Short-Term Forecasting of Air Cargo Demand from a European Airport Hub to the United States during COVID-19

B. Verhoeven
Department of Mathematics, Faculty of Sciences
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, 1081 HV
Email: brittverhoeven01@gmail.com

N.K. van Hout
Department of Mathematics, Faculty of Sciences
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, 1081 HV
Email: nimovanhout@hotmail.com

A. Devaraj
Department of Mathematics, Faculty of Sciences
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, 1081 HV
Email: archika.devaraj@gmail.com

H. Zwitzer
Director Development & Process Support Revenue Management Cargo
KLM Royal Dutch Airlines
Amsterdam, the Netherlands, 1117 ZL
Email: Hans.Zwitzer@klmcargo.com

T. Crapts
Project Manager Data Analytics
KLM Royal Dutch Airlines
Amsterdam, the Netherlands, 1117 ZL
Email: Terry.Crapts@klm.com

A. Ion
Project Manager Data Analytics
KLM Royal Dutch Airlines
Amsterdam, the Netherlands, 1117 ZL
Email: Andrei.Ion@klm.com

T.H.A Koch
Stochastics Research Group, Centrum Wiskunde & Informatica
Amsterdam Science Park, Amsterdam, the Netherlands, 1098 XG
Email: Thomas.Koch@cwi.nl

E.R. Dugundji
Department of Mathematics, Faculty of Sciences
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, 1081 HV
Email: e.r.dugundji@vu.nl

Word Count: 6280 words + 4 tables (250 words per table) = 7,280 words

Submitted Monday, July 20, 2020

ABSTRACT

Air cargo is mostly transported on passenger flights. During the COVID-19 outbreak, there have been worldwide restrictions on passenger transportation. Therefore, airlines experienced a capacity problem for air cargo. Better insight of air cargo demand during COVID-19 could contribute to the better arrangement of capacity by accordingly adapting flight schedules for cargo. The aim of this research was to make short-term predictions of air cargo demand between a major European airport hub and the United States during the COVID-19 pandemic. This was done for the month of May in 2020 by making 14-day predictions. The same was done for the year 2019 to observe whether the models performed well in the absence of the pandemic. The data set was compiled using data provided by a major commercial airline and exogenous features, such as stock market indices, foreign currency exchange rates and healthcare related predictions during COVID-19. To make the predictions, two classes of machine learning models for time series were compared: Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM). In the year 2020, the best performing model among the ARIMA-based models is the Seasonal ARIMA including the exogenous feature *Schedule*. During the year 2019 the Seasonal ARIMA model without exogenous features generates the most accurate predictions. Among the LSTM models, the multivariate LSTM models outperform the univariate LSTM models in both years. Nonetheless, the ARIMA-based models are more accurate than the multivariate LSTM model in this research.

Keywords: Air Cargo, COVID-19, Airlines, ARIMA, LSTM

INTRODUCTION

During the outbreak of COVID-19, many industrial sectors have faced exceptional challenges, among which the air cargo industry. Since the beginning of the COVID-19 crisis, air cargo has been an essential partner in shipping vital medical goods and equipment. Air cargo is indispensable, because transport by plane can be significantly faster than shipments by sea or overland, and thus plays a major role in sustaining global supply chains for time-sensitive materials (*1*). Fast transportation of medical goods and equipment has been crucial during this pandemic. Furthermore, there is a massively increasing demand for medical goods and equipment. The shipment of new supplies on a short notice is critical for the healthcare and survival of the infected.

However, approximately 80% of all transatlantic air cargo was transported in the belly holds of passenger flights (*2*) prior to the COVID-19 pandemic, with only the remaining share of 20% being transported on dedicated all-cargo planes. Since passenger transportation was restricted worldwide during COVID-19, a lack of transportation capacity was created. For example, travelers from most European countries were banned by the US. Whilst all-cargo flights were initially being operated at similar levels as the same period last year, they were unable to compensate for the loss of cargo capacity on passenger aircraft. Most of the originally scheduled air cargo could therefore not be transported during the pandemic. The European Commission responded 26 March 2020 by issuing relaxed guidelines for the duration of the COVID-19 crisis intended to assist Member States in maintaining and facilitating air cargo operations, until the exceptional air traffic and travel restrictions are lifted. It is certain that as the COVID-19 crisis continues, the airline industry will undergo a seismic shift.

Problem Description

The travel restrictions during COVID-19 led to a 70-90% or more decrease of passenger flights for many major commercial airlines. Accordingly, airlines have been losing income from passenger revenue as a result of the COVID-19 crisis, threatening the future of airline companies. However, while passenger transportation was nearly non-active, air cargo transportation remained a source of revenue given sufficient capacity could be arranged. Airlines have thus had to at least temporarily rely on their air cargo business operations and find a viable solution to the capacity problem. Therefore, it is relevant to be able to predict air cargo demand in order to add appropriate capacity for cargo in revised flight schedules.

Literature Review

This research deals with a demand forecasting problem for time series. Therefore, a literature research was conducted, investigating papers that discuss time series and air cargo demand forecasting. Several studies have examined correlations between economic factors and air cargo demand relying on different types of models. Marazzo et al. (*3*) investigated the relationship between air transport demand and economic growth in Brazil. Chi & Baek (*4*) adopt an Autoregressive Distributed Lag (ARDL) model to examine both the short- and long-run effect of economic growth and market shocks, like SARS epidemic and the 2008 financial crisis, on freight services. Hathurusingha & Mudunkotuwa (*5*) apply ARIMA modeling for constructing the forecast on air freight imports and exports.

Other studies focus on applying neural networks when forecasting air cargo demand based on economic growth. Chen et al. (*6*) implement back-propagation neural networks to enhance the accuracy of forecasting air cargo demand from Japan to Taiwan, whereas Baxter & Srisaeng (*7*) use an artificial neural network to predict Australia's export air cargo demand.

