USING DATA ANALYTICS & MACHINE LEARNING TO DESIGN BUSINESS INTERRUPTION INSURANCE PRODUCTS FOR RAIL FREIGHT OPERATORS

John F. Cardona
IN3 – Dept. of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
Juliana Castaneda
IN3 – Dept. of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
Leandro do C. Martins
IN3 – Dept. of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
Mariem Gandouz
IN3 – Dept. of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
Angel A. Juan
IN3 – Dept. of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain
Guillermo Franco
Guy Carpenter & Company LLC, New York, USA

ABSTRACT

This paper discusses a case study in which publicly available data of a rail freight transportation firm has been gathered, cleansed, and analyzed in order to: (*i*) describe the data using statistical indicators and graphs; (*ii*) identify patterns regarding several Key

Performance Indicators; (*iii*) obtain forecasts on the future evolution of these indicators; and (*iv*) use the identified patterns and the generated forecasts to propose customized insurance products that reflect the current and future freight transportation activity. The paper illustrates the different methodological steps required during the extraction and cleansing of the data --which required the development of Python scripts--, the use of time series analysis for obtaining reliable forecasts, and the use of machine learning models for designing customized insurance coverage from the identified patterns and predicted values.

1. INTRODUCTION

Data has become an essential element for the operational development and economic growth of many organizations throughout the world. As organizations across the world reach a level of economic abundance, their capacity to efficiently and effectively manage data has become a matter for concern (Parr-Rud, 2012). As a result, organizations from all industries and fields, and those which contribute to an emerging economy are welcoming innovative solutions and methods for managing their volumes of data. For instance, organizations from the railroad freight transportation sector in the United States (U.S.) have experienced substantial data growth throughout the years. This valuable information can be used for

identifying factors which can prevent an organization from reaching and maintaining economic growth. According to Samson and Previs (1999), the first train freight transport service contracted in the U.S. took place in 1827. Shipments of cargo were less frequent, and hence, the volume of standard goods being transported were minimal during this period. As a result, risks and incidents that occurred during transportation were minimal. On the other hand, train freight carrier services extend beyond transporting standard items and commodities. Train freight carriers have also committed to cautiously and effectively transport dangerous goods (DG) and hazardous materials (HAZMAT) which are essential for world economies. These materials are employed for powering vehicles, homes, and businesses, and for producing pharmaceuticals, sterilizers, disinfectants, fertilizers, and pesticides. In addition, these highly demanded items require cautious and effective handling and containment. Moreover, train freight carriers have encountered unforeseen challenges during transportation of standard items, DG, and HAZMAT. Increased risks of accidents, loss of goods, and adverse environmental consequences have been the most significant challenges. Therefore, sustaining risks awareness initiatives by assessing and anticipating risks which could occur during train shipments could prepare freight carriers for these unforeseen events. For this reason, the design of Business Interruption Insurance products for protecting the train freight industry from unforeseen circumstances and calamities is imperative.

The proposed insurance product or mechanism would parametrically pay (i.e., based on the collected data over time) when the actual performance falls below the agreed performance thresholds with respect to an established forecast. This type of mechanism would obviate the need to conduct any type of claim or forensic analysis and therefore permit the insured to recover without delay and without the peril of legal disputes.

Substantial data growth which is attributed to the railroad freight transportation sector implies that there is a high demand for the services provided by railroad freight operators.

Therefore, such organizations can profit from the abundance of valuable information extracted from databases by using data analytics not only for examining data, but also for attempting to identify patterns and for forecasting future events. In consideration of the foregoing, this paper will address the use of data analytics and machine learning methods for designing Business Interruption Insurance Products. As a result, data from a U.S. public train freight transportation organization has been extracted and analyzed by employing statistics and graphs. In addition, forecasts were processed for identifying patterns regarding key performance indicators (KPIs), and for supporting the design of business interruption insurance products aimed at safeguarding losses during train freight transportation.

Currently, data analytic methods could be considered the most efficient and effective solution for managing data (Ji and Wang, 2017). In the same manner, machine learning has contributed to the optimization of data analytic methods.

In order to create new knowledge and accomplish the goal of this study, the development of a forecasting model based on historical data that establishes the baseline performance of the freight business at the time of underwriting the policy is proposed. The predicted performance would then be compared with current performance developments.

The presented forecast method could be used for supporting the devising of customized parametric insurance policies that trigger a payout once verifiable conditions aimed at safeguarding losses during train freight transportation are satisfied. These policies could be devised by stipulating a payout that considers the real transported goods against expected transported goods over a certain period of time. In this manner, the policy would behave as a compensation for aggregate drops rather than as a payout for a momentary drop in transported cargo at a particular point in time.

The Methodology illustrates data extraction and cleansing based on the development of Python scripts, and consists of descriptive and predictive data analytics. Subsequently, time series analysis for forecasts on the evolution of these indicators have been provided, which together with previous results, are used with machine learning models for proposing customized insurance products based on identified patterns and predictive values, and for reflecting the current and future freight transportation activity. Subsequent to the Introduction, this paper is arranged as follows: Section 2: Literature Review presents a synopsis of related topics; Section 3: Problem Definition describes the problem contextualization; Section 4: Methodology introduces the tools and methodology for gathering and analyzing the data; Section 5: Results and Discussion provides a data and forecasting analysis of the future; and to conclude, Section 6: Conclusions and Further

Developments highlights the main conclusions of this investigation and proposes future research guidance.

