Team 58
Ariel Israel | Eddie Chen | Min Chun | Robert Mallon | Ryan Chung

Impact of COVID-19 on Port Processing Efficiency

Background Information

Containerization and trade liberalization paved the way for maritime trade for the last 100 years. Since its application after World War II, containerization revolutionized freight transport by increasing handling capacity, increasing labor efficiency, and reducing damage. These improvements led to decreased port congestion and kickstarted the era of accelerated growth and globalization. The North American Free Trade Agreement (NAFTA) was signed on Dec. 1992 by Canada, Mexico, and the United States. This agreement eliminated tariffs on goods traded among these countries and as a direct result, regional trade tripled from roughly $290 billion in 1993 to more than $1.1 trillion in 2016. The global economy boomed, and producers and consumers both enjoyed having more at ever-faster speeds without a second thought. Then, the COVID-19 pandemic hit. The shock of COVID-19 was sudden and severe, and according to Cullinane & Haralambides (2021), maritime shipping was one of the hardest hit industries. COVID-19 exacerbated congestion at many ports and canals worldwide as quarantine mandates and virus outbreaks amongst employees slowed down or in some cases halted processing times.

Problem Overview & Analysis Approach

In a separate 2021 study, Millefiori, et. al. (2021) concluded that shipping mobility had been negatively affected by the COVID-19 pandemic to varying degrees based on market and vessel size. The aim of this project is to determine whether it is possible to predict delays at shipping ports based on the connection Millefiori outlined between COVID-19 and maritime shipping delays. To do this, we developed an Ordinary Least Squares (OLS) linear regression model using average days at berth and anchor as dependent variables and COVID cases and cargo data as features. Additionally, Leonard (2021) references the vulnerability of shipping ports to climate forces, prompting the inclusion of weather features in our modeling efforts as well.

Analysis presented will focus primarily on The Port of Los Angeles. As the busiest seaport of the western hemisphere, it acts as the leading gateway for international commerce by handling 20% of all cargo that enters the United States. In the calendar year of 2021, this is equivalent to 10.7 million containers. Hence, any slowdowns as a result of COVID may be exacerbated at this location, allowing for easier impact detection, and inferences can be generalized across numerous other shipping ports.

Hypothesis

In creating this model, we wanted to evaluate our hypothesis that confirmed COVID-19 cases can effectively predict slowdowns in port processing. An increase in COVID-19 severity may result in a weakened labor force, and trigger social distancing or other efficiency reduction measures. Intuitively, we believe that larger vessel and cargo containers may reduce port efficiency by either requiring more manpower or as a result of increased dwell times.

However, further research has shown that shipping ports often have poor weather contingencies. Leonard (2021) mentions that during Superstorm Sandy in 2012, cargo was delayed by approximately eight days due to turbulent sea conditions. While this may be an extremity, weather can play a significant role in port processing and should not be discounted.

Hence, it is believed that combining COVID-19 statistics with cargo and weather data will enable accurate predictions for cargo dwell times.

Data Overview

There are three main datasets used in this analysis: COVID, cargo, and weather data.

COVID Data

For COVID data, we originally looked at the Centers for Disease Control (CDC) database. As the national public health agency of the United States, the accuracy of their data would be dependable. However, the data provided was aggregated at a state level, which, given the size of California, would not be representative of COVID near the Port of Los Angeles. Due to this, we explored data from the Systems Science and Engineering (CSSE) department at Johns Hopkins University. This was available at a zip code and county level, which provided sufficient coverage for our area of interest.

As Johns Hopkins University is one of the most prestigious research universities in the United States, their publicly available COVID data required minimal cleaning. The particular dataset examined was a time series of confirmed cases dating back to January 22nd 2020. This was ingested directly into a *jupyter* notebook, using the *pandas* csv reading capabilities when referencing a URL. Data manipulation involved filtering to Los Angeles County and unpivoting the date columns, as the original data set had provided a column for each date. This was done using *pandas* data frame manipulation, and can be observed in **covid.ipynb**.

Exploratory data analysis revealed two areas of concern: a delayed delta outbreak and volatility in confirmed cases throughout 2022. Both phenomena can be observed in Figure 1, produced using the *matplotlib* and *seaborn* libraries in *python*.