Another neural network model that is used in demand forecasting research is the Long Short-Term Memory (LSTM) model. Even though we did not find papers specifically using the LSTM model for air cargo demand, Su et al. (*8*) use this model for hourly natural gas demand forecasting, based on data that is chronologically arranged. Since LSTM is applied in this research as a demand forecasting model for time series, this model was interesting for our research.

However, none of the above mentioned papers consider the transatlantic route between Europe and the United States. Moreover, we have not yet experienced such an extreme situation like the COVID-

19 crisis before so severely impacting the airline industry, which makes this issue unusual and unique in its kind. There exists little to no literature yet addressing the consequences of the COVID-19 outbreak on the air cargo industry and forecasting the resulting air cargo demand.

The aim of this research paper is to make short-term, 14-day, predictions on the exported air cargo demand from a major European airport hub to the United States during the outbreak of COVID-19. In this paper, the description of the data is first given, followed by the methodology. Next, an overview of the results of the models implemented and the discussion are presented. Finally, the conclusion summarizes this research and recommendations for future research are provided.

DATA

Data received from a major commercial airline was used in this research. A dataset was compiled containing the weight (in kilograms) of all cargo transported by the airline from a major European airport hub to the United States per day. The weight only covers the air cargo tranported by passenger aircraft. The data set ranges from September 2016 to May 2020, containing 1,369 shipment days.

Exogenous features were added to the data in order to make predictions using the forecasting models. Based on Chou et al. (*9*), the foreign currency exchange rate (EUR/USD) and the stock market indices (AEX, NYA and DJI) were included. Furthermore, data sets containing healthcare related predictions during COVID-19 originating from The Institute for Health Metrics and Evaluation (IHME), were used. Their data included predictions concerning the number of hospital beds needed, COVID-19 deaths, hospital admissions, and new ICU patients in the US and the Netherlands. Furthermore, all holiday dates in the US from 2016 to 2020 were used. Moreover, four features were created by extracting and modifying them from the data set provided by the airline: *Shift Year*, *Da*y, *Month* and *Schedule.*

Seasonality

Seasonality refers to fluctuations in the data that occur periodically, such as weekly, monthly or yearly. This is crucial to adopt the appropriate forecasting model. Overall, until the outbreak of COVID-19 in March 2020, the weight follows approximately the same trend. To gain more insight into the seasonality, an Error-Trend-Seasonality decomposition was made. Figure 1 displays the seasonal component of the data from January 2020 until April 2020. The graph reveals that every month the same cycle is repeated about four times, which implies weekly seasonality.

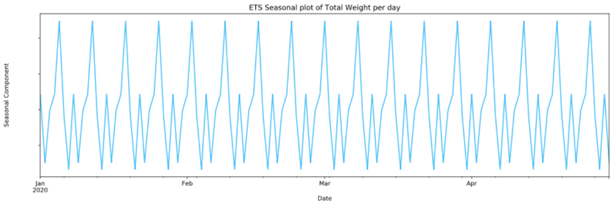


Figure 1 Seasonal component of total weight per day

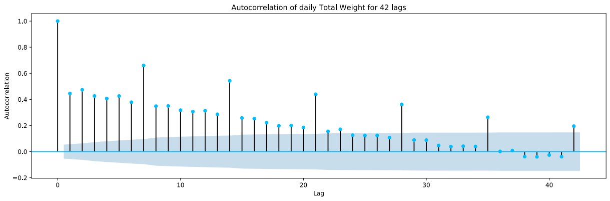


Figure 2 Autocorrelation for daily total weight for 42 lags

An alternative to check seasonality is an autocorrelation plot, which displays the correlation of the data with itself lagged by *x* time units. The blue shaded region represents the confidence interval. If the autocorrelation exceeds the confidence interval, it can be assumed that the autocorrelation value is statistically significant. Figure 2 presents the autocorrelation plot of the weight, shifted 42 times. The graph clearly reveals a high correlation each seventh lag, which, again, implies weekly seasonality.

Another autocorrelation plot was created for the autocorrelation of each individual data point in the data set. It was visible that the 365th lag is slightly higher than the surrounding lags and exceeded the confidence interval. Therefore, there is a significant correlation between the data now and the data shifted by one year. This could imply yearly seasonality.

Training, Validation and Test Sets

A training set, validation set, and test set were used to train the forecast models and make predictions on the daily total weight. The sets range from 2017-09-03 until 2020-04-16, 2020-04-17 until 2020-04-30, and 2020-05-01 until 2020-05-14 respectively. Moreover, two additional test sets were used. These range from 2020-05-08 until 2020-05-21 and 2020-05-15 until 2020-05-28. These additional test sets were used to study stability of the model. If the model's accuracy is close to each other in different time windows, we considered the forecast model stable. The same periods were taken for the year 2019, in order to see how the models performed without the COVID-19 pandemic. Thus, the validation and test set are set at two weeks and therefore the models generate 14-day predictions. The choice to predict 14 days into future was based on the often-changing policies and regulations during this pandemic. Moreover, it is customary in the air cargo industry to adhere to a booking window of two weeks, also in normal times outside fo the COVID-19 pandemic. Therefore, it is useful to make short-term, 14-day predictions.

METHODOLOGY

In order to make the predictions, ARIMA based models and LSTM models were used. For both type of models, the root mean square error (RMSE) was used as error measurement. To be able to compare accuracy of the different models, the RMSE as a percentage of the mean of the corresponding data was calculated. A low RMSE-percentage indicates a more accurate result.