2. LITERATURE REVIEW

The freight transportation industry continuously encounters risks such as loss of cargo due to business interruptions. Addressing these risks has posed challenges in carrying out logistics processes. Risk management has existed for decades and has been essential for logistics processes and systems. When transportation companies are faced with cargo losses, the consequences directly affect the cargo owners. For this reason, companies acquire cargo insurance that provides protection and entitlement to monetary compensation (Wu et al. 2017). Business interruptions affect both the insurance sector and global industries (Mizgier et al. 2018), forcing insurance companies to replace their business models with those that meet financial obligations as a result of an accident (Ganapathy, 2017).

However, few studies have focused on business interruption insurance although it has significant optimization and digitization potential through the use of big data and data analytics (Dong and Tomlin, 2012; Ganapathy, 2017). This literature review presents an overview of existing work on business interruption insurance and data analytics for the freight transportation industry.

According to Gagatsi et al. (2014), the devising of transportation insurance policies is an intricate process requiring extraordinary diligence from all stakeholders and affected economic sectors. Moreover, business interruption insurance products protect against losses incurred during company operational interruptions as a result of unexpected events occurring within their own sites or the suppliers' facilities. Business interruption insurance products are devised to include clauses describing explicit coverage limitations that define three crucial elements: (*i*) the premium which represents the price paid by the interested company for obtaining insurance coverage; (*ii*) the coverage limit which represents the maximum amount paid by the insurer in case of a loss; and (*iii*) the deductible which represents the monetary value of the loss absorbed by the insured company. Prior to establishing prices for customized business interruption insurance products, insurance providers must access the facilities of prospective insurance holders for assessing possible loss values (Dong and Tomlin, 2012).

Keller et al. (2018) have reported that data analytics has played a fundamental role in insurance policies allowing them to evolve from "intuitive bets" on the future to an industry based on logic calculus and decision making. They also considered that recent advances in big data analytics, artificial intelligence, and the internet of things promise to transform the role of data by understanding risks and protecting insurance product holders by providing compensation for incurred losses. According to Frees (2015), "insurance is a data-driven industry" which is linked to data and models of uncertainty. Insurance is also of a random nature, given by the very concept of payment, compensation, risks, as well as their grouping and dispersion. Understanding its benefits leads to the relationship of all factors involved giving rise to stochastic models that have correlations for measuring dependencies between random outcomes. A statistical data analysis pioneer who aimed at investigating the distribution of business interruption products was Zajdenweber (1996) from the French

Insurance Syndicate. Zajdenweber (1996) analyzed the consequences of Pareto's alpha exponent law when the tail was close to one on the actuarial risk. Pareto's law asserts that 80% of outputs results from 20% of all inputs. This implies that premium calculations for 100% insurance coverage require an uppermost limit corresponding to the highest potential loss. He also concluded that the insurer and reinsurers' financial stability requires funds for coping with annual fluctuations and unforeseen events. Dong and Tomlin (2012) explored the relationship between business interruption products and operational measures such as inventory and emergency sourcing as strategies for managing business disruption risks.

Moreover, an endogenous insurance pricing model was used for providing reliable insurance coverage and for supporting operational decisions. The results highlighted the importance of linking effective decision making and risk management practices into operational processes.

According to Frees (2015) description of the contributions of analytical and statistical methods for insurance market operations, analytical predictions are advanced data mining tools. Among the applied methodologies are the neural networks, the classification trees, and non-parametric regression statistical methods. An example of this is the methodology for calculating the Fuzzy net present value proposed by Neto et al. (2012). The net present value verifies the viability of purchasing a business interruptions insurance product for an offshore production unit. Moreover, it is the result of the discounted cash flow which gives rise to uncertainty with the available information. In Zurich, Switzerland, Mizgier et al. (2018) developed a collaborative project between Zurich Insurance and the Swiss Federal Institute of Technology. A large amount of data concerning business interruption claims from various data sources was extracted and analyzed. The results enabled Zurich Insurance to design and implement a relevant business interruption insurance service proposition for customers.

In the case of business interruption products for the freight transport industry, Zhen et al. (2016) proposed a model based on the work of Dong and Tomlin (2012). This model characterized the relationship between transport recovery and business interruption insurance when transport costs were uncertain and when transport recovery was deemed an endogenous factor. In the same manner, Li et al. (2018) developed a fine-grained transport insurance prototype based on blockchain and internet of things technologies. The insurance premium was evaluated on the basis of vehicle use and driver behavior. The insurance and payment model were implemented using an Ethereum framework by saving data from mobile sensors. Wang et al. (2018) contributed to the literature for the high-value transportation disruption including the value declared by the customer, the optimal insurance premium, and the strategy preference problem of the express logistics providers. Two types of contracts were developed, the additive and the multiplicative, which depend on actual probability of disruption where it is critical for the express logistics providers to be aware of the actual value of the load in order to maximize its profits. This proposal benefits both the transport company and the insured customers. In addition, studies have been based on data analysis that propose management improvements and support decision making in freight transport cases involving business interruption insurance. For example, Tatarinov and Kirsanov (2019) built an information support system aimed at managing the road transportation of dangerous goods based on a systematic approach that relies on guidance documents and that employs information and communication technologies for transmitting information from moving vehicles to duty vehicles. This system includes the load insurance during the transport by road organization stages. Wu et al. (2017) also addressed the management of logistics risks based on knowledge discovery in databases procedures with business analytics of descriptive, predictive, and prescriptive analysis to address load loss