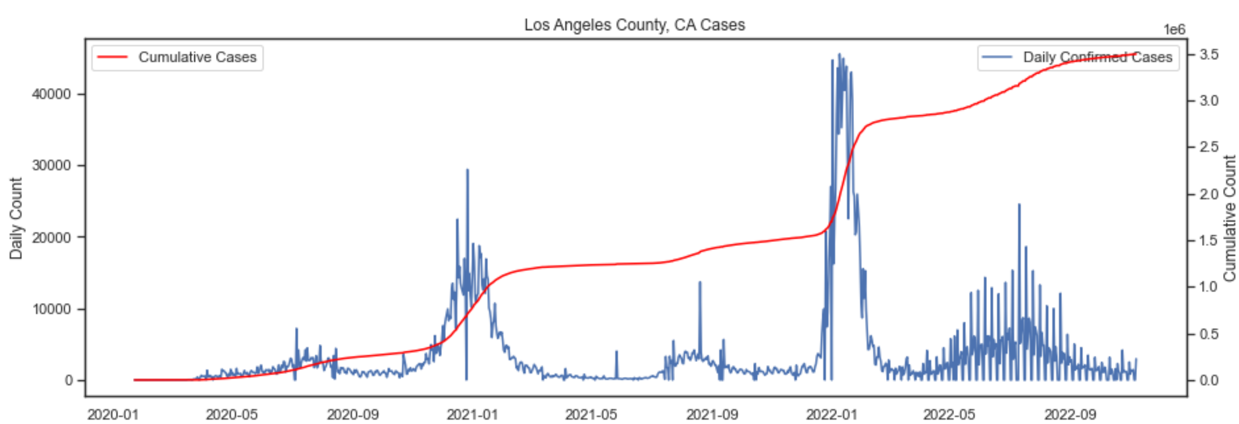

Figure 1 – Los Angeles County, CA Cases.

The delta COVID outbreak had first been reported in May of 2020. Hence, it was deemed that there was a slight lag in the impact until the end of the year. Omicron was represented well, with a spike at the expected time frame. Volatility in 2022 was due an update in reporting schedules, where data was no longer published on weekends. Hence, the combined valleys and peaks were a result of delayed weekend cases being reported on Monday.

Cargo Data

The Port of Los Angeles provided container and vessel statistics on their business website. Container data was available at a monthly cadence dating back to January 2020. The standard in the shipping and maritime industry for containers is a "Twenty-foot Equivalent Unit" (TEU). Imports and exports of both loaded and empty units were provided at this monthly level. As our COVID data was provided at a daily level, we made the decision to uniformly distribute the container data across all days of the month. While perturbation was also considered to provide some variation amongst the days, this was deemed unnecessary and immaterial in the broader picture of prediction.

Figure 2 illustrates the number of loaded export TEUs against the rolling seven day average of covid cases. As per our expectations, the number of exported units was decreasing over time, with our assumption being resultant of COVID. The sudden drop in September indicates an end of data availability.

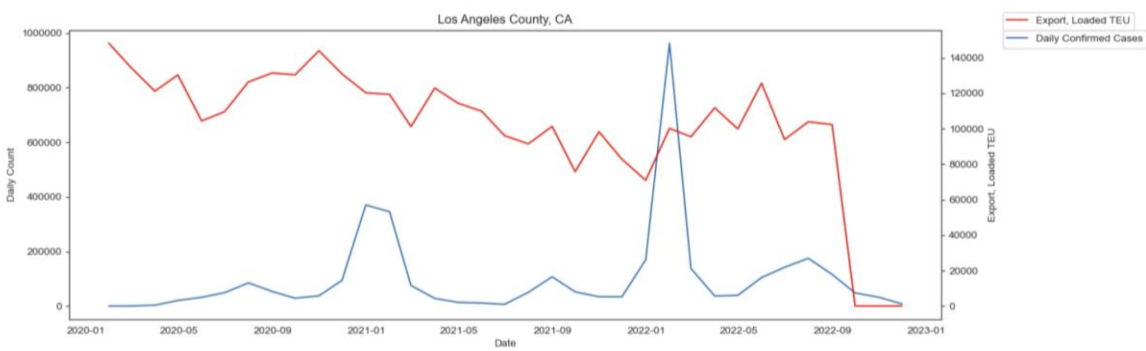

Figure 2 – Loaded export TEUs against 7-day rolling average confirmed COVID cases

Vessel statistics were provided on the Port of Los Angeles website in *pdf* format. We used the *tabula* python library to convert this to a data frame, which can also be observed in **covid.ipynb.** The data provided included Vessels at Anchor, Berth, and Departed. At Anchor refers to being anchored at any open space, while at Berth represents a fixed wharf or location in the port. We ultimately decided to use a combined metric, average days at anchor and berth, as our response. This will be discussed further in the modeling section. Figure 3 displays this metric from the start of 2020.

