Loyola University

2nd Annual Supply Chain Data Analytics Competition

Charlie Marshall & Aarij Rehman

Assumptions

Throughout our analysis, we had to make certain assumptions about the Freighwaves and Shipper data. Those assumptions and their motivations are described below.

Dates:
We were often asked to conduct analysis within a specific year. However, in certain cases, we found that the Book Date, Pickup Date, and Delivery Date were not all within the same year. This would usually happen when a shipment was booked near the end of December, and it wasn't picked up or delivered until the following January. To address this, the 'date' an order falls into relies exclusively on the Book Date. This is because we're assuming all the details regarding an order are determined on that day as well.

Truckload:
We interpreted the "Truckload" category in question 1 to mean only Trucks or Dry Vans. We found evidence to suggest that some parties would include Flat Beds or Temperature Controlled, but this is not how we interpreted it.

When the Mileage or Revenue was Zero:
We considered cases where either the revenue or mileage of a load was either zero or blank to be an error in input. Often, these cases would cause the RPM to be calculated as infinity or zero. To address this, we removed all loads where the mileage or revenue was blank or zero.

Volatility:
In one question, we were asked to consider lanes by volatility. As is standard in finance, we calculated volatility by using the standard deviation.

Volume:
We interpreted lane volume in question 1b as the count of the number of loads which used a specific lane.

Rate Decision:
In our analysis, we assumed that all rate decisions were made on the most recently reported FreightWaves indexes based on the Book Date. For instance, if a load has a Book Date of March 10th, it would reference the February 28th report of many of the Macro indexes such as ISM.

Computational/Statistical Questions:

These questions involved relatively simple calculations, so we won't go into detail in how they were determined.

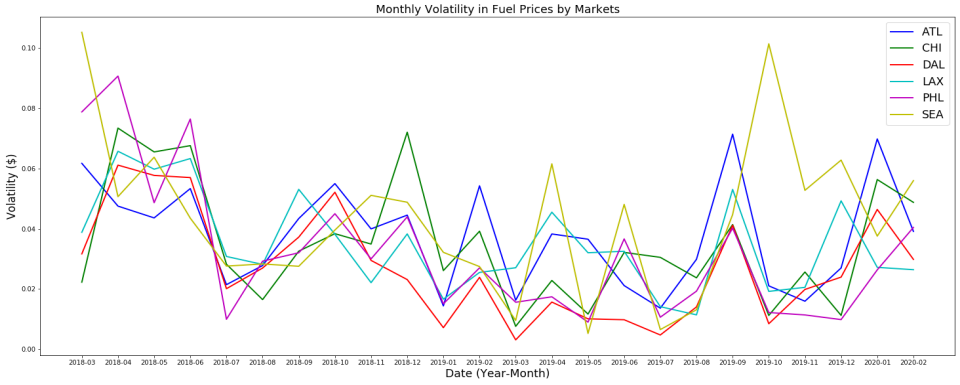
What was the overall average Truckload rate per mile in 2019?

$3.69/mile

Here we rank the 7 FreighWaves Futures Lanes by 2019 rate volatility, Truckload Volume, and Intermodal Volume from highest to lowest.

Rate Volatility	**Truckload Volume**	**Intermodal Volume**
VPC	VCA	VLD
VCA	VSL	VDL
VLS	VDL	VCA
VAP	VPC	VPC
VLD	VLS	VLS
VDL	VLD	VSL
VSL	VAP	VAP

Here, we graph monthly volatility in fuel prices by markets provided in the FreightWaves data.

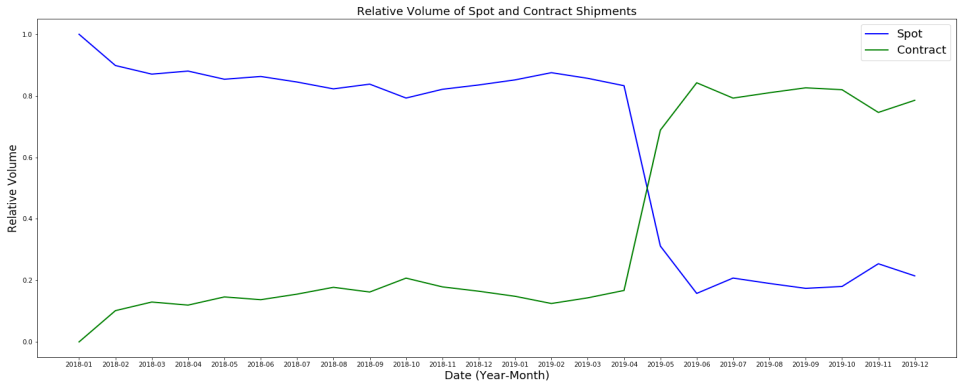


One market that stands out is Seattle because it has a large spike in fuel price volatility around October 2019 when all other markets were seeing decreasing price volatility. However, it is interesting to see that in most cases the price volatilities of different markets increase and decrease at similar times, which suggests that a Macroeconomic scale which measures volatility across the country might be accurate for larger specific markets as well.