incidents. In relation to business interruption insurance, they conceived strategic cargo loss insurance policies where they used insurance company claim databases for not only preventing the financial losses caused by incidents, but also to avoid jeopardizing the company's competitiveness. Currently, investigative research on business interruption insurance and data analytics for the freight transportation industry remains limited and inadequate. Nonetheless, the insurance industry seems to be clear about the importance of integrating big data and data analytics into activities such as product development, portfolio analysis, underwriting operations, pricing, and loss and control. As insurers venture beyond the analysis of structured transaction data to incorporate external data of all kinds, the combination and analysis could be challenging (Breading and Garth, 2014).

3. PROBLEM DEFINITION

The case study in this research is a recognized train freight transportation company servicing the U.S. railroad industry. Its railways cover thousands of miles across the U.S. eastern contiguous territory. It operates up to 1,300 trains per day, and it transports some 6.5 million carloads of products per year. Moreover, HAZMAT constitutes 7.5% of the company's yearly cargo. As with all organizations, train freight transport has not been exempted from encountering challenges in the course of its organization's development. Train freight carriers that assess, anticipate, and are informed of the various risks which could occur during train shipments are less affected by unforeseen circumstances.

More recently the managing of emerging or accelerated data growth has become a trending concern. Train freight transport companies generate and store excessive volumes of data in their organization's internal database. Consequently, the data becomes valueless if it is not analyzed. However, by employing the analyzed railway transportation data through descriptive and predictive analytics, a forecasting model was developed. This forecasting model supports the devising of customized insurance policies by calculating the probability of transportation statistics deviation from the forecasting model. Had the railway operator and insurer accepted the customized model terms as a valid trading tool, this calculation could allow the insurer to stipulate a compensation amount if the forecasted drop materialized. This mechanism would allow the devising of parametric business interruption policies that compensate the railway operator if certain transportation KPIs deviate from expectations.

This descriptive and predictive data analytic model based on historical data will identify the patterns and will generate a forecast in order to define KPIs and thresholds. Therefore, the resulting information will support decision making strategies and present insights for devising customized business interruption insurance products. In addition, it will define the baseline performance at the time of signing the policy intended for minimizing losses occurred during train freight transportation and will protect the rail freight transportation industry from unforeseen circumstances and calamities.

4. METHODOLOGY

With the exponential growth of data from businesses, data analytic methods have been considered by experts as one of the most efficient and effective solutions for managing information (Ji and Wang, 2017). Recently, this process has been optimized by the use of machine learning algorithms that have contributed to the management of large amounts of valuable information. The machine learning algorithms are implemented for searching through a set of possible predictive models and for identifying the model that best captures the relationship between the descriptive traits and the objective feature of a data set (Kelleher et al. 2015).

Moreover, the objective of this investigation is to examine all past and current information derived from an active freight transportation company by employing a descriptive and predictive analytical methodology. With the examination results, customized insurance products that consider identified patterns and predictive values that reflect the current and future freight transportation activity could be proposed. This methodology is based on the development of Python scripts that perform data extraction, cleansing, and descriptive and predictive data analyses through data analytic techniques. Subsequently, machine learning methods will be applied for predicting the evolution of the gathered indicators.

4.1 Data Wrangling

The methodology starts with gathering and processing a large amount of raw data from the active freight transportation company which is publicly available online in a PDF format.

This process, referred to as *data wrangling*, consists of cleansing, structuring, and enriching raw data into a desired format for effective and prompt decision making (McKinney, 2012).

The following tasks were performed for generating the final structured data: (*i*) data gathering from the web; (*ii*) correction of typographical errors and standardizing product titles (also referred to as *metrics*); (*iii*) deletion of empty and duplicate entries; and (*iv*) categorizing and structuring the pre-processed data according to the metrics, weeks, and years. Subsequent to the gathering of data, the methodology approach will address performing a descriptive analysis and performing forecasts. Figure 1 presents an example of the resulting structured data gathered from previous steps. Notice that the data is categorized by metric, year, and week. For the year 2020, the gathered structured data was collected only up to the 38th week of the year and does not reflect the entire year.

	Metric	Year	Value	Week
0	Grain	2013	2672	1
1	Grain Mill Products	2013	2082	1
2	Farm Products, Ex. Grain	2013	255	1
3	Food Products	2013	1674	1
4	Chemicals	2013	10087	1
...	...	...	...	...
10045	Total Carloads	2020	64783	38
10046	Trailers	2020	2109	38
10047	Containers	2020	57904	38
10048	Total Intermodal	2020	60013	38
10049	Total Traffic	2020	124796	38

[10050 rows x 4 columns]

Figure 1 – Structured data example.

