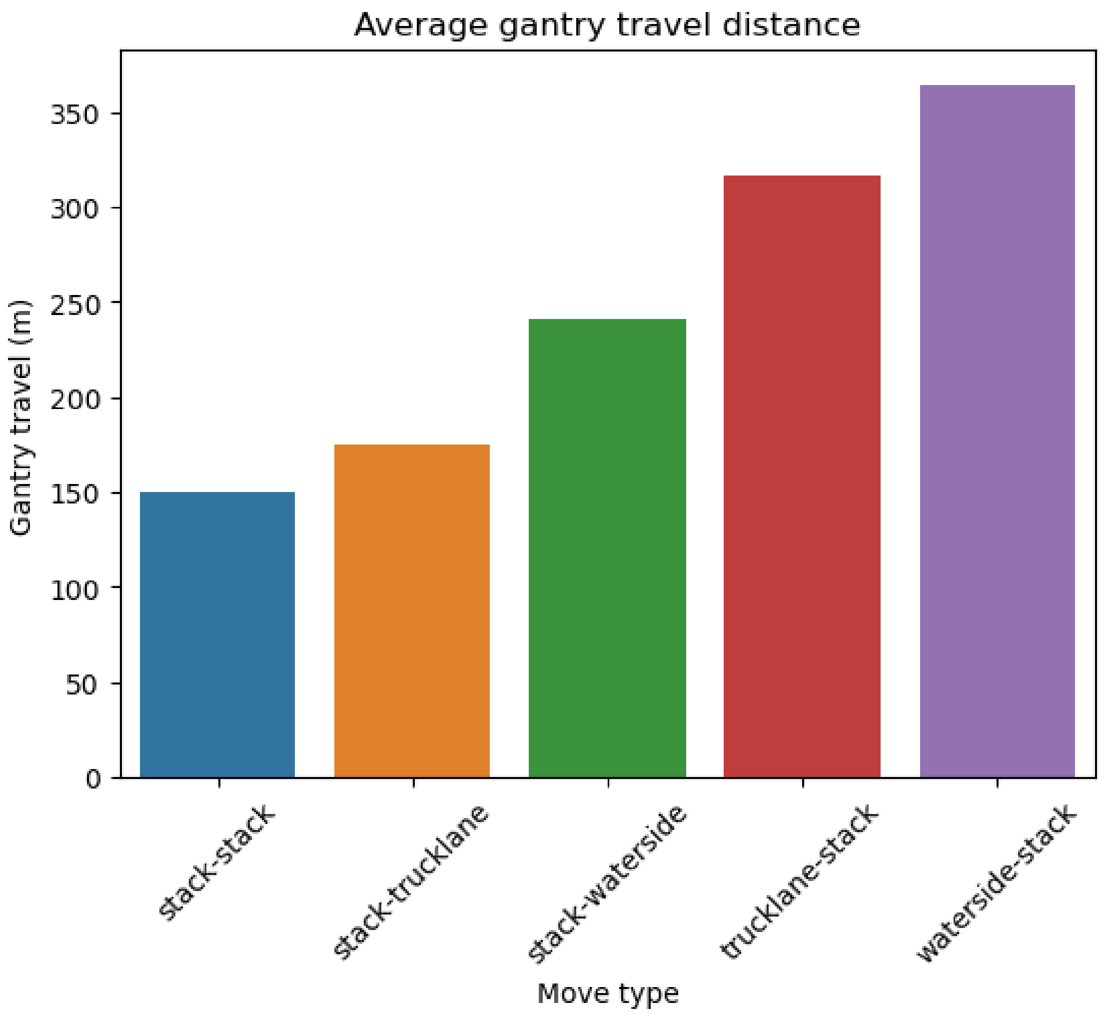

Figure 11. Average gantry travel distance

4.2.8 Preparation results

There remain approximately 270 000 rows of training data after the initial filtering the most obvious bad moves from the dataset. It is likely that some of the filtered job cycles had normal characteristics even with the faults and mros-calls. Nevertheless, there is still enough data to continue by training the linear model. Therefore, the linear model will be trained without the filtered job cycles. Thus, finding good training candidates out of the removed rows is not worthwhile. Row amount has increased from 54 to 65 after enriching the data.

4.3 Predicting cycle times

Now the linear regression model for the cranes will be prepared. Process described by Montgomery et al. will be followed (2012, p. 11). The process will be repeated few times. During the process, methods of quality control shall be utilized. Training of the regression will be performed using the filtered dataframe. The dataframe will be further filtered during the process.

Python library Statsmodels will be used for estimating the ordinary least squares linear model (*Statsmodels.Regression.Linear_model.OLS*, n.d.). Statsmodels will also calculate multitude of quality-related information for the regression model. Python library Sklearn will be used for calculating RMSE and MAE.

4.3.1 Initial linear model

The regression model building process will be started with simple linear model to see the initial results. This initial regression model will only include 20 random samples from one crane. For this initial model only moves waterside-to-stack shall be utilized. The input for the regression will be based only on three independent variables. The variables are distances in meters that gantry, hoist, and trolley have travelled during a job cycle. The initial model is simple and filtered down for initial impressions of the problem setting.

The results of the regression can be observed from Appendix 10 which is made with Seaborn regplot method (*Seaborn.Regplot*, n.d.). The plot represents actual vs fitted cycle time of the 20 samples. The results are promising. RMSE is 9 seconds. For an ASC crane, predicted error of 9 seconds in cycle time is satisfactory. Models R-squared is 0,96. It means that the model can explain variance of the independent variable very well. It can also be observed that the residuals of the regression are not skewed in any direction. This is a good starting point to continue evaluating the model by utilizing whole dataset.

The regression training will now be repeated for the whole filtered dataset. This time the data will be for all the cranes and all types of moves. Move types will be inputted as categorical dummy variables to the model. When training a linear model with categorical dummy-variables, there must be one dummy variable that is left out of the model as a base category (Siegel, 2016, p. 396;

Young, 2017, p. 143). One dummy variable will be used as a baseline-category and including all would impose problems with multicollinearity. The baseline category can be chosen freely (Siegel, 2016, p. 396). However, baseline category should be chosen as a point for comparison for the rest of the dummy variables. Therefore, base category used for the regression is stack-to-stack type of moves. The distances of the main moves will also be split between empty and loaded distances, instead of summed values.

The regression results can be observed from Appendix 11. The plot is made with Statsmodels OLS regression results summary method (*Statsmodels.Regression.Linear_model.OLSResults.Summary*, n.d.). The results are poor compared to the previous model. The models R-squared has dropped to 0,616. MAE for the model is 28 seconds. Calculated RMSE is 1774. High RMSE compared to a much lower MAE is an indication of outliers in model. The reason being that since with RMSE the errors are squared. Therefore, big residual errors lead to high RMSE. It can also be observed that there are meaningful differences between empty and loaded coefficients of the main moves. However, there seems to be strong multicollinearity in the model. The condition number is larger than the advised 30 for strong multicollinearity (J. H. Kim, 2019, p. 559; Young, 2017, p. 113). Thus, it is advised to not make any conclusions of the coefficients (Siegel, 2016, p. 374).

The initial plan was to use stepwise search to verify the input variables. However, due to multicollinearity, stepwise search cannot be used. The results of stepwise search depend on the p-values of the input variables. Since multicollinearity exists, the p-values of the input variables cannot be relied upon (Siegel, 2016, p. 374). Only one high p-value over 0.05 can be observed with the categorical dummy-variable, waterside-stack movement. However, definite conclusions of the variable will not be made at this stage because of the problems in the model and because the unreliable p-values.

The problems in the model can be further inspected by observing the versus fits plot in Figure 12. It is also called the diagnostic plot and it is common method for troubleshooting a multiple regression (Siegel, 2016, p. 382). The versus fits plot has blue dots indicating fitted value and the resulting residual. In red colour there is a local regression line associated with the plotted values. The local regression line was calculated using Seaborn lowess function of the residplot method (*Sea-*

born.Residplot, n.d.). Lowess stands for locally weighted scatterplot smoother which is often interchangeably used with term loess (Young, 2017, p. 316). Plot has 1 000 randomly selected samples from the model.