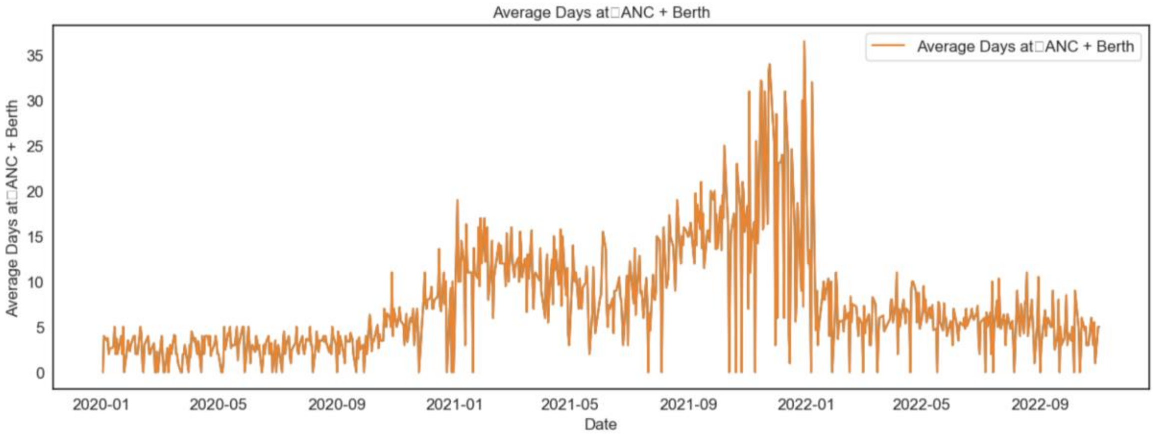

Figure 3 - Average days at anchor and berth. This was used as our dependent variable.

Weather Data

The National Centers for Environmental Information is a government body that contains the largest archive of climate data available for the United States. Yearly weather data is available dating back to 1929, and required very minimal cleaning. Weather stations have been set up across the nation, and the closest one to the Port of Los Angeles was at Long Beach Airport, which is approximately 8.65 miles away.

Given the institute's rigor in collecting and maintaining data, multiple factors were available. However, the Global Surface Summary of the Day was deemed the most reasonable and likely to impact port processing. In particular, the variables we retained were:

Variable	Description
VISIB	Mean visibility (.1 miles)
WDSP	Mean wind speed (.1 knots)
MXSPD	Maximum sustained wind speed (.1 knots)
PRCP	Precipitation amount (.01 inches)
FRSHTT	Indicator for occurrence of: Fog Rain or Drizzle Snow or Ice Pellets Hail Thunder Tornado/Funnel Cloud

For simplicity, the FRSHTT variable was converted into a binary indicator for inclement weather. The rationale is that there was not enough distinction between each event regarding their impact on port operations.

Model Overview

Model Preparation

The hypothesis is that a combination of COVID cases, cargo data, and inclement weather effects can effectively predict a port's slowdown in processing efficiency. In order to determine the latter, a suitable metric must be used. A number of proxies were available in the Port of Los Angeles vessel statistics, but we ultimately decided Average Days at Anchor and Berth was the most suitable metric to use. Vessels at Anchor, vessels at Berth, and vessels departed were not comprehensive measures, as they represented only a single part of the ship docking process. Hence, the selected metric was deemed the most holistic. Figures 4 and 5 show a histogram and box plot of the response respectively. Both indicate right skewed data, with potentially high outlier values.