4.2 Descriptive Analysis

Once the data was structured as presented in Figure 1, the descriptive analysis was initiated.

The exploration of the data was performed with data visualization techniques such as pie charts, bar charts, histograms, and time series charts. For this purpose, the methodology approach filtered the data according to the desired analysis to be performed. For example, the data was filtered by specific year(s) and/or by a specific number of weeks. However, only the first 38 weeks of the year 2020 were filtered in order to provide and perform approximate comparisons with other years within the same time period.

Subsequently, a series of statistical data analyses that included graphs and metrics were performed. Such analyses were used for extracting information that reveals the following criteria for each year: (*i*) the histogram of the total transported carloads for each product in 2019; (*ii*) the volume percentage of the 10 most transported carloads in 2019; (*iii*) the time series of transported carloads by product from 2013 to 2020; (*iv*) the mean volume of transported products per season (winter, spring, summer, and autumn of 2013 to 2020).

4.3 Predictive Analysis

The structured data was composed of a sequence of values representing each type of transported products from 2013 to 2020. As a result of the data having been categorized into weeks and years for each metric, this categorization resulted in a sequence of values over time, i.e., a time series forecasting. Given this particularity for making projections about future performance on the basis of the gathered historical and current data, i.e., the forecast, a predictive model was proposed by applying the Holt-Winters forecasting model (Chatfield and Yar, 1988), also known as the *triple exponential smoothing* for time series forecasting.

Another reason which supports the use of the Holt-Winters forecast modeling method refers to the fact that this data reveals trend and seasonality over an entire year of (52 weeks). As a result, the introduction of an additional parameter to handle seasonality is required

(Kalekar, 2004). When the model is trained with the historical data, the model becomes a machine learning method that uses previously transported volume values that support the designing of customized business interruption insurance products based on identified patterns and predictive values that protect the train freight industry form unforeseen circumstances and calamities.

Figure 2 summarizes the methodology steps beginning from the raw data extraction to its analysis and prediction of future events.

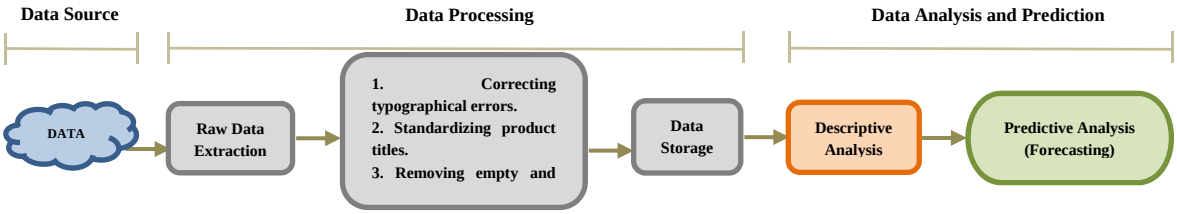


Figure 2 – Methodology Steps.

5. RESULTS AND DISCUSSION

This section provides a descriptive and predictive analysis of the train freight transportation data. In addition, a predictive model validation has been included. Moreover, this section will reveal the proposed customized insurance product functionality once all models are created. In accordance with the data structure, the following criteria was assessed for each year: (*i*) the total transported carloads for each product in 2019; (*ii*) the volume percentage of the 10 most transported carloads in 2019; (*iii*) the time series of transported carloads by product from 2013 to 2020; (*iv*) the mean volume of transported products per season (winter, spring, summer, and autumn) of 2013 to 2020; (*v*) the forecast for the year 2021, and (*vi*) the predictive model validation which included real data vs a forecast from 2019. Descriptive and predictive analytic methods were applied for completing this investigation. Furthermore, this analytical method involved monitoring data trends and patterns. The database representing the year 2019 was queried by employing a descriptive analytic method. The parsing process and descriptive analysis were developed using Python scripts.

5.1 Train Freight Transportation Descriptive Analytics

The descriptive analysis revealed the total transported carloads for each product in 2019 as depicted in Figure 3. According to the bar chart, the largest total volume carload was Coal with a total volume of 783,777 carloads. However, the least total volume carloads were Farm Products (excl. Grain) with a total volume of 9,514 carloads.