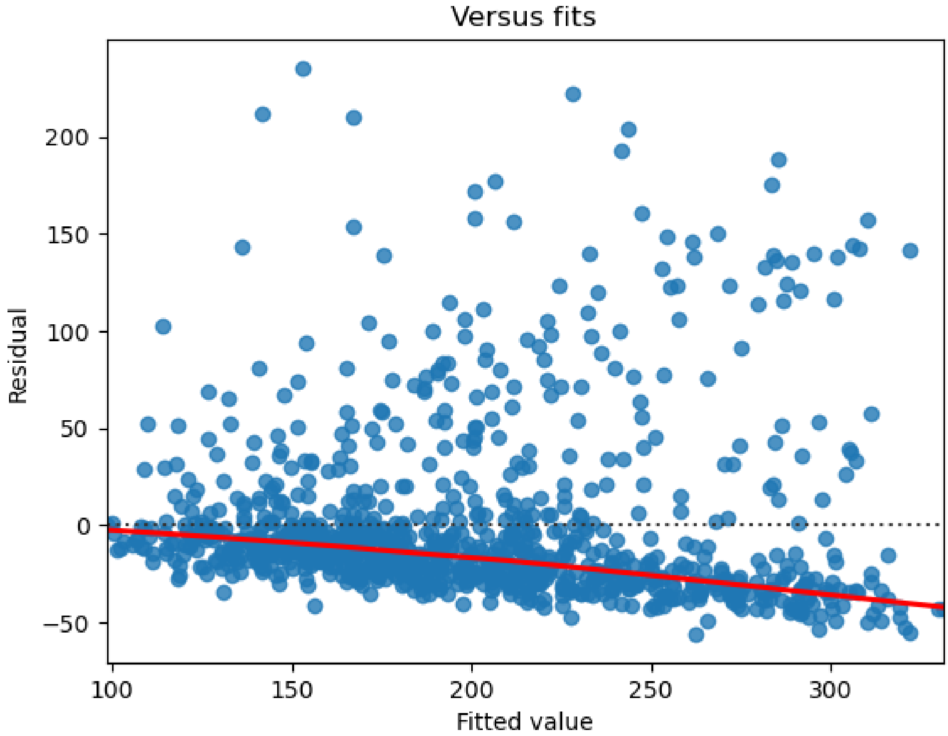

Figure 12. Residual versus fitted plot

There is a clear main group visible near the regression line. It can be observed that there are only positive outliers in our data. All the sparsely scattered residuals are on the positive side. It is also important to note that the sparse residuals are along whole X-axis. It means that there is unexplained variance in the cycle times throughout the fitted values.

The model is also predicting too long cycle times since the residual is generally negative. The error grows larger as the fitted value increases. This was to be expected. Appendix 6 of gantry travel distance versus cycle time must be reviewed. The same unexplained variance in the cycle times was visible in Appendix 6 as well. It means that there are job cycles where the cycle time cannot be reliably predicted by the input variables. Siegel advises not to intervene with the regression unless there a clear and definite problem is shown in the residual plot (2016, p. 382). However, a clear and definite problem is shown in the plot. It can be concluded that the current model is insufficient and further filtering to the data must be done.

4.3.2 Further filtering

The outliers are a problem for the ordinary least squares regression. The cost method of linear regression, SSE, is putting heavy weight on the outliers. The outliers are abundant in the data, as was previously observed from Figure 12 and Appendix 6. The method of linear regression could be exchanged for example to robust regression. Robust regression is less sensitive to outliers than ordinary least squares due to Huber-loss function. Another example when robust regression is often used, is when the data is suspected to be heteroscedastic. It was previously observed from Appendix 6, that the data is homoscedastic. The outliers are abundant at both small and large travel distances of the main moves.

The purpose of this work is detecting the ASCs with differing performance, and, in this case, using the regression residuals. The target is to predict normal and uneventful moves well with the regression. Including the outliers does not bring any value to the model. On the contrary, the regression model performs worse if it explains the bad moves in any way. All the linear regression methods put some weight to the outlier calculation. It is therefore better to remove the unexplained variance completely from the model. Ordinary least squares will be kept as the regression model.

Few of the moves with large positive residuals were manually inspected from the dataframe. These are the moves that are skewing the model. There does not seem to be any common explanation in the data for the large residuals. Large residuals are visible on all cranes and on all types of moves. Although with such a small review it is too early to make any definitive conclusions. Therefore, it is the best approach to apply statistical methods to the model. The kind of methods that take input variables into account are the most suitable for the model.

Two different diagnostics measures will be evaluated to remove the outliers from the data. The compared methods are Cook's distance and DFFITS. Both will be calculated by Statsmodels OLS Influence-method (*Statsmodels.Stats.Outliers_influence.OLSInfluence*, n.d.). The comparison will be started by using the general guidelines of the cutoff-values. Final cutoff-formula for both methods will be acquired visually. The plotted data amount is limited to 2500 rows for better visual representation. At first, Cook's distance will be calculated and checked.

There are guidelines for choosing the cutoff value with Cook's distance. One common rule of thumb is that observations with Cook's distance between 0.5 and 1.0 is a good threshold for an outlier (Frost, 2019, p. 233). However, there is a problem when there is a large number of observations. Cook's distance for each individual observation is dependent on the number of observations. With large number of observations, the Cook's distances become small. This is because Cook's distance measures the effect of removing individual observation from the model. Therefore, the effect becomes smaller when there are larger number of observations.

Another proposed method for choosing the cutoff value for Cook's distance is using formula 4/n, where n is the number of observations (Van Der Meer et al., 2010, p. 175). However, it is advised in many sources that the cutoff value depends on the application. Therefore, few different cutoff values will be evaluated for the model. The different values will be visually assessed by using plots of the data. After evaluating with different values, a suitable Cook's distance was found. The result can be observed from Appendix 12. The plot was made using Seaborn scatterplot method (*Seaborn.Scatterplot*, n.d.). Cutoff candidates are on darker hue. Observations with smaller value than cutoff are on lighter hue. In the plot there is also a hand-drawn red ellipse.

The most suitable cutoff value for this application became with formula 0,1/n. N is the number of observations. Therefore, for this plot, the cutoff value is 0,00004. Observations greater than the cutoff-value are candidates for removal. This is an aggressive cutoff-value. However, a clear pattern can be observed from the plot. The density of observations is clearly higher inside the red ellipse than on the outside. The observations inside the red ellipse are assumed to be the normal and uneventful moves. It can also be observed that the outliers are skewing our regression. The observations inside the red ellipse are predicted to have higher cycle time than they had.

However, Cook's distance does not capture all the moves inside highlighted by the red ellipse. Cook's distance has problems with many moves where the cycle time was shorter than predicted. This could be resolved by keeping the observations where the regression residual is less than zero. It means that moves that were completed faster than predicted would be kept. Appendix 6 will be reviewed and from there it can be confirmed that there is not negative outliers in the data. It means that keeping the negative residuals and utilizing the Cook's distance is viable option.

However, the DFFITS-method will be evaluated and compared against the Cook's distance. A common proposed formula for DFFITS-cutoff is 2*sqrt(k / n) (Belsley, 2004, p. 237; Frost, 2019, p. 233; *Identifying Influential Data Points*, n.d.; Montgomery et al., 2012, p. 218). In the formula, n is the number of observations and k is the number of regression parameters. It must be noted that this formula is also just a guideline. Actual cutoff value must be chosen for the application (Montgomery et al., 2012, p. 218). Different cutoff-values will again be assessed, and the results visually observed.

After evaluating different formulas with the data, a suitable cutoff-formula for the data was found. The resulting formula is sqrt(k/n)/5. This is ten times smaller than the rule-of-thumb. Like the Cook's distance, smaller DFFITS values represents a smaller outlier-score. The results can be visually inspected from Appendix 13. The plot was made using Seaborn scatterplot method (*Seaborn.Scatterplot*, n.d.). The removal-candidates are marked with a cross. Hue of the observations is indicating DFFITS-value. DFFITS identifies the inlier moves inside the same red ellipse better than Cook's distance. Additionally, DFFITS does not require any additional conditions on regression residuals for removal-candidates. Therefore, DFFITS will now be utilized to remove the outliers from the training data.