ARIMA Models

A frequently used time series forecasting technique is the ARIMA (Autoregressive Integrated Moving Average) technique and its variations. One of the variations on the ARIMA model is the SARIMA model, which is used when the data is seasonal. Another variant is the SARIMAX model that is created by adding exogenous features to the SARIMA model. For both models the data has to be stationary. The SARIMA and SARIMAX models both use trend and seasonal elements. The trend elements are p, d and q. The parameter p represents the order of the autoregressive part, d represents the degree of the first difference involved, and q represents the order of the moving average part. The seasonal elements include P, D, Q, and m, in which P represents the seasonal autoregressive order, D represents the seasonal difference, Q represents the seasonal moving average order, and m represents the seasonal period.

LSTM Models

The LSTM (Long Short-Term Memory) model is a special kind of Recurrent Neural Network (RNN) and is able to learn long-term patterns in sequential data (*10*). After feeding the LSTM model with the historical observations of the target value (and exogenous variables), the model blends the information of the variable(s) into the memory cells and hidden states (*11*). This research used the Univariate and Multivariate LSTM models, also called the U-LSTM and MV-LSTM models. The U-LSTM model has the same input and target variable; thus, it looks at the feature's historical observations to predict the next time step. MV-LSTM has Multivariate time series data as input. This implies that there is more than one observation for each time step, which is the case after adding exogenous features. For a more descriptive explanation of the LSTM models Goel et al. (*10*) and Guo & Lin (*11*) could be consulted.

Both LSTM models contain hyperparameters. With the use of a grid search method, the values for the parameters were determined in this research. The sets for the number of neurons, number of epochs and the dropout rate were selected by trial and error. The grid search used all the available activation functions and optimizers form Keras. The used batch size for the grid search was 64, based on Guo et al. (*12*).

Feature Selection

Inspired by Karagiannopoulos et al. (*12*), Forward Selection (FS) was used to select the features. FS is a wrapper method that uses a greedy search approach by evaluating feature combinations against an evaluation criterion. The evaluation criterion used in this feature selection is the RMSE.

Since the SARIMAX model can only take exogenous features that contain future values for the to be predicted period of the target feature, the feature selection for this model is done among the COVID-19 related features, the variables *Day*, *Month*, *Shift year*, the US holiday dates feature, and *Schedule.*

Unlike the SARIMAX model, the MV-LSTM model uses exogenous features that do not contain future values. Therefore, the feature selection for the MV-LSTM model is done among all the exogenous features, except the variable *Schedule* due to inconvenience for the model. For the selection of the features, first the features were ranked by eXtreme Gradient Boosting (XGBoost). Thereafter, a top 1, 2 and 3 features were combined, and the model is run with the different combinations of features. After this procedure, the combination with the lowest RMSE is selected as the best set of features for the MV-LSTM model.

RESULTS: ARIMA

Parameter Selection

The first step in building a SARIMAX forecast model is choosing the right parameters for the ARIMA terms *p*, *d,* and *q,* and the seasonal terms *P*, *D*, *Q* and *m*. The auto ARIMA function was used to determine the values of these parameters. The value for *m*, denoting the number of observations per seasonal cycle, was already known. During the data exploration, a weekly seasonality was found for the daily weight. Therefore, *m* equals 7. Using this value for *m*, the auto ARIMA function fits the best SARIMA model according to the Akaike Information Criterion (AIC). This is a performance metric which estimates the quality of the model, relative to the other models. The function performs a search over possible parameters and selects the parameters that minimize the AIC. The values found for *p*, *d*, *q*, *P*, *D* and *Q* are 2, 1, 2, 2, 0 and 2 respectively.

SARIMA

A 14-day prediction was generated using a SARIMA model with the parameters described. For the validation data, the last two weeks of April 2020, the predictions gave a RMSE% of mean of 28.55%. The RMSE% of mean on the test data, the first two weeks of May 2020, is 26.49%.

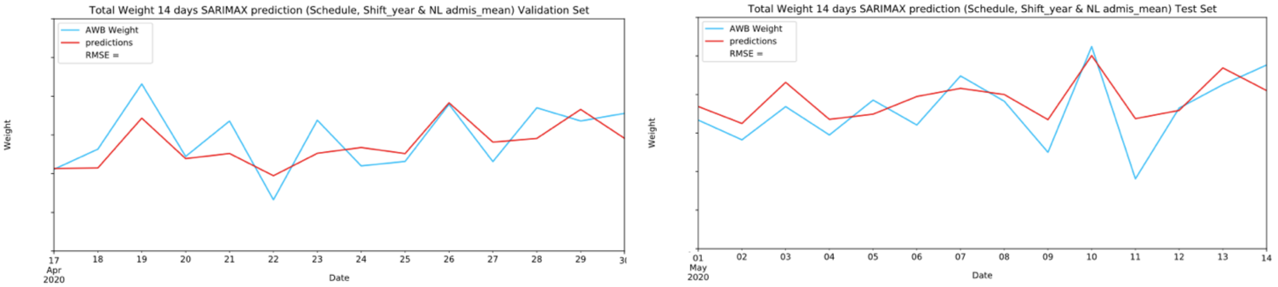
Feature Selection

The next step is to add the relevant exogenous variables to the SARIMA model and thereby, making it a SARIMAX model. Only variables for which future data is available can be added to the model. These variables should be stationary. Each possible variable was checked for stationarity using the Augmented Dickey-Fuller test. Since the data exploration of the daily weight suggested potential yearly seasonality, the variable *Shift Year* was included in the feature selection. To make the non-stationary features stationary, the features were differenced. Differencing involves calculating the differences between consecutive observations until the data is no longer non-stationary.

As discussed in the section *Methodology,* forward selection was used to decide on the most optimal features for the SARIMAX model. The final SARIMAX model, which returned the lowest RMSE on the validation data, includes only one feature, namely *US deaths_mean.* This model gave an RMSE% of mean of 25.77% on the validation data and 22.02% on the test data.