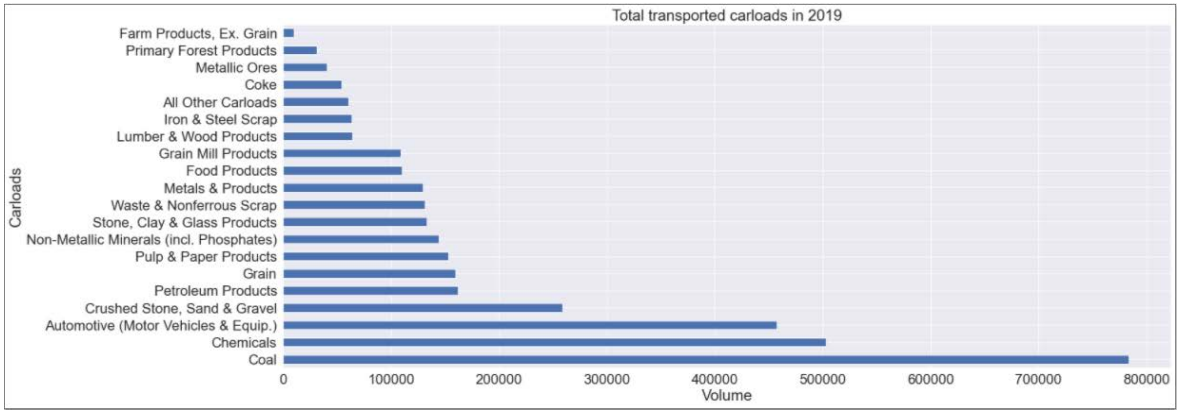


Figure 3 – Total transported carloads for each product in 2019.

Moreover, the 10 most transported carloads of 2019 are presented in Figure 4. Those carloads representing the most significant carloads were Coal with 27.1%, Chemicals with 17.4%, and Automotive (Motor Vehicles & Equipment) with 15.8%. The transport of Petroleum Products, Grain, Pulp & Paper Products, Non-metallic Minerals (Incl. Phosphates), Stone, Clay and Glass Products, and Waste and Nonferrous Scrap represented an approximate proportion of about 5% of the total.

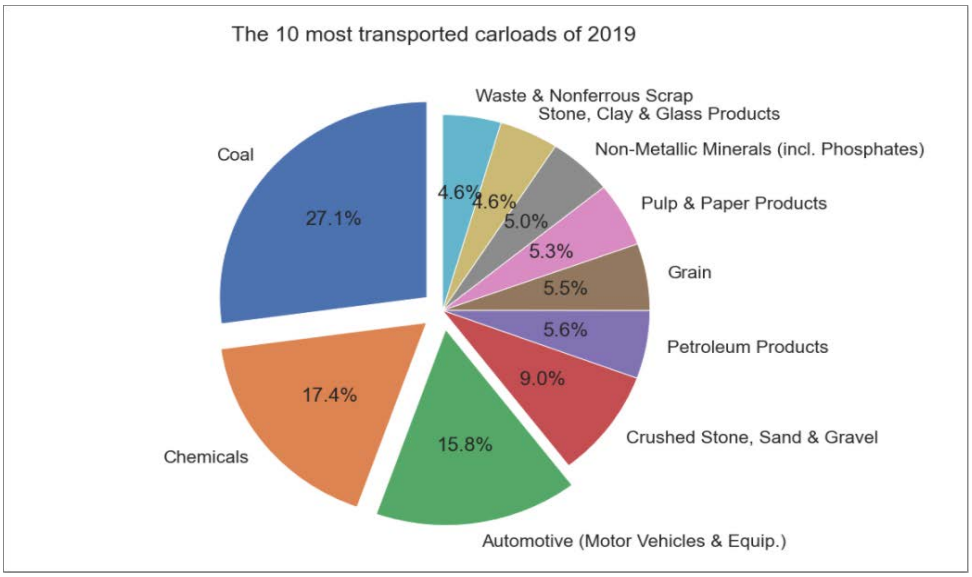


Figure 4 – Volume percentage of the 10 most transported carloads of 2019.

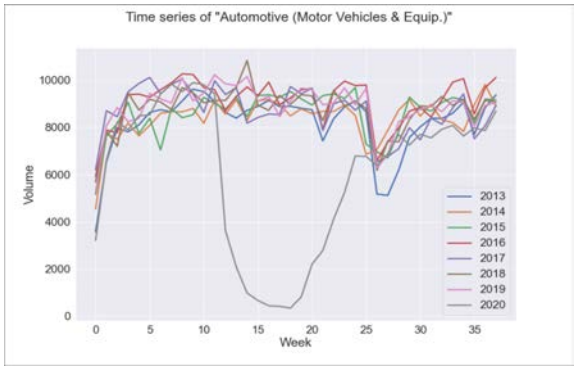
Figure 5 represents the weekly time series of four distinct transported carloads from 2013 to 2020: (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food

Products. The equivalent volumes of each carload per week provide a clear representation of the overall trend in the variability of each carload over the years. These graphs allow for a year-to-year behavior comparison for each week. For example, from 2013 to 2019, the Automotive (Motor Vehicles & Equipment) transported carloads maintained the same behavior week by week with approximately 2,000 units when compared to previous years. A similar trend can be observed for Food Products. Contrary to this, Coal volumes descended

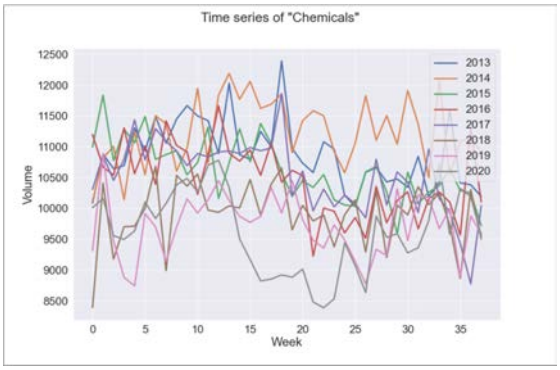
by almost 10,000 units in 2015. From then onward, volumes remained between 10,000 and 17,000 units until 2019. In the case of Chemicals, transported carload volumes had ascending and descending movements. During mid-2016, a decline in Chemical carloads became evident. Then in 2017, Chemical carloads began to ascend. However, during 2018 through the 38th week of 2020, Chemical carload volumes descended. Although the year 2019 collectively experienced the lowest Chemical carload activity when compared to other years, both 2017 and 2018 experienced the lowest Chemical carload points within a fractional period.