DFFITS is a method that is based on the influence of the observations. However, influential cases are not necessarily outliers and outliers are not necessarily influential cases (Nieuwenhuis et al., 2012, p. 38). However, it can be observed that these influential observations are heavily skewing the regression model. It is possible that within the removal candidates there are influential observations that are not outliers. Nevertheless, there is small number of methods to verify influential onliers. Even if they could be detected it is likely that they would not have major effect on the linear model. There can also be outliers that are not influential within the data that was not captured by DFFITS. However, the dataframe is large and the influence is small. Thus, it is unlikely that the low influence outliers would have large effect on the model.

The number of rows removed is approximately 64 000. Therefore, in every fourth move, there is unexplained variance that the model cannot predict accurately. The removed rows will be reintroduced back when fitting the model.

4.3.3 Quality assurance

There are approximately 206 000 rows left in the dataframe. It means that one thirds of the moves have been removed compared to the original dataframe. Original dataframe will now be compared to the filtered dataframe. This will be done as a method of quality assurance to verify what types of moves have been removed.

The results can be observed from Figure 13. The figure is presenting the proportions of move types. It can be verified from the figure that the original dataframe is nearly matching the filtered one. Biggest difference between the two is in stack-to-trucklane type of moves. The amount has clearly decreased in the filtered dataframe. Additionally, the trucklane-to-stack type of moves have slightly decreased. This was to be expected. The most challenging moves for the cranes are related to the truck lane. Therefore, it can be concluded that there was more unexplained variance in truck lane-related moves. These rows were removed manually using the domain knowledge and the DFFITS-method.

Conclusion can be made, that the abnormal moves happen on all types of moves. They are most frequently existing in trucklane related moves. However, the difference is small. Therefore, there is no reason to inspect any move type further.

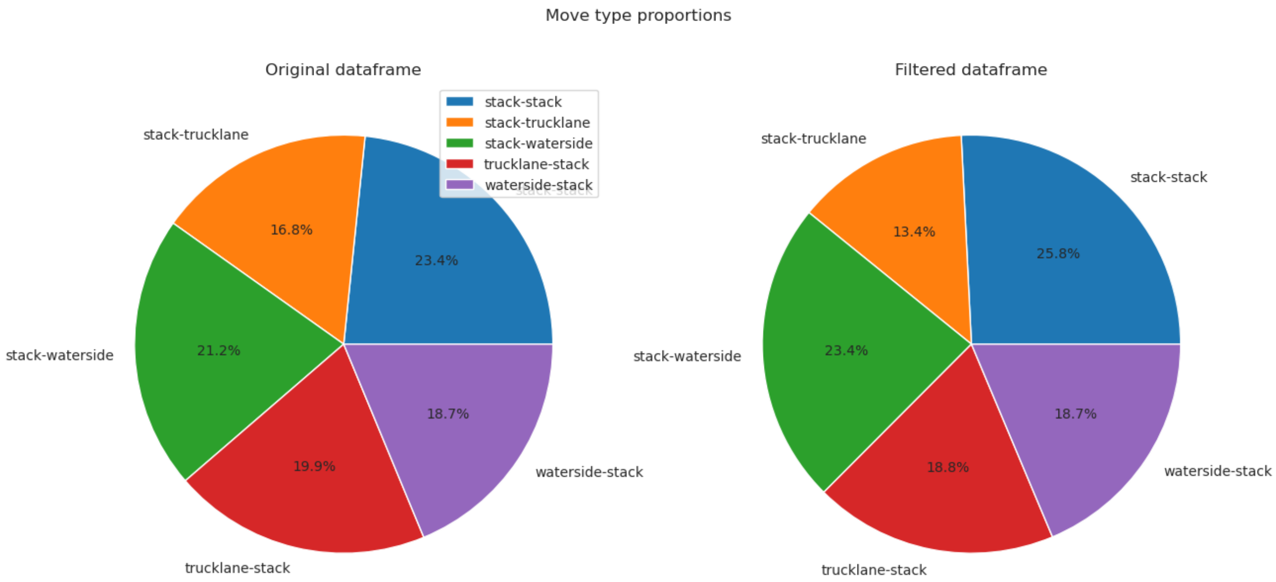

Figure 13. Move type proportions

4.3.4 Second linear model

The model will now be trained by utilizing the remaining rows using the ordinary least squares regression. Separated main move travel distances will used. The move type and the crane wait will be as binary independent variables. Therefore, the input variables are the same as in previous regression results. The results of regression can be inspected from Appendix 14. It is common practise to compare regression model with and without outliers (Frost, 2019, p. 234). The results confirm that the model has improved. RMSE is 10,0 and MAE is 7,6 seconds. These values are adequate for predicting the cycle time. The model also has nearly the same R-squared as with the initial model, 0,96. This high R-squared value could be an indication that the model is biased (Frost, 2019, pp. 124–125). However, high R-squared value can be expected from an ASC.

The model still has problems with multicollinearity. This is a problem if the requirement is to make conclusions out of the coefficients. The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations (Kutner, 2005, p. 283). The target is to make good predictions. Thus, multicollinearity is not problematic for the predictions themselves. However, multicollinearity is problematic for the significance levels. Currently all the variables are of the model are calculated as significant. Due to multicollinearity, conclusions of significance levels of the coefficients can't be made (Siegel, 2016, p. 401). Therefore, stepwise regression cannot be utilized for selecting the input variables since it relies on the significance levels.

However, the coefficients are mostly aligned what can be expected from an ASC. The hoist movement speed is decreased when it is loaded, and it can be observed from the coefficients. Trolley and gantry, however, should have same coefficients loaded and empty. Minor differences with the trolley coefficients can be observed. Additionally, a minor difference can be observed in gantry coefficients.

The coefficients for binary values also seem to be dependable compared to the domain-knowledge. Trucklane related moves have their move times higher compared to stack-to-stack -type of moves. It is known, that the stack-to-trucklane -type of move is the most challenging one to a crane. When an ASC picks or places a container, the placement is scanned with lasers and machine vision. The containers are standard and therefore picking a container is easy for the automation. Additionally, all the waterside vehicles are like each other, and the scanner are tuned for the specific vehicles. Therefore, waterside vehicles are easily detected by the crane. However, the trucks have the most variance. Height, width, location, and their appearance vary a lot. This is the reason why stack-to-trucklane type of moves have clearly the highest coefficients, as was to be expected. Waterside-related moves have their coefficients close to what they are in stack-to-stack-moves.

The previous regression model was trained with waiting times subtracted from the cycle time. In the regression results at Appendix 14, crane wait binary coefficient is negative. It is an indication that if the crane stopped to wait during the job cycle, cycle time is predicted to be few seconds smaller. This raises concerns regarding waiting time and binary variable credibility. Both variables will be assessed by training the linear model four different times and inspecting the errors of the regression. The comparison of the variable accuracy can be observed from Table 1.

Table 1. Regression accuracy

	RMSE	MAE

Wait time subtracted and crane wait -binary	10,0	7,6
Wait time not subtracted and crane wait -binary	10,1	7,8
Wait time subtracted and without crane wait -binary	10,0	7,7
Wait time not subtracted and without crane wait -binary	10,7	8,3

Each of the compared models would be viable for our purposes. Most importantly, it is confirmed that the wait time in our dataframe is dependable because it increases the prediction accuracy of the model. However, the binary wait time will be excluded from the model since the negative co-efficient with waiting time subtracted is not dependable. The cycle time cannot be predicted to take less time when the crane must wait. Even though it increases the accuracy of the model very slightly. The chosen method is bolded in Table 1.

4.3.5 Final linear model

Separated and combined main move travel distances will now be compared as input variables. Previously, separated travel distances of all the main moves were used. Separated travel distance is most important for the hoist since the speed is reduced with load. However, it might not be necessary to gantry and trolley. Speeds of both gantry and trolley should remain the same with and without the load.

In Table 2 is a comparison of the regression results with separated and combined the travel distances. Combined distances means that for both gantry, and trolley, the travel distances of empty and loaded are summed together. Hoist travel distance is left as separated loaded and empty distances in both models.