Schedule

A new feature selection was done, where the feature *Schedule* was selected at first before continuing the selection procedure. *Schedule* equals the total number of flights per day of the airline. It was added to study its influence on the SARIMAX model. Including this feature in the new SARIMAX model and using forward selection among the other features generates a final SARIMAX model, which contains the three features *Schedule*, *Shift Year*, and *NL admis_mean*. Adding these features led to a RMSE% of mean of 19.44% on the validation data. The graph in Figure 3 shows the predictions of this SARIMAX model on the validation set and on the test set. The predictions made on the test data gave a RMSE% of mean of 17.44%.



(a) Validation set

(b) Test set

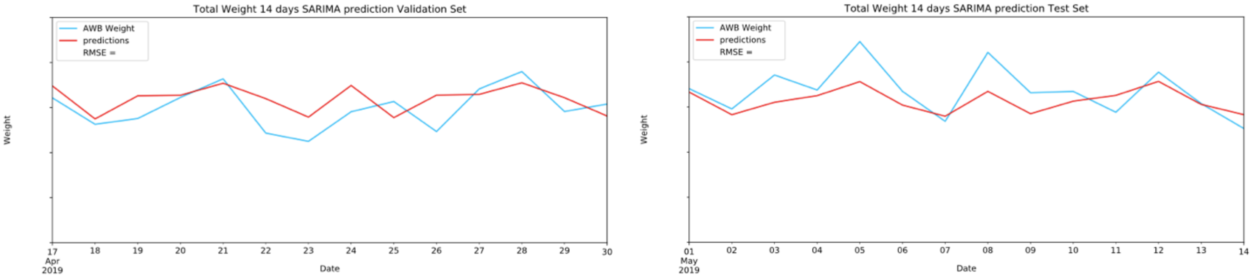
Figure 3 Total weight 14-day SARIMAX Prediction 2020

Since this SARIMAX model generates the most accurate predictions, this model was also used on two new test sets, which range from 2020-05-08 until 2020-05-21 and 2020-05-15 until 2020-05-28. The two new models adopt the same features and parameters as mentioned previously. The predictions of these SARIMAX models on the new test sets gave an RMSE% of mean of 18.84% and 18.24% respectively, indicating stable results under different test sets.

Excluding COVID-19

The same approach of the SARIMAX model was also tested in an everyday situation. Thus, excluding the outbreak of COVID-19. As a result, the train data ranges from 2017-09-01 until 2019-04-16, the validation data from 2019-04-17 until 2019-04-30, and the test data from 2019-05-01 until 2019-05-14.

First, the parameters are again determined using the auto ARIMA function. The values found for *p*, *d*, *q*, *P*, *D* and *Q* are 1, 1, 2, 1, 0 and 1 respectively. Before building the SARIMAX model, the SARIMA model was made. The predictions of this SARIMA model for the validation and test set are plotted in Figure 4 below. The corresponding RMSE% of mean on the validation data is 14.38% and 12.78% on the test data.



(a) Validation set

(b) Test set

Figure 4 Total weight 14-day SARIMAX Prediction 2019

Since this SARIMA model is the best performing model for the year 2019, it is also used on two new test sets mention in *Traning, Validation and Test Sets*. The SARIMA model on these new test sets work with the same parameters as mentioned before. The predictions of this model on these test sets gave an improved RMSE% of mean of 11.34% and 8.80% respectively.

Because of the absence of the pandemic, all COVID-19 related features are irrelevant in this approach. Therefore, the exogenous variables taken into consideration for the SARIMAX model are *Schedule*, *Shift Year*, *IS_HOL*, *Day*, and *Month*. Adding the feature *Month* to the initial SARIMA model decreased the RMSE. However, including any other exogenous variable in this model did not improve the RMSE. Therefore, the final SARIMAX model includes only the feature *Month*. This gave an RMSE% of mean of 13.16% on the validation data. The RMSE% of mean on the test data is 13.00%.

RESULTS: LSTM

Before showing the results, the parameter settings will be explained. The results for both U-LSTM and MV-LSTM consist of a training and validation part, and a testing part. Since MV-LSTM contains exogenous variables, a feature selection for this model will also be given. Within the results we will discuss the 'mean%' and the 'std.%'. With the 'mean%' we refer to the mean RMSE after 10 runs divided by the mean of the actual data for the selected period. With the 'std.%' we refer to the mean standard deviation of the RMSE after 10 runs divided by the mean of the actual data for the selected period.

Parameter Settings

The models use a single LSTM layer, single dropout layer and a single dense layer. During the validation and testing of the models we looked at 10 runs per model and did a grid search. The following sets for the parameters were used: Batch Size = {64}, Epochs = {1000}, Neurons = {10, 20, 40, 80, 100}, Dropout Rate = {0.0, 0.1, 0.2, 0.3, 0.4, 0.5}, Optimizers = {Adadelta, Adagrad, Adam, Adamax, Nadam, RMSprop}, Activation = {hard_sigmoid, linear, ReLu, sigmoid, softmax, softplus, softsign, tanh} and kernel initializer = {glorot_normal, glorot_uniform, he_normal, he_uniform, lecun_uniform, normal, uniform, zero}.

The first grid search was done on the data ranging from 2017-09-01 until 2019-04-16 and the second grid search was done on the data ranging from 2017-09-01 until 2020-04-16. Both grid searches were based on the U-LSTM. However, the best performing grid search result on the U-LSTM validation was also used on the MV-LSTM validation.

In total, the grid search returned 11521 different results. However, we solely analyzed the 5 best grid search results during the validation process. During the validation process for every set of parameters 10 runs were recorded. From these 10 runs the mean RMSE and the standard deviation of the 10 RMSE's were analyzed. The results are reported as percentages of the mean of the data.

Univariate LSTM

U-LSTM: Training and Validation

The parameters that came out the grid search for the validation set were also used on the trianing set. These grid search results are displayed in Table 1 and 2. For the validation on the year 2019, the parameters in Table 1 are used. For the validation on the year 2020, the parameters in Table 2 are used.