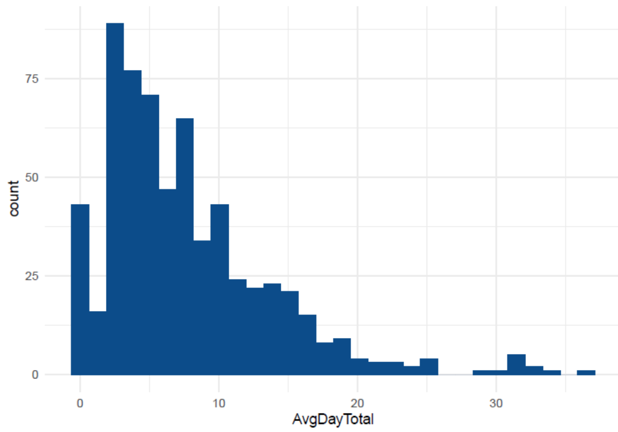

Figure 4 - Histogram of Average Days at Anchor and Berth

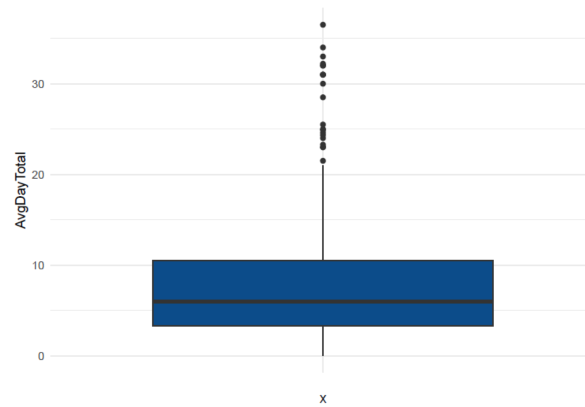

Figure 5 - Box plot of Average Days at Anchor and Berth

Throughout the modeling process, we also discovered a number of outliers and high leverage points. Utilizing Cook's distance, we found several outliers with a distance greater than 3. This corresponds to what the above distributions show.

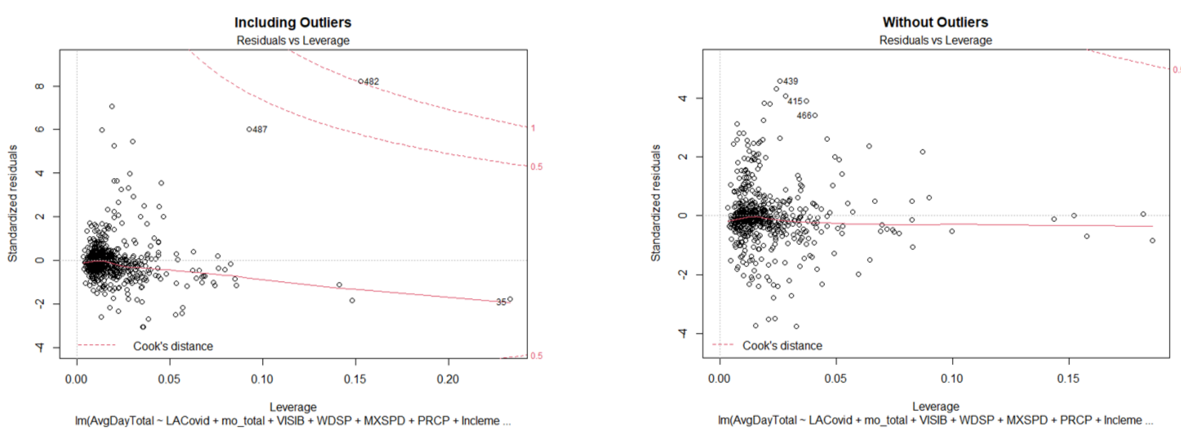

Figure 6 – Residuals vs Leverage with and without outliers

The split between training and test data was contentious. Our initial intentions as stated in the project progress were to use 2020 and 2021 as our training set and use 2022 as our test set. However, this was deemed inappropriate due to autocorrelation factors being present. We ultimately decided that a random 70 / 30 split between train and test would be a better representation of our model. This would allow us to make predictions based on varying degrees of port capacity, covid cases, and weather details.

In terms of model selection, it was agreed that Ordinary Least Squares (OLS) Linear Regression would be the method of choice. This is a technique that has been highly emphasized throughout this course, with implementations being demonstrated across retail, finance, and also media & advertising. As a result, hyperparameter tuning was not applicable in this instance.

Model Iterations

An overview of our models can be found in **model_projectV2.R.**

We began with a naïve baseline model that would be quick to implement just to get started with the iteration process. This included all the variables that were available. Multiple failed implementations in R revealed that vessel statistics were not actually populated for non-business days. This was missed in the EDA, as the rows had been omitted completely. Our options were toe either keep all calendar days and forward fill the empty values, or omit non business days. The benefit to keeping all calendar days would be that we would have extra data points, approximately 30% of our entire data set, each with their own COVID and weather data points. Our model with triple weighted Fridays ended up only producing an adjusted R^2 value of 0.06 on the test data set. With this abysmal score, we decided the omission of non business days was preferred.

We iterated through the following models, each with varying results. Models 1 – 7 all omit non business days.