Moreover, the COVID-19 pandemic and its impact on carload volumes was also considered. The impact was more evident during week 10 of 2020. In particular, the Automotive (Motor Vehicles & Equipment), Chemicals, Coal and Food Products carloads were the most affected as a result of the COVID-19 preventive restrictions imposed by local governing authorities.

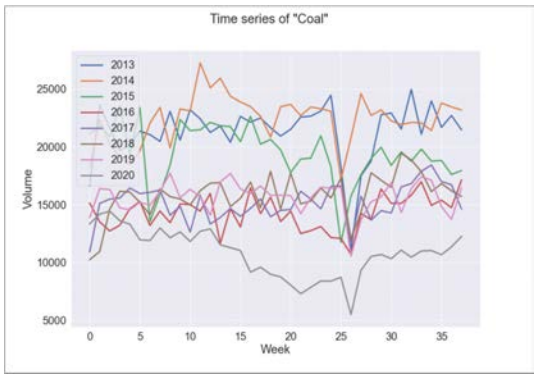
Despite the impact COVID-19 had on carload volumes, it can be observed that the need of Chemicals for producing sanitizers against COVID-19 resulted in a fast recovery for Chemicals' carload volumes. Furthermore, first necessity goods such Food Products carload volumes recovered faster than other transported goods.



(a) – Automotive (Motor Vehicles & Equip)



(b) – Chemicals



(c) – Coal

(d) – Food Products

Figure 5 – Time series of transported carloads from 2013 to 2020: (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food Products.

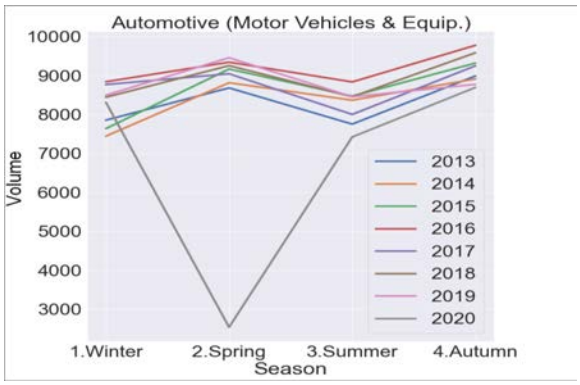
To complement the descriptive analysis, a seasonal analysis was completed. Figure 6 represents the volume of four distinct transported carloads per season (winter, spring, summer, and autumn) of 2013 to 2020 for (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food Products. The values for each season correspond to the "Mean" value of each carload. This compact structure generates a visualization of the year-by-year changes reflected in the columns. Figure 6 (a) illustrates a variability of 2,000 units between the 7,000 to 10,000 volume unit range for the Automotive (Motor Vehicles & Equipment) carload. This variability consists of comparable seasonal behavior where carload volumes increased during winter to spring and decreased during spring to summer.

The atypical behavior of transported carloads during spring of 2020 is exhibited in Figure 6 (*a*, *b*, *c*, and *d*). This atypical behavior resulted from the unexpected COVID-19 pandemic.

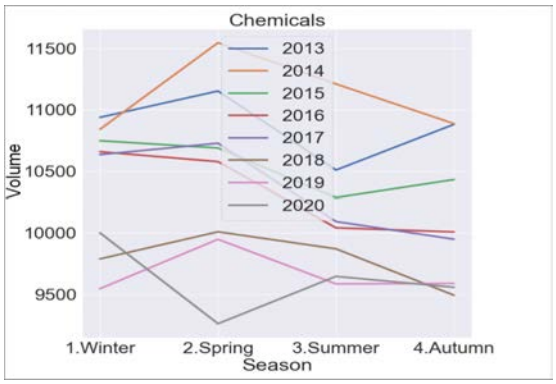
With the exception of 2020, spring was the season with the highest transported Food products carload volumes as illustrated in Figure 6 (*d*). Figure 6 (*a*, *b*, *c*, and *d*) illustrates the reactivation and recovery trend of the world economy following the COVID-19 pandemic. A similar carload volume recovery trend can be observed during autumn with the appearance of a second phase in some countries. The difference in volumes between the analyzed years remained stable since 2013 by approximately 350 units. On the other hand, Coal and Chemicals experienced greater carload volume variability from one year to the next as depicted in Figure 6 (*b*) and (*c*). Coal carload volumes for example remained at a high level from 2013 to 2015.