TABLE 1 Top 5 Grid search results used on the validation set 2019. The Mean% and the Standard Deviation % are based on 10 runs

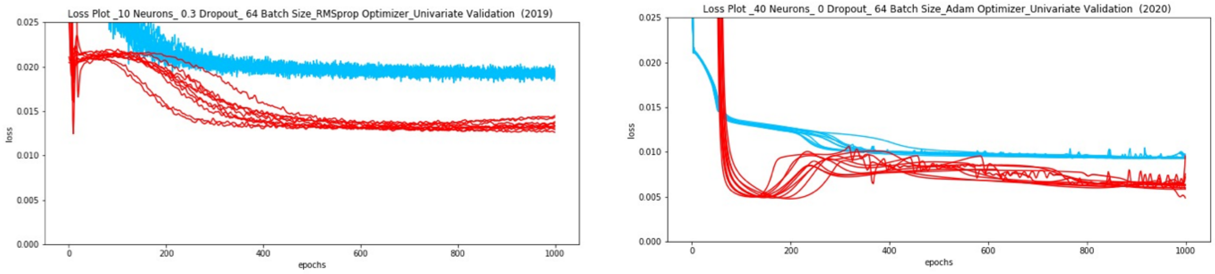
Epochs	Rank	Neurons	Dropout	Kernal Init.	Activation	Optimizer	Mean%	Std. %
1000	1	100	0.4	He_normal	Sigmoid	Adam	12.40%	0.11%
1000	2	80	0.3	He_normal	Sigmoid	Adam	12.49%	0.18%
1000	3	100	0.4	Lecun_uniform	Hard_Sigmoid	Adam	12.49%	0.11%
1000	4	100	0.3	He_uniform	Hard_Sigmoid	Nadam	12.64%	0.16%
1000	5	10	0.3	Lecun_uniform	Tanh	RMSprop	**12.22%**	**0.20%**

TABLE 2 Top 5 Grid search results used on the validation set 2020. The Mean% and the Standard Deviation % are based on 10 runs

Epochs	Rank	Neurons	Dropout	Kernal Init.	Activation	Optimizer	Mean%	Std. %
1000	1	40	0	Uniform	Softplus	Adam	**32.39%**	**1.27%**
1000	2	20	0	He_uniform	Softplus	Adamax	30.55%	2.44%
1000	3	40	0	Uniform	Softsign	Adam	35.27%	0.95%
1000	4	100	0.2	He_uniform	Softplus	Adam	29.91%	8.92%
1000	5	40	0.3	Glorot_uniform	ReLu	Adamx	31.45%	2.75%

Primarily we look for the lowest mean%. When the std.% is above 1.5%, we do not feel confident that 10 runs were enough to get stable results. With 10 runs and a high std.%, the mean can shift a lot if the 10 runs are repeated. A more reliable mean% can be obtained by choosing a low std.% or increasing the number of runs. However, increasing the runs takes more computing power. Therefore, a low std.% is considered here. Thus, in this case another set of parameters were selected with a lower std.% and a slightly higher mean%. This way, the predictions are more reliable.

After finding the best performing parameters for the validation set, we looked at the validation loss against the training loss in Figure 5. This was done to see if there is a significant better performing number of epochs with a smaller runtime, since more epochs result in more runtime. However, based on Figure 5 there is no significant indication to investigate other numbers of epochs for both 2019 and 2020.

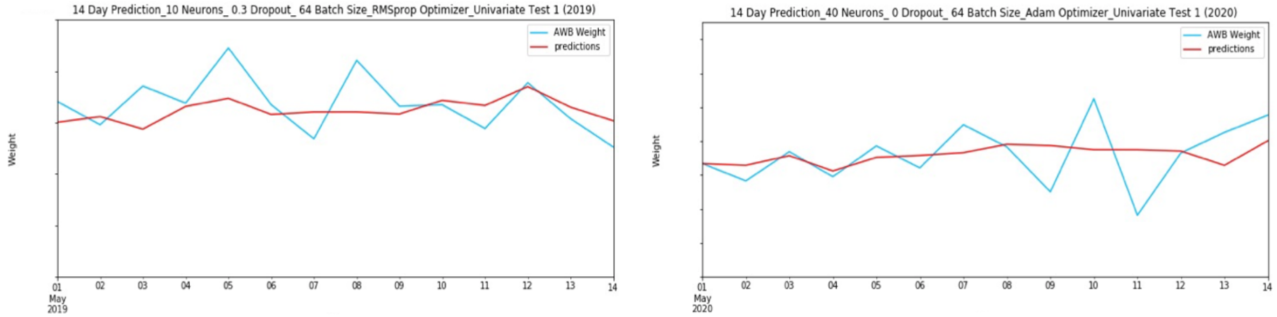


(a) 2019 (b) 2020

Figure 5 Training loss in blue versus the validation loss in red for the chosen parameters

U-LSTM: Testing

For testing on 2019 and 2020 the highlighted parameters in Table 1 and 2 were considered. Again, 10 runs were taken into account for the randomness in the model. Feeding the LSTM models with the values from Table 1 and Table 2, the following results were given. Figure 6 shows an example prediction for both 2019 and 2020 on test set 1. For 2019 the mean% is 15.30% and the std.% is 0.14%. For 2020 the mean% is 23.89% and the std.% is 0.52%.