Table 2. Regression result comparison, travel distances

	RMSE	MAE	R^2
Separated distances	10,0	7,7	0,957
Combined distances	10,1	7,7	0,956

It is confirmed from the results, that the combined distances is performing as well as the separated distances. Additionally, since it makes the model more simple, combined distances is the better option. From now on, the combined gantry and hoist distances will be used for the regression. This will reduce the number of input variables by 2. The chosen approach is bolded in Table 2.

The accuracy of the regression must be confirmed with all the move types. The comparison results can be observed from Table 3. The model is clearly predicting stack-to-waterside -type of moves most accurately. Highest errors are with stack-to-trucklane -type of moves. However, the differences in regression accuracy are minor. Based on the results in Table 3, it can be confirmed that the linear regression model is performing adequately with all kinds of moves. Therefore, there is no reason to further refine to model with any specific move-type.

Table 3. Regression accuracy of different move types

	RMSE	MAE

Stack-to-stack	10,2	7,7
Waterside-to-stack	10,7	8,4
Stack-to-waterside	8,7	6,4
Trucklane-to-stack	10,3	7,8
Stack-to-trucklane	11,2	8,8

The linear regression accuracy will now be verified using the residuals. The residual count plot histogram can be observed from Figure 14. The figure was plotted with Seaborn histplot method (*Seaborn.Histplot*, n.d.). In an adequate linear model, regression residuals are normally distributed (Young, 2017, p. 52).

From Figure 14, it can be visually confirmed that the distribution is bell-shaped and thus, normally distributed (Frost, 2020, p. 129). There is slightly more deviation on the positive residuals, and they are distributed little bit longer than the negative residuals. However, the distribution is adequate for the regression model. The peak of the residuals is a few seconds' negative. This a sign that we were not able to filter out all the unexplained variance with DFFITS. The longer tail of positive residuals supports this hypothesis. It means that the model's fitted value on normal moves is greater than the actual value. Therefore, the model predicts few seconds longer cycle times than the actual cycle times on training data.

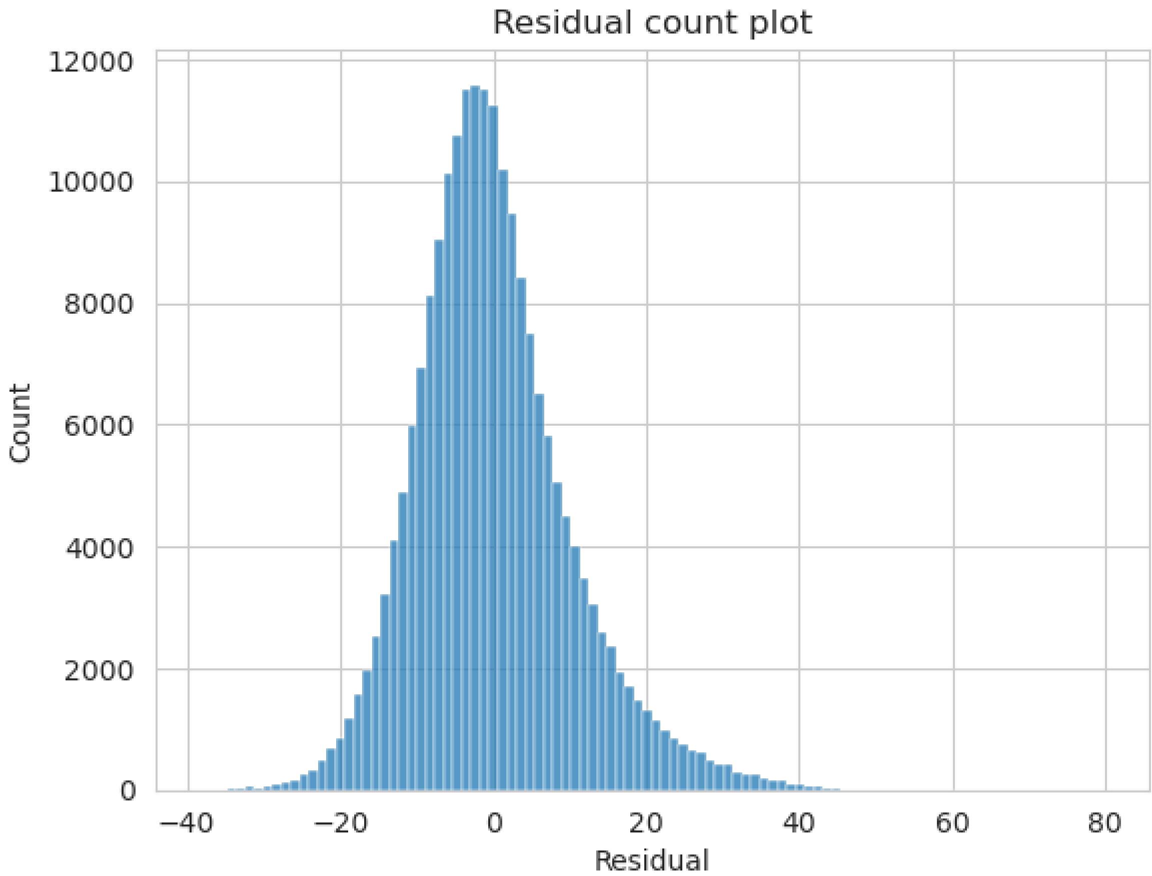

Figure 14. Residual count plot

The residuals will be visualized in two additional ways. First, a residual plot will be made which can be observed from Appendix 16. The figure is plotted with Seaborn residplot method (*Seaborn.Residplot*, n.d.) The plot was made using only 2500 observations for better visual presentation. The Individual observations are marked with a blue dot. A local regression line for the plot is in red color. One of the important properties of successful OLS-regression is homoscedasticity (Frost, 2019, p. 201; Yan & Su, 2009, p. 195). There is slightly more variance in the errors at higher cycle times. Heteroscedasticity would appear as a fanned shape in this plot. However, it can be verified from the Appendix 16 that the center of observations, where the number of observations is highest, is clearly homoscedastic.

An unbiased model has residuals that are randomly scattered around zero (Frost, 2019, p. 125). The residuals on Appendix 16 are randomly scattered. Therefore, our model is unbiased. The data in Appendix 16 is not presenting the whole dataset. However, we can conclude that the model is unbiased to training data.

The distribution of errors will be further examined by making a plot of regression actual vs fitted value. The plot will be made with Seaborn regplot method (*Seaborn.Regplot*, n.d.). The plot can be observed from Appendix 17. Perfect regression result with zero residual is marked with a red line. Blue dots represent the observations. From Appendix 17, a very minor skew in the regression model can be observed. The model is slightly predicting move times higher than the actual cycle time on the lower end of cycle times. On higher cycle times, the model is predicting very slightly lower cycle times than the actual cycle time. However, the highest density of observations follows the red line. The model is making adequate predictions at both ends of cycle time spectrum.

In addition to constant, there are eight independent variables to predict the cycle time. Four independent variables represent the travel distances of the main moves out of which the hoist is separated as empty and loaded distance. Another four independent variables represent the move type. Stack-to-stack move is used as a baseline dummy-variable for the regression model.

It could be argued that waterside related dummy-variables could be left out. They do not increase the accuracy of the model. The coefficient is the same as in the baseline move type. However, they make the model easier to interpret. The dummy-variables also give information to the observer of the differences between the move types or the lack thereof.

The average error of our model is also acceptable. In conclusion, it has been confirmed that the linear model is adequate for predicting ASC cycle times for the purpose of move residual. More detailed report of the regression results is in Appendix 15. The summary was made using Statsmodels OLS summary (*Statsmodels.Regression.Linear_model.OLSResults.Summary*, n.d.). Condition number is still large, 1.730, indicating that there are still problems with multicollinearity in the regression model. However, since we want to make accurate predictions, the multicollinearity is not a problem.