Although Coal carload volumes decreased between 2016 and 2017, carload volumes began ascending in 2018 and onward. This ascending behavior was particularly observed during spring of 2019 when Coal carload volumes reached its highest average volume throughout the analyzed years. Chemicals' carload volumes exhibited similar behavior in terms of year-to-year volume variability. Between 2015 and 2017,

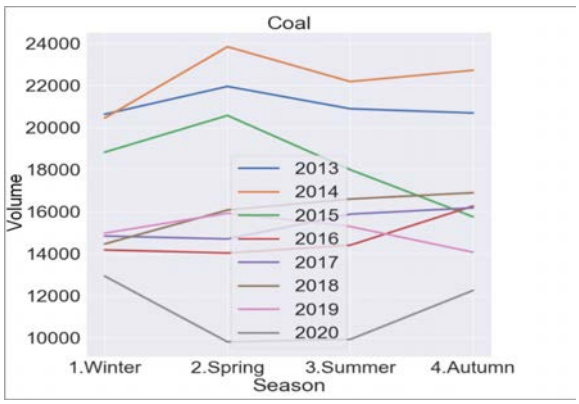
Chemicals' carload volumes followed a similar trend. A significant decrease in transported carload volumes was observed for the years 2018 and 2019. Figure 6 (*a*, *b*, *c*, and *d*) illustrates a significant decrease in transported volumes during the spring of 2020 due to the COVID-19 pandemic.



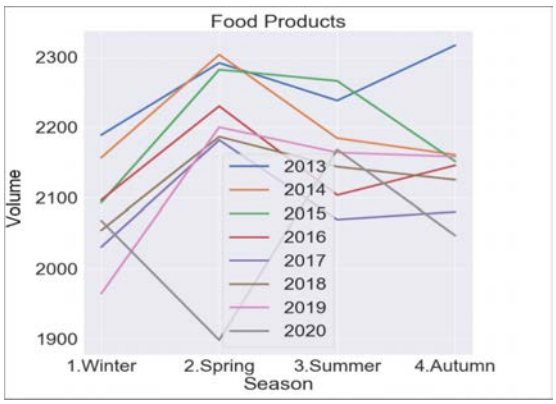
(a) – Automotive (Motor Vehicles & Equip)



(b) – Chemicals



(c) – Coal



(d) – Food Products

Figure 6 – Mean volume of transported carloads per season (winter, spring, summer, and autumn) of 2013 to 2020: (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food Products.

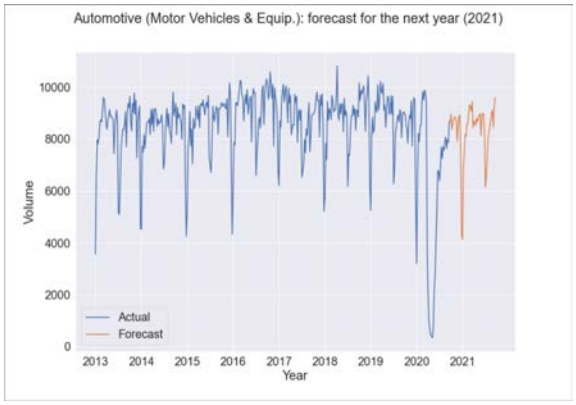
5.2 Train Freight Transportation Predictive Analytics

The Predictive analysis provided a transport carload volumes 2021 forecast for (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food Products. The forecast can be observed in Figure 7 (*a*, *b*, *c*, and *d*).

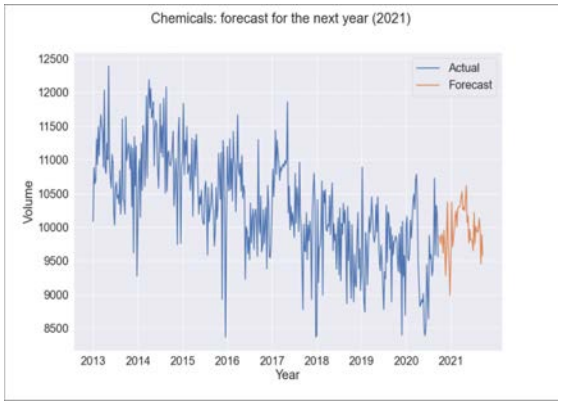
According to the one-year forecast, from week 38 of 2020 to week 38 of 2021, two significant decreases in transported carload volumes were anticipated in the Automotive (Motor Vehicles & Equipment) carloads. A decrease in transported carload volumes was expected towards the end of 2020. However, an increase in carload volumes was expected during the subsequent weeks. This increase of carload volumes is expected to maintain a stable flow until June 2021 when a decrease in carload volumes is anticipated. In the case of Chemicals, carload volumes were expected to remain stable during 2020. Then in 2021 onward, the carload volumes of Chemicals would begin an ascending trend. Chemicals' carload volumes similar to the previous year can be expected. In contrast, the transportation of Coal carload volumes descended since mid-2014. The forecast for Coal carload volumes anticipates a descending trend falling under carload volumes from the previous years.

In the case of Food Products, a comparable trend to Coal is anticipated, however, with less variability as has occurred during the previous years.