(a) 2019 (b) 2020

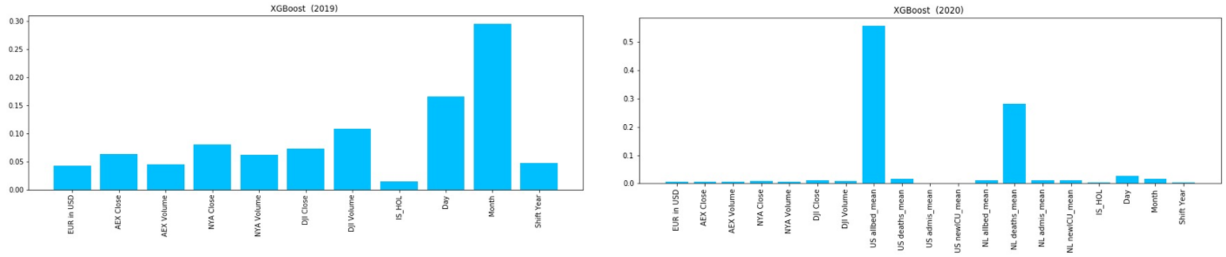
Figure 6 The 14-day prediction plot on test set 1 for U-LSTM

To verify the performance of the models, two additional test sets were added. For test set 2, the mean% is 11.22% and the std.% is 0.13% in 2019. For 2020 the mean% is 24.55% and the std.% is 0.76%. In 2019 for test set 3, the mean% is 9.93% and the std.% is 0.17%. For 2020 the mean% is 17.05% and the std.% is 0.62%.

Multivariate LSTM

Feature Selection

Before adding variables to the LSTM model, we performed feature selection. For our feature selection the XGBoost weights were taken. This was done for both 2019 and 2020, which can be seen in Figure 7. During our research we looked at the features with the highest score within XGBoost and made a top 1, 2 and 3 out of the best scored features.



(a) 2019

(b) 2020

Figure 7 XGBoost scores for the exogenous features

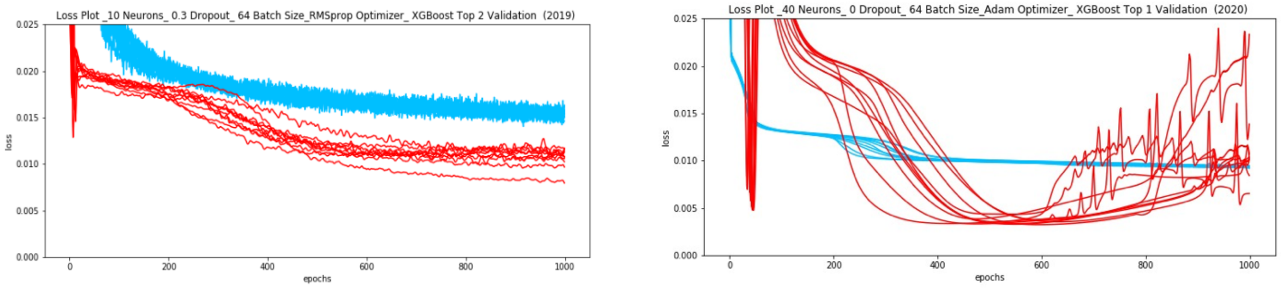
MV-LSTM: Training and Validation

The training for the MV-LSTM was done in the same manner as the U-LSTM. Therefore, the set of parameters used here are the same. In Table 3 the feature selection is recorded on the validation data.

TABLE 3 Feature selection on validation sets for both years based on XGBoost score

Year	**Tops**	**Features**	**Mean%**	**Std.%**
2019	Top 1	Month	11.37%	0.42%
2019	Top 2	Month, Day	11.04%	0.51%
2019	Top 3	Month, Day, DJI Volume	11.92%	0.69%
2020	Top 1	US allbed_mean	42.42%	4.63%
2020	Top 2	US allbed_mean, NL deaths_mean	51.22%	5.77%
2020	Top 3	US allbed_mean, NL deaths_mean, Day	44.50%	6.55%

After finding the best performing parameters for the validation set, we looked at the validation loss against the training loss in Figure 8. This is done to see if there is a significant better performing number of epochs with a smaller runtime, since more epochs result in more runtime. Based on Figure 8 there is no significant indication to investigate other numbers of epochs for 2019. However, Figure 8 indicates a possible better solution around 550 epochs for the year 2020. By using 550 epochs the new mean% becomes 24.57% and the new std.% becomes 1.14%. As can be noticed the standard deviation is below the 1.5%, which makes the spread of the predictions more reliable.

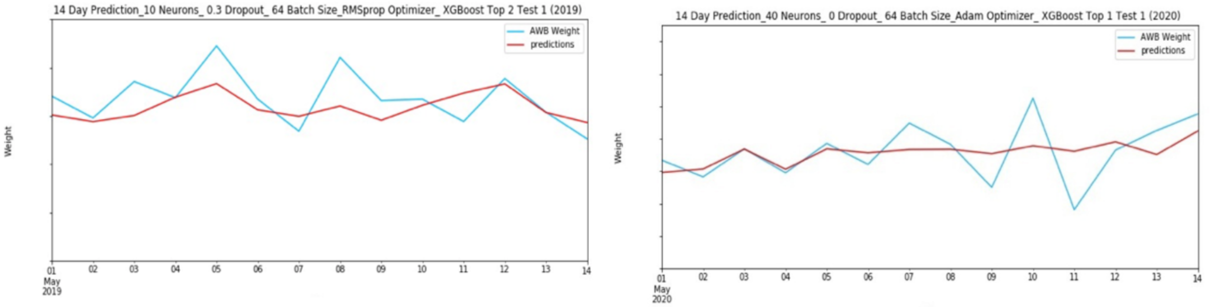


(a) 2019 (b) 2020

Figure 8 Training loss in blue versus the validation loss in red for the MV-LSTM

MV-LSTM: Testing

During the testing of the MV-LSTM the highlighted parameters in Table 1 and 2 are considered. In addition, the added exogenous features are highlighted in Table 3. Again, two additional test sets were added. Figure 9 shows an example prediction for both 2019 and 2020 on test set 1. For 2019 the mean% is 13.66% and the std.% is 0.69%. For 2020 the mean% is 21.54% And the std.% is 0.92%.