Spot vs Contract Mixture

When looking at the mix of Spot and Contract rates for 2018 and 2019, we considered the mix of those rates for each individual month. Specifically, we parsed through the data to look only at Dry-Van and Truck freight for every month, and we would calculate what percentage of

shipments in each month would come from Spot Rates and what percentage would come from Contract Rates. We plotted the ratios for each month as seen below:



One important consideration is that for every month, each of the data points will always sum up to 1 because they are relative frequencies. This is especially apparent in January of 2018 when all the shipments in the data were conducted at spot rates and none at contract rates.

Promising External Factors

In order to find promising external factors from the FreightWaves data for predicting future shipping rates for Dry Vans, we created a data frame which related rates from loads in the Shipper Data to the indexes reported in the FreightWaves data. In order to do this, we operated under the Rate Decision assumption, that the negotiated rates referenced the most recently reported FreightWaves indexes. In order to weigh all factors equally, we used a subset of the full data which included only loads in which all FreightWaves indexes were relevant. Then, we used a machine learning feature selection technique called Recursive Feature Elimination (RFE) to select the three external factors which best predict future shipper rates.

RFE works by fitting a simple linear regression model and recursively removing the features based on feature importance attributes until the specified number of features, in this case 3, is reached. RFE tries to account for dependencies between features when eliminating, so the final features are not highly correlated. In our case, Recursive Feature Elimination found IPRO.USA, TLT, and DATVF to be the most representative features.

We chose to use RFE over other commonly used feature selection techniques such as Backwards Elimination, Forward Elimination, or Lasso because it gave us the ability to specify the number of factors, whereas the other feature selection methods do not.

Additional Insights

For additional insight into the data, we took the opportunity to develop statistical models to predict future rates. In order to do this we subsetted the data into 5 categories based on the length in miles of the shipment: city (less than 100 miles), short (between 100 and 250 miles), mid (250 - 450 miles), tweener (450 - 800 miles), and long (over 800 miles). These distances were not arbitrarily selected; they are commonly used distances to subset many of the FreightWaves indexes. The data was broken up because statistical models do not perform well when fed partial information in the form of NA's in the data set. These show up because the vast majority of the trips in the Shipper Data set did not occur in one of the seven specific Freight Futures Lanes, and therefore did not have values for the lane-dependent indexes. Splitting the data into sets based on distances allowed us to remove the lane-dependent indexes from some datasets, while keeping it in others.

Additionally, a few categorical variables were added to the dataset from the Shipper Data set. These include the type of contract, which is either Contract or Spot, and the Load Type, which is either Ad-hoc, Consistent, Critical, Dedicated, Standard, or White Glove. These factors were added to improve the models ability to correctly predict the rates.

Next, Recursive Feature Elimination was utilized to find the optimal features to include in the model. Unlike in the previous question when we used RFE, we do not need to specify a particular number of factors because we are checking every combination of numbers of factors by utilizing a for loop. RFE found the following 21 factors to be best at predicting future rates distances without lanes were evaluated:

ism, ipro.usa, ipro.fbevt, ppi, cosp, reslg, orders, otvi, otri, haul, tlt, dts, cstm, contract_Contract, contract_Spot, load_type_Ad-hoc, load_type_Consistent, load_type_Critical, load_type_Dedicated, load_type_Standard, load_type_White Glove

RFE found the following 23 features to be best at predicting feature rates when shipping was done on one of the 7 Freight Futures Lanes:

ism, ipro.usa, ipro.fbevt, ppi, cosp, reslg, otvi, otri, haul, tlt, dts, cstm, datvf, orail53l, intrm, contract_Contract, contract_Spot, load_type_Ad-hoc, load_type_Consistent, load_type_Critical, load_type_Dedicated, load_type_Standard, load_type_White Glove

After the correct features for each model were found, the model was trained on 80% of the data, then it was tested on the other 20% of the data it had never seen before. Each statistical model's ability to predict the rates on never before seen data for a specified distance is reported below in Table 2.

Distance Qualifier	Mean Absolute Error (MAE)	Mean Squared Error (MSE)
Lanes	0.422	0.591
City	4.319	92.374
Short	1.108	2.617
Mid	0.730	1.419
Tweener (Non-Lane)	0.540	0.739
Long (Non-Lane)	0.492	0.413

The Mean Absolute Error measures the average absolute between the predicted value and the actual value of the rate. Whereas the Mean Squared Error measures the average squared difference between the predicted value and the actual rate. It is similar to variance. In both cases, it is ideal for the value to be close to zero. From the table, it is obvious that the model performs significantly worse on City shipping based on larger MAE and MSE values. Steps were taken to improve the model including adding polynomial terms, using a log transformation, and adding more categorical variables. None of these additions saw significant improvements to the model.

Future Developments

One limitation in our assumption that shippers and carriers used the most recent FreightWaves index values to make rate decisions is with contracts. With a contract, the rate is determined at the beginning of the contract based on FreightWaves index values at that time and predictions for the future. However, we found it impractical to identify loads belonging to the same contract due to the structure of the data we were given, and therefore, we were not able to find the first instance of a contract. If we were able to, we would've made all loads belonging to the same contract reference FreighWaves indexes which their rate decisions were made on when the contract was signed.

In the future, we would take the opportunity to try different features to see if we could improve the statistical models. For instance, we might analyze how some of the indices changed over specific periods of time including compared to a year prior, a month prior, or a week prior. The percent increase or decrease of particular factors might give a better insight into shipper sentiment than the indexes we are currently examining.