Model name	Description
Model 0	All variables, triple weighted Fridays
Model 1	All variables, omitted weekdays
Model 2	All variables with an indicator for calendar months to account for seasonality
Model 3	Model 2 + Days of the Week
Model 4	Model 2 + interaction term between COVID cases and monthly TEU
Model 5	Model 2 + linear regression model with the accumulative prolonged effect of COVID cases
Model 6	Model 5 – COVID cases
Model 7	Significant Variables from Model 2 only

Results and Insights

We ultimately decided on Model 4 as it contained the most descriptive factors without sacrificing too much on the adjusted R^2, RMSE and MAE values, as seen in Figure 7 below.

	testdata (N = 181)
Adj R-Squared	
model 2: base model accounting for seasonality (months)	0.653
model 4: model with interaction term	0.653
model 7: model with only significant variables	0.66
RMSE	
model 2	2.856
model 4	2.855
model 7	2.828
MAE	
model 2	2.141
model 4	2.142
model 7	2.126

Figure 7 – Comparison of model metrics between model 2, 4, and 7

We also used k-fold cross validation, with results very similar to those above.

	df (N = 601)
Adj R-Squared	
model 2	0.685
model 4	0.685
model 7	0.688

Figure 8 – k-fold cross validation results

Our final model had the following features:

<u>COVID</u>

- Daily confirmed cases

<u>Weather</u>

- Mean visibility (.1 miles)
- Mean wind speed (.1 knots)
- Maximum sustained wind speed (.1 knots)
- Precipitation amount (.01 inches)
- Inclement weather indicator (binary)

<u>Cargo</u>

- Total monthly TEU

- Vessels at anchor
- Vessels at berth
- Vessels departed

Other

- Month dummy variable
- Interaction term between COVID and Total monthly TEU

These were intuitively correct, and the model results validated our hypothesis.

We did explore one further model, which was without the inclusion of the COVID variable at all. As evidences in Figure 2, there did not seem to be a strong correlation with and without the COVID feature. Results shown in the table below reveal that the performance of those models was comparable.

Metric	Model 4	Model 4 without COVID
Training Adjusted R^2	0.6782	0.6788
Test Adjusted R^2	0.6885	0.6876
RMSE	2.635	2.639
MAE	1.866	1.864

This was the most interesting finding from our analysis!

Conclusion

Overall, we can conclude that we have created a great model for predicting a port's processing efficiency, but there was no significant correlation with COVID cases. In terms of predicting power, Model 4 is especially accurate and therefore could be used to predict slow downs in goods processing capabilities of the Port of Los Angeles. However, this would not be correlated with COVID cases in the area.

With more time we would further refine our models by testing out new features and additional data sources so that we could more accurately isolate and predict potential slow downs in the Port of Los Angeles processing capacity. In terms of public health data, an investigation into why there was a slight delay in the Delta outbreak, in addition to vaccination rollout would be helpful. Furthermore, local restrictions and lockdowns would also provide useful information for our prediction model. Additional cargo data such as harbor pilot data and vessel size would also be useful features, as intuitively larger containers would require longer dwell times.

Additionally, as retailers shifted to an e-commerce focus during the pandemic there may have been automation efforts in the maritime shipping industry. Cranes, automated assembly lines and other advancements in the industry may have resulted in a reduced reliance on the labor force, and so COVID cases may no longer be as significant.

Time Series analysis would also be an area for further investigation. The presence of autocorrelation within the data argues that ARIMA would be a more suitable technique. The use of the "AdStock" concept mentioned in the course as well, could also be employed and was briefly attempted in Model 5 to no avail.

Ultimately, further analysis would allow us to convert our predictive capabilities into uses for prescriptive analytics. Techniques such as exponential smoothing and CUSUM may allow us to proactively suggest avoidance from a particular port if a slowdown is expected.

References

Millefiori, L.M., Braca, P., Zissis, D. et al. COVID-19 impact on global maritime mobility. Sci Rep 11, 18039 (2021). https://doi.org/10.1038/s41598-021-97461-7

Cullinane, K., Haralambides, H. Global trends in maritime and port economics: the COVID-19 pandemic and beyond. Marit Econ Logist 23, 369–380 (2021). https://doi.org/10.1057/s41278-021-00196-5

Rivero, Nicolás (2021-09-28). "Cargo ships are so stuffed that ports are struggling to unload them". ("Cargo ships are so full that ports are struggling to unload them ...") Quartz (publication). Retrieved 2021-09-29.

https://www.joc.com/maritime-news/container-lines/new-oligopoly-container-shipping_20190704.html

https://www.nature.com/articles/s41467-022-32070-0

https://sgp.fas.org/crs/row/R42965.pdf

https://lloydslist.maritimeintelligence.informa.com/LL1136229/Suez-Canal-remains-blocked-despite-efforts-to-refloat-grounded-Ever-Given

https://www.nytimes.com/2021/03/26/business/suez-canal-blocked-ship.html

https://www.bbc.com/news/business-58643717

Leonard, M (2021) "As storms become more frequent and volatile, some ports plan for the risk — but most do not" https://www.supplychaindive.com/news/port-climate-change-sea-level-rise-change-sandy-new-york-los-angeles-pianc/600285/