(a) 2019 (b) 2020

Figure 9 The 14-day prediction plot on test set 1 for MV-LSTM

To verify the performance of the models, two additional test sets were added. For test set 2, the mean% is 12.21% and the std.% is 0.58% in 2019. In 2020 the mean% is 22.44% and the std.% is 0.37%. Test set 3 has a mean% of 9.34% and std.% of 0.65%. For 2020, the mean% is 17.14% and the std.% is 0.41%.

Overview of Results

Table 4 gives an overview of the above-mentioned results per model for test set 1.

TABLE 4 Overview results

Model	**RMSE% of mean (ARIMA) / Mean% (LSTM)**	**Std.% (LSTM)**
SARIMA 2019	12.78%	-
SARIMA 2020	26.49%	-
SARIMAX 2019	13.00%	-
SARIMAX 2020	22.02%	-
SARIMAX 2020 *Schedule*	17.44%	-
U-LSTM 2019	15.30%	0.14%
U-LSTM 2020	23.89%	0.52%
MV-LSTM 2019	13.66%	0.69%
MV-LSTM 2020	21.54%	0.92%

DISCUSSION

This discussion will elaborate on the differences between the results of the models. First the results of the ARIMA-based models will be discussed. Therafter, a discussion for the LSTM models will be given, followed by a comparison between the ARIMA-based models and the LSTM models.

ARIMA

SARIMA versus SARIMAX

It can be interpreted that the RMSE-percentages of the 2019 models are very close to each other. For 2019 the SARIMA model (12.78%) performs slightly better in comparison to the SARIMAX model (13.00%). This indicates that the exogenous variable added to the SARIMAX model does not contribute to more accurate predictions. For 2020 however, the SARIMAX model (22.02%) generates better results than the SARIMA model (26.49%). Therefore, the inclusion of the exogenous features, which included the healthcare related predictions during COVID-19, does result in more accurate predictions.

Furthermore, both the SARIMA and the SARIMAX model perform better in 2019 than in 2020. This is a result of the outbreak of COVID-19, which has caused instability within the data. Since the data in the year 2017, 2018 and 2019 differ substantially from the year 2020, the models do not predict the unexpected peaks.

Schedule versus No Schedule

The SARIMAX model for the year 2020 including the flight schedule of the airline (17.44%) is more accurate than the regular SARIMAX 2020 model excluding the flight schedule (22.02%). Since SARIMAX 2020 with *Schedule* is the best performing model, it is also tested on two additional test sets. The RMSE-percentages (18.84% and 18.24%) do not significantly fluctuate using the new test data and therefore, this model performs well and is stable.

U-LSTM versus MV-LSTM

Since the std.% for all models in both years were low, we will only compare the models by looking at the mean%. For 2019, MV-LSTM (13.66%) performs better than the U-LSTM (15.30%) on test set 1. When looking at the percentages of the additional test sets (12.21% and 9.34%) for MV-LSTM, it can be seen that the model even performs slightly better. Moreover, the percentages do not significantly fluctuate and therefore this model could be considered well-performing and stable. For the year 2020, again the MV-LSTM (21.54%) performs better than the U-LSTM (23.89%). The percentages of the additional test sets (22.44% and 17.14%) for MV-LSTM show that the model performs worse for test set 2 and better for test set 3. Even though the MV-LSTM performs better in both years, it is clear that the predictions are less accurate in 2020, where COVID-19 has caused irregular peaks.

SARIMA(X) versus MV-LSTM

Among the ARIMA based models, SARIMA and SARIMAX with *Schedule* gave the most accurate predictions. These models are compared to the MV-LSTM, which performed the best among the LSTM models. In 2019, the SARIMA model (12.78%) is slighty more accurate than the MV-LSTM (13.66%). For the year 2020, the SARIMAX with *Schedule* (17.44%) outperforms the MV-LSTM (21.54%). Hereby, it can be said that the ARIMA based models perform better than the LSTM models in this research.

CONCLUSION AND RECOMMENDATIONS

Conclusion

The aim of this research was to make short-term predictions of the air cargo demand between a major European airport hub and the United States during the COVID-19 pandemic. The data in this report were provided by a major commercial airline. The data consisted of the daily weights of air cargo transported from the airport hub in Europe to several cities in the United States. In addition to the data facilitated by the airline, exogenous variables have also been used in order to improve the performance of the forecast models. These exogenous variables included, among others, stock market indices, foreign currency exchange rates, healthcare related predictions during COVID-19, and the Airline X flight schedule.

The two types of models built and analyzed in this report were ARIMA-based models and LSTM models. For 2020, the best performing model among the ARIMA-based models is the SARIMAX with *Schedule*. During the year 2019, where the outbreak of COVID-19 is excluded, the SARIMA model generates the most accurate predictions. Among the LSTM models, the Multivariate LSTM is more accurate than the Univariate LSTM in both 2019 and 2020. However, in this research it can be concluded that the ARIMA-based models perform better than the LSTM models.

Recommendations

This paper generated predictions by using different models. However, more extensive research could be done by improving or expanding the existing models. This could be done by taking the suggestions mentioned below into consideration.

Scale Down to Regions

In future research one could look at a specific region in the USA when predicting the demand for air cargo transported from the European airport hub. When predicting the air cargo demand to the USA, there are both advantages as well as disadvantages in considering the total demand to the entire USA instead of scaling back to a certain region. The advantage of considering the entire country is that the demand will fluctuate less. Moreover, many publicly available features are based on the whole country. However, the disadvantage of considering the entire US is that the situation in one region could be very different than the situation in another part of the country. Therefore, studying local demand can help to tailor air cargo operations to a specific situation in a specific region.