4.3.6 Initial move residuals

The finalized model will now be utilized to calculate the move residuals. First, only one crane is selected, and results presented. Then the utilization of move residual will be expanded to other cranes. Initial example of move residual is in Figure 15. The plot was made using the lineplot method in Seaborn (*Seaborn.Lineplot*, n.d.). Red dots represent individual moves, or observations, in their respective residual value on y-axis. The y-axis is inverted to indicate that lower observation is located, worse the move is. Thus, larger the residual, worse the move is. Blue line is a lineplot presenting the trend in the residuals. The time period on the x-axis is approximately 14 hours.

It can be observed that the crane has had both active and passive periods. A passive period has long time without any observations. There is one observation with large residual where the residual has been especially high, approximately 1 150 seconds. The variance of the residuals seems to increase after the one large residual.

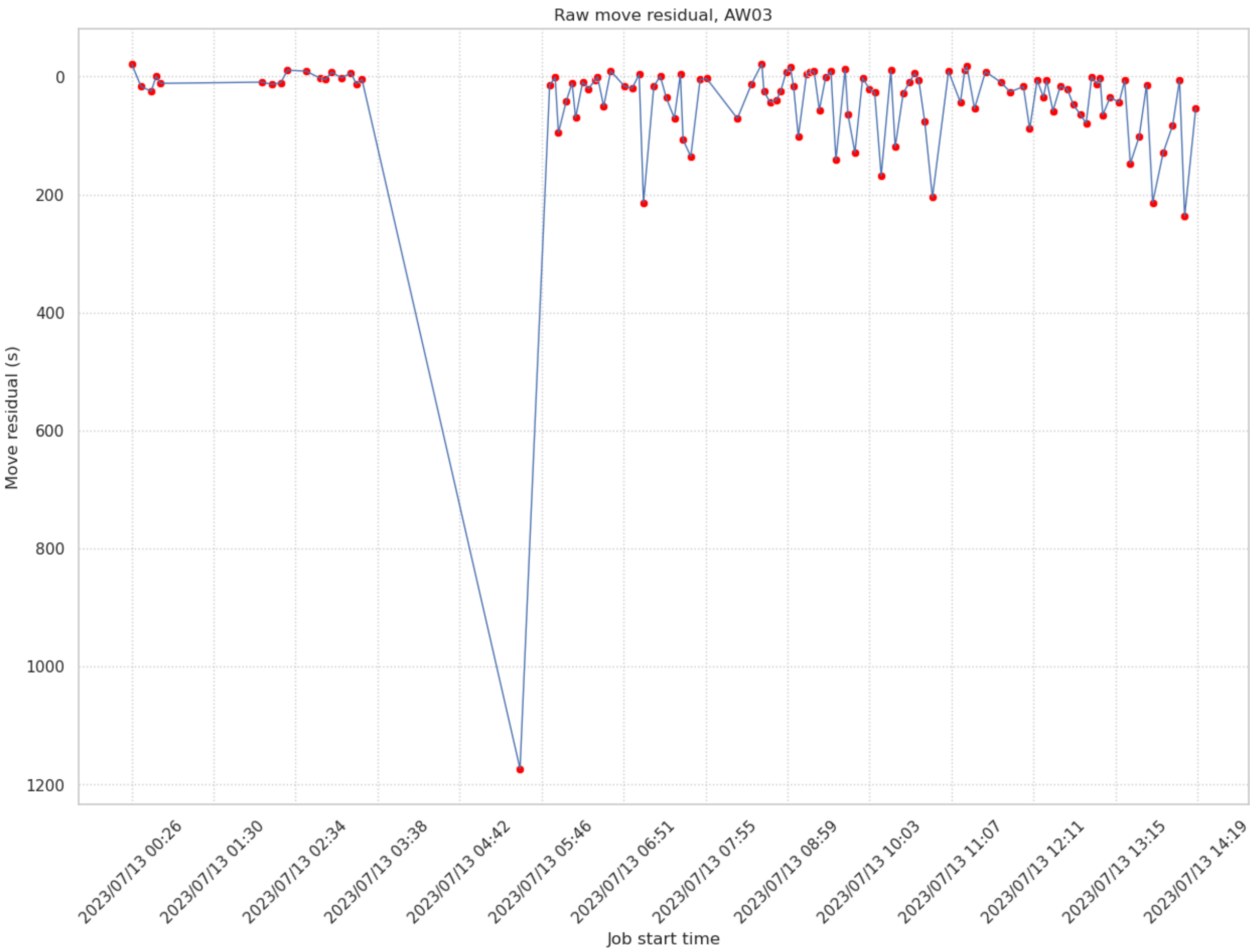

Figure 15. Raw move residual

One large residual makes the other residuals seem small. The largest residual does not necessarily mean that it is the most problematic move or something to inspect further. One bad move can be caused by numerous reasons, and it might not even be fault of the crane. What is more interesting in the plot are the points that have residuals approximately 200 seconds. This is already an exceptionally large residual for a move. When the move residuals are presented, they are in a way the scores of moves. There is no reason to give such a high score to one bad move. Otherwise, it will inflate the values of other high residuals. The highest residual recorded was for the move that took approximately 21 hours. The resulting residual was approximately 75 000 seconds. To solve the inflation, large residuals will be modified. All the residuals with value over 200 seconds will be selected and replaced with value 200. This way uniform bottom limit can be had for the residuals. The selected value was chosen by trial and error.

There is another reason for the fixed limit of 200. The timescale of 14 hours in the x-axis of Figure 15 is small. Normally, the period to monitor an ASC is from one week to several weeks. Occasional high residual would make the scale of y-axis so large, that lot of information would be lost in the visualization. It would be hard to identify well performing cranes and compare them to cranes that are not performing so well. Thus, information is gained, rather than lost, by replacing information the extreme residuals with fixed values.

Additional point for consideration for the representation is the value of single residual. Crane performs dozens of moves daily. There are some bad moves on daily basis. A single bad move isn't in any way alarming in crane operations. There was already significant amount of variation in the residuals in Figure 15. Additionally, visualizing several cranes over the period of few weeks would result in even larger variation. Raw move residuals will indubitably vary. This, in turn, will result in highly spiky plot.

The move residual also does not explain the cause for the large residual in any way. Therefore, it is better suited to present a general picture of performance of multiple cranes. Thus, a filter will be applied to the residual to get a better overview of the crane performance. The rolling average will be used.

5 Results

5.1 Move residual

The linear model has now been prepared to predict the cycle times. Additionally, initial the results for the move residuals have been calculated. We have also concluded methods to refine the raw residuals for better presentation and to present more information. We will now make a move residual visualization for five cranes with the model. It's important to have additional cranes for comparing their performance sidelong. The data will include especially long period of several months to get a good overview of the method. Reliability of the findings will also be confirmed using other methods of crane performance.

The results of the move residuals can be observed from Figure 16. The figure was made using Seaborn lineplot method (*Seaborn.Lineplot*, n.d.). The legend and color are indicating the crane in question. The colored lines are the rolling average of move residual of individual cranes. The move residual increases downwards along y-axis. The y-axis is inverted for better overview. We can clearly observe changes in the performance on different cranes. We will now interpret the resulting move residuals for each crane that we have calculated using Figure 16.

AW01 was maintaining satisfactory performance until the end of May. At the end of May, there a period when the crane has not performed moves. This is explained by the straight line of the residuals. At the beginning of June something has happened to the crane which has resulted in degradation of performance. The crane had month long idle in July, after it regained the previous performance. AW02 has not performed many moves during the time-period. Therefore, the interpretation is not convenient.

AW03 has had an extended period of lacking performance between April and June. There were a few weeks in June when it did not perform any moves. In July, whatever the problem that the crane had, was repaired. However, AW03 still has high residuals in August. AW04 has steadily lacked performance during the time-period. It has not suffered from as high residuals as other cranes. Nevertheless, it is something that should be looked upon.

AW05 is clearly the best performing crane in this graph. The regression model's average error was little less than 10 seconds. There are times when the residual is close to 20 seconds. Residuals over 20 seconds give no point for concern when comparing to the other cranes. There are also periods when the residual is slightly over zero. However, as this seems to be the best performing crane, we will use it as a benchmark to compare to the other cranes.

AW07 seems to be the worst performing crane in this comparison. Residuals grew at the end of May and have sustained high since then. This crane also has generally the highest residuals amongst all cranes.

We have also found the cause for high amount of unexplained variance that was not filtered manually. Earlier, we utilized DFFITS-method to remove the unexplained variance from the training data. We removed approximately one fourths of the remaining rows in chapter 4.3.2. We can make a general observation from Figure 16, that there are many cranes that have had high move residuals at some point. The conclusion from this is that all the high residual moves were removed by DFFITS and the manual removal methods with domain knowledge. We have already inspected the residuals with the training data in Figure 14. Residuals over 40 were exceedingly rare. We have rolling average applied in Figure 16. Therefore, some data will be lost in the visualization. However, we can still make a general conclusion that every time the move residual is over 40, it has not been included in our training data. It is easy to observe from Figure 16 the cranes and time periods when a crane has not performed adequately. For instance, generally the moves of AW07 since the end of May has not been included in the training data. Thus, they also represent the data that was not included in the training data.

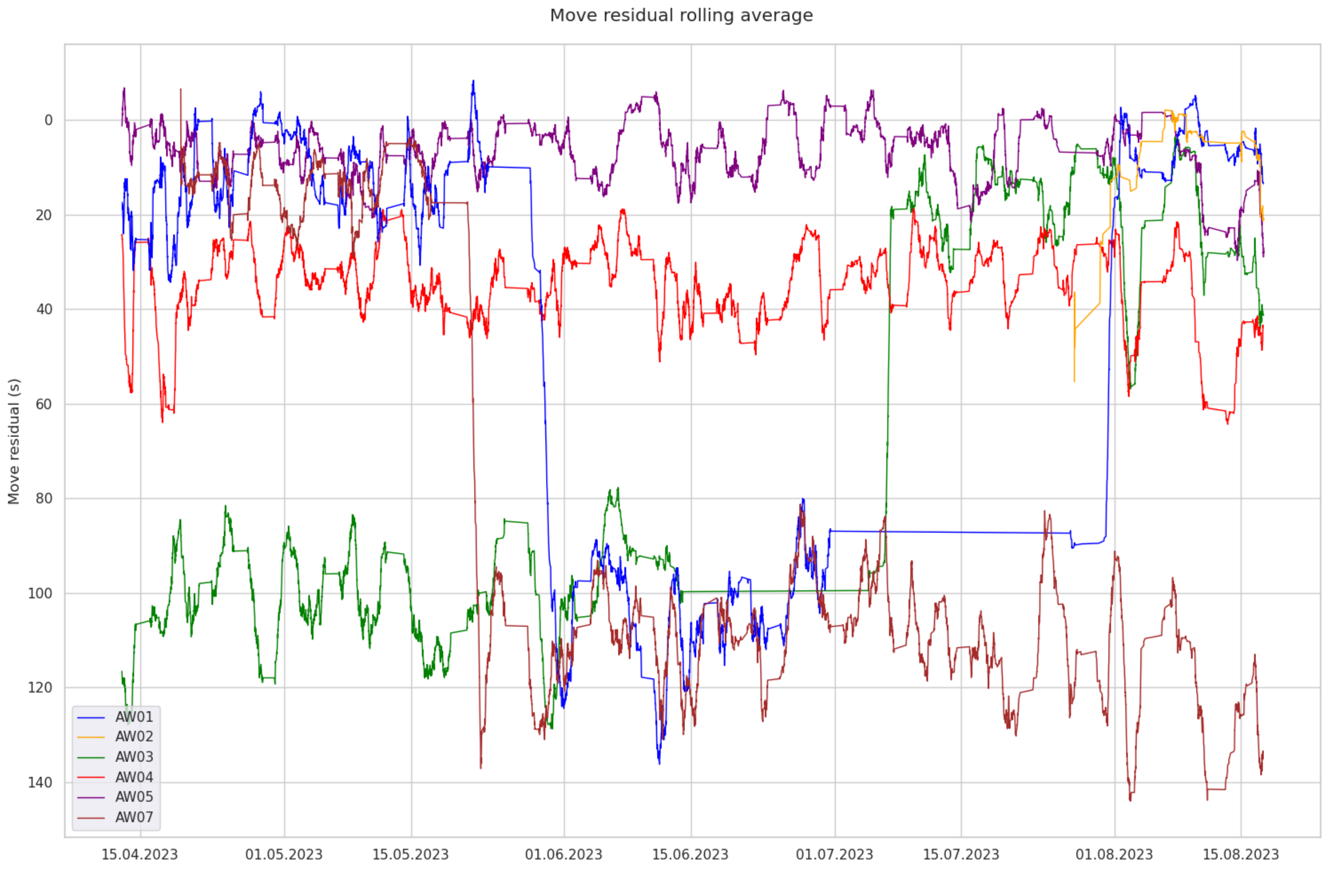

Figure 16. Move residual rolling average

5.2 Move residual reliability

Reliability of the results will now be confirmed. Therefore, we will use other methods to verify that the crane AW07 is performing less adequate than the other cranes. The performance of other cranes will also be verified. The methods will not be related to the regression model. This way we can verify reliability of the results. The methods will use the same data and time period as the regression model.

Earlier we confirmed that the most significant explanatory variable for cycle time is the gantry movement. Thus, it is the most natural place to examine first. Therefore, we will calculate the average gantry speed for each crane. The speed is calculated by summing the gantry travelled distance for each crane separately. Gantry travel time is also summed together for each crane. Total distance is then divided by the gantry travel time. Result is average speed in m/s for each crane.

The average speed for the cranes can be observed from Appendix 18. The figure was made using Seaborn barplot method (*Seaborn.Barplot*, n.d.). We can confirm that crane AW07 is moving slower than the other cranes. Crane AW03 also seems to have low gantry travel speed. However, we will review the Figure 16. There seems to be a significant drop in the move residuals for crane AW07 in May. There also seems to be a significant increase in the move residual for crane AW03 in June. This is an indication of correlation between gantry speed and the residual. The same trend can be observed from crane AW01 that has a period of large residuals. AW01 also has slightly low gantry speed. Highest observed gantry speed is of AW05, which has the best overall move residuals.

However, gantry speed can vary, and the gantry is not used during all moves. Gantry speed for individual move vastly differs when the crane travels long distance. On moves where gantry travels short distances, the gantry speed is lower because of the acceleration and deceleration times. Therefore, gantry speed alone is not sufficient metric for monitoring crane performance on detailed level. It is best suited to get an overview of the speeds from a longer period of time.

We will further explore the important ASC KPI’s to verify our findings. We have calculated average cycle times for cranes in Appendix 19. The direct comparison of cycle times is not normally done. The cranes in different blocks can be utilized differently, which may result in vastly different cycle times. The cycle times can also change depending on the situation at port. Normally ports want to handle the cargo ship as efficiently as possible. Thus, the operational mode of the cranes can change, which in turn leads to different cycle times. However, for the purpose of validating our model they are useful.

Cranes AW07, AW03 and AW02 have the highest average cycle times. The results for AW02 cycle times are not as dependable as on the other cranes since it has exceptionally low usage compared to the other cranes. We can observe, that the AW05 is also best performing crane when measured with cycle times. The average cycle time results are in line what was observed from the move residuals in Figure 16. Thus, move residuals are also correlating to the average cycle times in Appendix 19.

6 Discussion

6.1 Reliability and ethicality

The usage of predicted cycle times has drawbacks related to credibility. It raises questions regarding the accuracy of the predictions. The reliability of move residual is completely dependent on the accuracy of the predicted cycle time. Without accurate cycle time predictions there cannot be reliable move residuals. When the cranes are visually inspected at port during operations, they can sometimes appear to be inconsistent. However, we were able to demonstrate that the cycle times can be predicted reliably using simple linear regression. The move residuals also proved to be dependable when comparing to other measurements of crane performance. This further confirmed the accuracy of the predicted cycle times.

We excluded one thirds of the research data when we trained the linear model. There were clearly moves with faults and mros-calls that were easy to identify as bad moves to exclude from the training data. However, in the rows that were excluded from training, there was also considerable amount of moves with unexplained variance. It is likely that our data is missing information that could have explained more of the variance. Filling missing would lead to increased accuracy in cycle time predictions. Consequently, move residuals would also be more accurate.