LSTM

During this research, future variables have been added to the SARIMAX model, but not to the LSTM model. When the exogenous variable *Schedule* was added to the SARIMAX model, the model performed better. The most optimal forecast model created thus far is this SARIMAX model with future variables added. Therefore, adding future variables to the LSTM models could improve the results and potentially create a more optimal model.

In the discussion the spread of the predictions is already mentioned. Therefore, multiple runs are considered. However, more than 10 runs are desired to make the outcome more reliable. Especially for the MV-LSTM model more runs could improve the results with a great deal.

In this paper we used a limited grid search. Increasing the number of parameters would increase the runtime, but this is necessary to cover a more representable grid search. We would suggest adding more epochs to the gridsearch and adding higher numbers of neurons. Moreover, we only used two different grids searches. These grid searches were solely performed on the U-LSTM model. It can be useful to use the grid search for the MV-LSTM model as well.

Furthermore, it would be interesting to take Granger causality into consideration during the feature selection in future research. Granger causality is used to investigate causality between two variables in time series. For example, changes on the stock market indices on time t-x could have causality with the freight demand on time t, where t and x are in days. Therefore, the features related to the stock market indices could be shifted with x days and used in the model.

Long Term Predictions

The predictions included in this report are all 14-day predictions. In addition, the data is aggregated by day, which means that all graphs show a single data point per day. When predicting further into the future, the predictions could be aggregated by week or by month. This leads to a longer-term prediction period. Prolonging the prediction period could generate more stable results, as the fluctuation declines. However, the longer the prediction period, the less accurate the forecast. Thus, this has to be taken into account when going through this decision-making process.

Multiple Validation and Test Periods

In order to work with the forecast models, the data was divided into a training, validation, and test set. Throughout this research, there was only one validation period used per model. Adding more validation periods could increase the accuracy of results. Namely, these extra periods would make sure that the models anticipate better on unexpected peaks in the data. Regarding the test periods, this research considers three different test sets. However, these test sets are all within the same month and therefore, it would be more reliable for future research to analyze different time periods.

ACKNOWLEDGMENTS

We would like to thank the major commercial airline for their cooperation during this research. We would also like to thank the Faculty of Science at the Vrije Universiteit Amsterdam for giving us the opportunity to work on the Business Case, which resulted in this paper. Hereby, we also thank C.E.L. Jonkman and T.H.W. Stultiens, who were our team members. Furthermore, we are grateful for the knowledge A. Abdi and C. Amrit provided about LSTM models. Finally, we also appreciate the help of E. de Jong and L. Kunz, who gave us insights into the transportation sector.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: Study conception: E.R. Dugundji, T.H.A Koch; Study design: A. Devaraj, N.K. van Hout, B. Verhoeven, H. Zwitzer, T. Crapts, A. Ion, E.R. Dugundji, T.H.A Koch; Data collection: A. Devaraj, N.K. van Hout, B. Verhoeven, H. Zwitzer, T. Crapts, A. Ion; Analysis of results: A. Devaraj, N.K. van Hout, B. Verhoeven; Interpretation of results: A. Devaraj, N.K. van Hout, B. Verhoeven, H. Zwitzer, T. Crapts, A. Ion, E.R. Dugundji, T.H.A Koch; Draft manuscript preparation: A. Devaraj, N.K. van Hout, B. Verhoeven. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. IATA. (2020). Action Cargo: COVID-19. https://www.iata.org/en/programs/cargo/. Accessed May 5, 2020.
2. Leigh, G. (2020, March 23). The Latest On Which Airlines Are Still Flying And Why. https://www.forbes.com/sites/gabrielleigh/2020/03/23/the-latest-on-which-airlines-are-still-flying-and-why/#3ea349ad1ffc. Accessed May 5, 2020.
3. Marazzo, M., Scherre, R., & Fernandes, E. (2010). Air transport demand and economic growth in Brazil: A time series analysis. *Transportation Research Part E: Logistics and Transportation Review*, *46*(2), 261-269.
4. Chi, J., & Baek, J. (2013). Dynamic relationship between air transport demand and economic growth in the United States: A new look. *Transport Policy*, *29*, 257-260.
5. Hathurusingha, C. J., & Mudunkotuwa, M. R. S. (2015). Time Series Approaches to Forecast Air Freight Imports and Exports: Empirical from Sri Lanka.
6. Chen, S. C., Kuo, S. Y., Chang, K. W., & Wang, Y. T. (2012). Improving the forecasting accuracy of air passenger and air cargo demand: the application of back-propagation neural networks. *Transportation Planning and Technology*, *35*(3), 373-392.
7. Baxter, G., & Srisaeng, P. (2018). The use of an artificial neural network to predict Australia’s export air Cargo demand. *International Journal for Traffic and Transport Engineering*, *8*(1), 15-30.
8. Su, H., Zio, E., Zhang, J., Xu, M., Li, X., & Zhang, Z. (2019). A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced Deep-RNN model. *Energy*, *178*, 585-597.
9. Chou, T. Y., Liang, G. S., & Han, T. C. (2011). Application of fuzzy regression on air cargo volume forecast. *Quality & Quantity*, *45*(6), 1539-1550.
10. Goel, H., Melnyk, I., Oza, N., Matthews, B., & Banerjee, A. (2016). Multivariate Aviation Time Series Modeling: VARs vs. LSTMs. *Unpublished manuscript. Retrieved from https://www. semanticscholar. org/paper/Multivariate-Aviation-Time-Series-Modeling*, *3*.
11. Guo, T., & Lin, T. (2018). Multi-variable LSTM neural network for autoregressive exogenous model. *arXiv preprint arXiv:1806.06384*.
12. Guo, T., Lin, T., & Antulov-Fantulin, N. (2019). Exploring interpretable lstm neural networks over multi-variable data. *arXiv preprint arXiv:1905.12034*.
13. Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). Feature selection for regression problems. *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece*, *2022*.