We were also able to demonstrate, that many cranes had suffered from bad periods at some point. These bad periods caused rows of data to be discarded from training data by DFFITS. These moves were in fact bad moves without clear explanation in the data. Further inspection of the data might reveal further insights of the unexplained variance. Additionally, it is likely that the performance drops were caused on purpose. High speeds of the main moves are demanding for the mechanic structures of the cranes and the rails where the ASC travels on. Therefore, it is possible that the speeds of some cranes were lowered on purpose.

With the move residual, bad performing cranes can be identified more quickly. Therefore, existing equipment can maintain higher capacity. Ports are actively measuring their co2-emissions (Cobo, 2016, p. 16; Port of Rotterdam Authority, 2022, p. 4). Increasing the performance of existing cranes leads to decreased emissions since higher capacity can be achieved by existing cranes instead of acquiring additional ones.

The research data does not include any data that could be linked to an operator or other person involved in the operation. There are ROS and MROS-operations with timestamps that indicate human intervention. There are also truck drivers that provide landside logistics for the ASCs. However, the driver or the truck is not included in the data. All human related operations are completely anonymous. There is no possibility to link this information to any individual operator or truck driver. The details of both ROS/MROS and truck handling are also out of scope of this paper. Therefore, there are no issues with privacy or ethicality towards any individual persons.

The research cranes in question or the port operators using the cranes will not be disclosed. Disclosing the information would be unethical due to privacy of commissioner's customers. Additionally, performance metrics have not been disclosed as they are trade secrets of the commissioner.

6.2 Discussion of the main results in view of the theoretical framework

We were able to demonstrate that simple linear regression can be used to predict cycle times of the ASCs. The causality between travel distances of the main moves and cycle time was exceedingly high. There was minor scepticism whether the cycle times can be reliably predicted. The conclusion is that the move cycles of ASCs can be predicted reliably with simple linear regression. There is always room for improvement. However, when more sophisticated methods to predict the cycle times are considered, the objective for the method should be designated first. For the purposes of move residual with rolling average, simple linear regression is adequate.

We have also observed that there is a vast number of alarms and faults that the crane can encounter. Some of the alarms are caused by an actual problem on the crane. Others are caused by external factors, such as interference with safety areas that will trigger the light curtains. Additionally, there can be false alarms. There are faults that disrupt the operations but also alarms that do not intervene with crane operations in any way. Any definitive conclusions cannot be made from the alarms.

Utilization of mathematical input variable selection methods was not successful. The stepwise method could not be used for calculating the important input variables mathematically. This was due to multicollinearity and the resulting significance levels of the coefficients. However, adequate independent variables were found even without statistical models.

Both Pearson and Spearman correlation could be utilized to get an initial overview of the input variables of the cranes. The coefficients of correlation and the linear regression vastly differed from another. The coefficient for gantry distance was much smaller with linear model than the correlation calculation. The coefficients for trolley, and hoist travel distances were also smaller but not to same extent as the gantry travel distance. The cause for this is because the gantry travels much longer distances than trolley and hoist.

Multicollinearity in the regression presented further challenges with the interpretation of the coefficients. The coefficients could not be reliably examined due to multicollinearity. However, the coefficients still were aligned what could be expected from an ASC. The coefficient values of empty and loaded distances in the regression model were highly correlated with what was to be expected for an ASC.

We were also able reach remarkably high R-squared. Adjusted R-squared was identical to R-squared on all regression results. Therefore, adjusted R-squared could not be utilized to reduce the number of independent variables for the regression. This indicates our linear model did not have too many independent variables. It is likely that the model would benefit from more variables. However, the exact variables are not clearly defined for the moment. What is known from the missing variables is that they are related to TOS signals.

We also discovered the importance of utilizing both R-squared along with other metrics such as RMSE and MAE. Even though the R-squared was exceedingly high, both RMSE and MAE were only relatively high. R-squared of 0,97 seems a bit too high value. However, RMSE and MAE of approximately 10 seconds make the model seem less overfitted. The performance metrics are adequate but still improvements could be made to the accuracy of the model.

DFFITS proved to be particularly useful for removing unexplained variance from the model. The resulting filtering formulas were aggressive. However, a clear pattern was observed from the versus fits plot. It is likely, that there exist three reasons why the formula became so aggressive. First, the solution required to negate the effects of any unexplained variance. Second, there was high number of observations. Even though the formula included the observation amount many guidelines in sources had low observation number compared to the research data. Third, there were still missing input variables which resulted in higher cycle times than predicted. All three things considered, the used formula for DFFITS was adequate.

6.3 Conclusions and development proposals

We have demonstrated a new way to quantitatively measure the performance of an ASC. It can be used to present an overview of the crane performance in general. The concept is easy to understand without profound knowledge of the domain or machine learning. It enables easy identification of cranes performing worse than others. The method can be used by both the operators and manufacturers of the crane. It also remains unaffected by the external influence factors at port, such as a vessel port call, contrary to other KPIs for ASCs.

The move residual is an important addition to the current performance metrics of ASCs. Problems with the crane performance can easily go unnoticed due to variation in the prevailing situations at port. It can be applied to all existing and new ASCs. There is also other equipment at ports that could benefit from the usage of move residual, such as the ARTG. Additionally, more ACSs there are at any given port, more important the performance measurements become. More ASCs also means that there is more training data available. Thus, more accurately the cycle times can be predicted and more accurate the move residual becomes.

The move residual is closely related to anomaly detection. Chandola et al. surveyed different outputs of anomaly detection and score method of residuals was one of them (Chandola et al., 2009, sec. 15:10, 15:32). Move residual can also be described as an anomaly score for the moves. However, crane moves quite are different compared to other domains where anomaly detection is used. Often the anomaly detection is used to detect and inspect individual observations. With cranes, the number of moves is high. Thus, the number of anomalies is also high. Additionally, all the necessary explanatory variables in the data for individual observations might not exist. One move with high anomaly score is not particularly interesting to either port operators or the machine suppliers. Therefore, it is better to get an overview of the move residual for each crane with the rolling average.

The move residual is an absolute value. However, we can argue that the benefit of the move residual is more on the relative value rather than on absolute value. Inspecting move residuals of an individual crane without points for comparison is normally not particularly useful. For an individual crane, the observer should have adequate knowledge of normal and abnormal residuals. Therefore, the move residual is best utilized when comparing cranes to each other. Then there are clear points for comparison.

For instance, the move residuals in Figure 16 of AW04 appear to be slightly large. If we were to remove cranes that have residuals over 100 from graph, the performance of AW04 would appear to be much lower. The relative comparison gives clear indication how much worse or better cranes are performing. This gives a good indication for anyone inspecting the crane performance of which crane to focus on. The points of comparison also lessen the need of having knowledge how the move residual is calculated.

The move residual can also be compared to other KPIs related to an ASC. During the research we were able to verify that the differences in residuals reflect closely to other measures of productivity of the crane, such as moves per hour. Additionally, the results from move residual reflected the issues on the gantry movement speed on the cranes that had high move residuals.

Move residual with linear regression is not the only way to detect the cranes that are performing worse than other cranes. Diverging ASCs could also be detected using classification methods. However, this would bring challenges regarding the definition of diverging ASC. Diverging ASC is ambiguous and very subjective term. When a crane is down due to fault, it will easily be noticed without any use of machine learning. However, the situation changes when the crane can continue the operation. It is hard to clearly define whether the crane is operating normally or abnormally. There are various metrics that are be utilized, none of which focus on the pure crane performance without external factors.

It is also possible to represent the move residuals in diverse ways. The timescale on the x-axis on Figure 16 presents the changes in cranes rapidly. If we were to do changes to the cranes, we would see them quickly. However, there can be dozens of cranes operating at a port. All the cranes cannot fit nicely to a same visualization in this way. Therefore, it would be possible to calculate cumulative value of the residuals for each crane of same time period. The residuals would be presented in a barplot. Therefore, higher cumulative value would indicate worse performance. In this way all the cranes would fit into a same visualization.

Move residual is a concept that could be used in other applications other than ASCs as well. There is considerable number of automated equipment and machinery whose performance depends on the external factors. For instance, a delivery robot. Both equipment share the waiting time due to external sources and the distance travelled during a move. Additionally, path optimization is done externally. The performance metrics and external factors have a lot in common. However, the challenge to a machine supplier is the same. How to negate the external factors and measure the overall equipment productivity?

References

Abdelmagid, A. M., Gheith, M., & Eltawil, A. (2022). Scheduling External Trucks Appointments in Container Terminals to Minimize Cost and Truck Turnaround Times. *Logistics*, *6*(3), 45. https://doi.org/10.3390/logistics6030045

Arimie, C. O., Biu, E. O., & Ijomah, M. A. (2020). Outlier Detection and Effects on Modeling. *OALib*, *07*(09), 1–30. https://doi.org/10.4236/oalib.1106619

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Balliauw, M. (2021). Time to build: A real options analysis of port capacity expansion investments under uncertainty. *Research in Transportation Economics*, *90*, 100929. https://doi.org/10.1016/j.retrec.2020.100929

Belsley, D. A. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley-Interscience.

Burnham, K. P., Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). Springer.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chu, F., Gailus, S., Liu, L., & Ni, L. (2018). The future of automated ports. *McKinsey & Company*. https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/the-future-of-automated-ports

Cobo, P. T. (2016). *Optimization of Yard Operations in Container Terminal From an Energy Efficiency Approach* [PhD, Universidad Politécnica de Cataluña – Barcelona Tech]. https://www.tdx.cat/bitstream/handle/10803/620639/TPTC1de1.pdf

De, S. K., & Baron, M. (2012). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *Journal of Statistical Planning and Inference*, *142*(7), 2059–2070. https://doi.org/10.1016/j.jspi.2012.02.005

Enders, F. B. (n.d.). Coefficient of determination. In *Encyclopedia Britannica*. https://www.britannica.com/science/coefficient-of-determination

Freedman, D. (2009). *Statistical models: Theory and practice* (Rev. ed). Cambridge Univ. Press.

Frost, J. (2019). *Regression analysis: An intuitive guide for using and interpreting linear models*. Statistics By Jim Publishing.

Frost, J. (2020). *Introduction to statistics: An intuitive guide for analyzing data and unlocking discoveries*. Statistics by Jim Publishing.

Golbabaie, F., Seyedalizadeh Ganji, S. R., & Arabshahi, N. (2012). Multi-criteria evaluation of stacking yard configuration. *Journal of King Saud University - Science*, *24*(1), 39–46. https://doi.org/10.1016/j.jksus.2010.08.010

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, *26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, *12*(1), 69–82. https://doi.org/10.1080/00401706.1970.10488635

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Identifying Influential Data Points. (n.d.). Regression Methods. Retrieved 16 October 2023, from https://online.stat.psu.edu/stat501/lesson/11/11.5

Interreg Europe. (2020). *Reduction of CO2-Emissions ejected by Heavy-Duty-Vehicles*.

Javatpoint. (n.d.). *Linear Regression in Machine Learning*. https://www.javatpoint.com/linear-regression-in-machine-learning

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.4097/kja.19087

Kim, M. G. (2017). A cautionary note on the use of Cook’s distance. *Communications for Statistical Applications and Methods*, *24*(3), 317–324. https://doi.org/10.5351/CSAM.2017.24.3.317

Konecranes. (n.d.). *History*. https://www.konecranes.com/about/history

Konecranes. (2023). *Automated RMG (ARMG) System*. https://www.konecranes.com/port-equipment-services/container-handling-equipment/automated-rmg-armg-system

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kutner, M. H. (Ed.). (2005). *Applied linear statistical models* (5th ed). McGraw-Hill Irwin.

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, *71*(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

Meester, T. (2022). Automation Trends. *Port Technology International*, *122*, 18–20.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed). Wiley.

Morales-Fusco, P., Saurí, S., Lekka, A. M., & Karousos, I. (2016). Assessing Customs Performance in the Mediterranean Ports. KPI Selection and Best Practices Identification as Part of the MEDNET Project. *Transportation Research Procedia*, *18*, 374–383. https://doi.org/10.1016/j.trpro.2016.12.049

Navigating stormy waters. (2022). United Nations.

Nieuwenhuis, R., Grotenhuis, M., te, & Pelzer, B. (2012). influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *The R Journal*, *4*(2), 38. https://doi.org/10.32614/RJ-2012-011

Pandas.DataFrame.corr. (n.d.). Pandas Documentation. Retrieved 16 November 2023, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

Port of Rotterdam Authority. (2022). *Highlights of 2021 annual report*.

Rintanen, K. (2018, June 12). *THE PEMA PAPERS: Automatic Stacking Crane Performance*. TOC Europe, Rotterdam. https://www.tocevents-europe.com/content/dam/Informa/toc-events-europe/presentations/tech-toc-presentations-/the-pema-papers/Automatic_Stacking_Crane_Performance.pdf

Rovetta, A. (2020). Raiders of the Lost Correlation: A Guide on Using Pearson and Spearman Coefficients to Detect Hidden Correlations in Medical Sciences. *Cureus*. https://doi.org/10.7759/cureus.11794

Saleh, A. K. M. E., Arashi, M., & Kibria, B. M. G. (2019). *Theory of ridge regression estimation with applications*. Wiley.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients: *Interpretation of regression coefficients*. *Methods in Ecology and Evolution*, *1*(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Scipy.stats.gaussian_kde. (n.d.). SciPy Documentation. Retrieved 16 November 2023, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html#scipy.stats.gaussian_kde

Seaborn.barplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.barplot.html

Seaborn.histplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.histplot.html

Seaborn.kdeplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.kdeplot.html

Seaborn.lineplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.residplot.html

Seaborn.regplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.regplot.html

Seaborn.residplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.residplot.html

Seaborn.scatterplot. (n.d.). Seaborn Documentation. Retrieved 16 November 2023, from https://seaborn.pydata.org/generated/seaborn.scatterplot.html

Siegel, A. F. (2016). *Practical business statistics* (Seventh edition). Academic Press.

S&P Global Market Intelligence. (2022). *Maritime and Trade Talk: Trade outlook response to geopolitical headwinds*. https://www.spglobal.com/marketintelligence/en/mi/research-analysis/maritime-and-trade-talk-trade-outlook-response-to-geopolitical.html

Stahlbock, R., & Voβ, S. (2008). Vehicle Routing Problems and Container Terminal Operations – An Update of Research. In B. Golden, S. Raghavan, & E. Wasil (Eds.), *The Vehicle Routing Problem: Latest Advances and New Challenges* (Vol. 43, pp. 551–589). Springer US. https://doi.org/10.1007/978-0-387-77778-8_25

Statsmodels.regression.linear_model.OLS. (n.d.). Statsmodels Documentation. Retrieved 16 November 2023, from https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

Statsmodels.regression.linear_model.OLSResults.summary. (n.d.). Statsmodels Documentation. Retrieved 16 November 2023, from https://www.statsmodels.org/devel/generated/statsmodels.regression.linear_model.OLSResults.summary.html

Statsmodels.stats.outliers_influence.OLSInfluence. (n.d.). Statsmodels Documentation. Retrieved 16 November 2023, from https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.OLSInfluence.html

Thode, H. C. (2002). *Testing for normality*. Marcel Dekker.

Van Der Meer, T., Te Grotenhuis, M., & Pelzer, B. (2010). Influential Cases in Multilevel Modeling: A Methodological Comment. *American Sociological Review*, *75*(1), 173–178. https://doi.org/10.1177/0003122409359166

Visser, D. (2017, November 30). *Container Terminal Developments*. Feport Third Stakeholders' Annual Conference.

World Bank. (2023). *The Container Port Performance Index 2022: A Comparable Assessment of Performance Based on Vessel Time in Port*. World Bank. https://doi.org/10.1596/39824

Yan, X., & Su, X. (2009). *Linear regression analysis: Theory and computing*. World Scientific.

Young, D. S. (2017). *Handbook of regression methods*. CRC Press.

Appendices