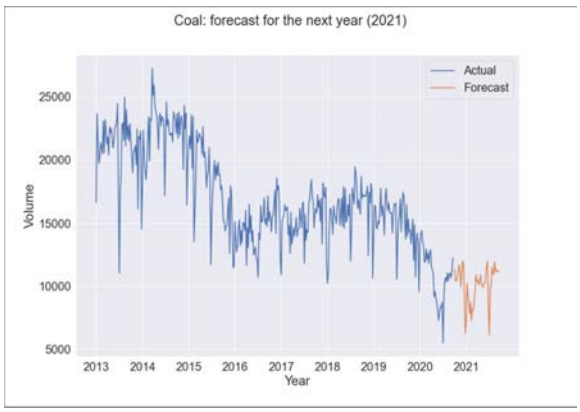
Moreover, a Predictive analysis could support decision-making by analyzing data patterns. As a result, business interruption insurance product thresholds can be set and defined with premiums established by the Insurer and the Insured in the event of exceeding thresholds.



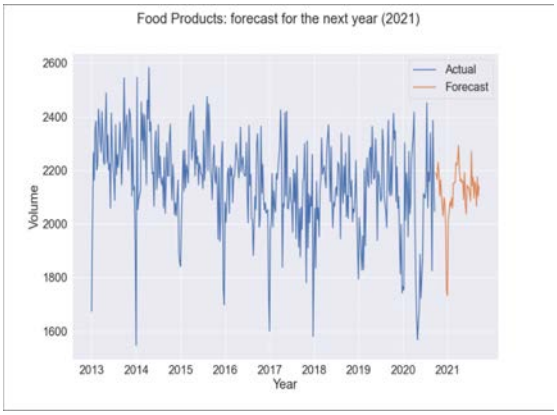
(a) – Automotive (Motor Vehicles & Equip)



(b) – Chemicals



(c) – Coal



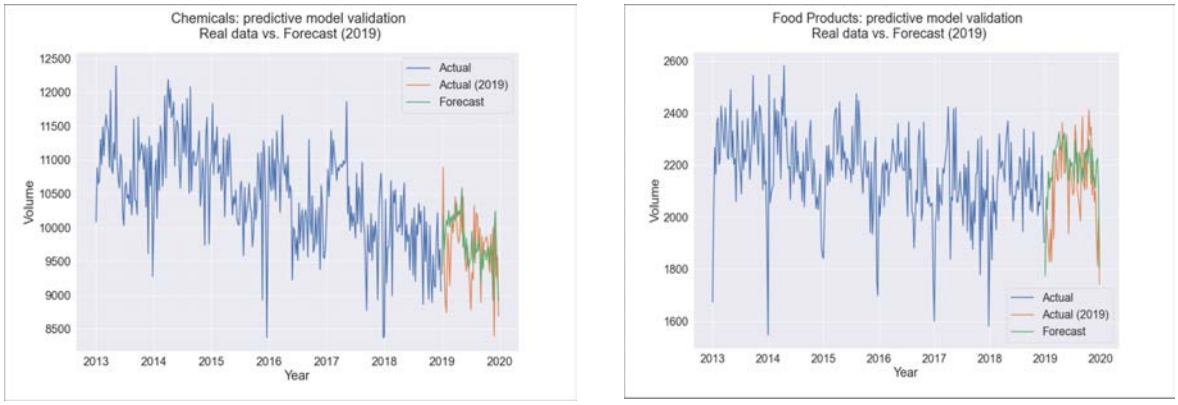
(d) – Food Products

Figure 7 – 2021 Forecast for (*a*) Automotive (Motor Vehicles & Equipment), (*b*) Chemicals, (*c*) Coal, and (*d*) Food Products.

5.3 Predictive Analytics: Model Validation

The predictive model was validated with the cross-validation method which is a statistical method used for estimating the skills of machine learning models. The cross-validation method examined data from 2013 to 2019. The examined data was subsequently used as training data for this model. The validation data corresponds to the year 2019. Figure 8 (*a*) and (*b*) illustrates the predictive model validation for (*a*) Chemicals and (*b*) Food Products carloads, and compares Real Data with that of a 2019 Forecast. The comparison distinguishes actual 2019 data from predictive data that indicates what was expected to occur during 2019.

The predictive model graphs of both Chemicals and Food Products carloads exhibit the transported values for each period of the year. However, the predictive model graphs illustrate that during the first weeks of 2019, both Chemicals and Food Products carloads would have higher values than those actually obtained. Despite this difference, both the predictive model and the actual values indicate similar trends. In the second half of the year, the model values were closer to the actual values and accurately described the transported quantities of each product.



(a) – Chemicals (b) – Food Products

Figure 8 – Predictive model validation: Real data vs. 2019 Forecast for (*a*) Chemicals and (*b*) Food Products.

Altogether, the values predicted by the model do not exactly match the actual values. In addition, as illustrated in Figure 8 (*a*) and (*b*), the model does not predict statistical data points or outliers which differ from the actual values such as volume increases and decreases.

Therefore, the validation revealed that predictive models based on the Holt-Winters forecasting method allow for predicting data behavior. Nevertheless, the model does not anticipate unexpected events. The model does however accurately predict behavior of values for each week.

5.4 Train Freight Transportation Customized Insurance Product

Once all models are created, the forecasting system will support the devising of customized insurance policies based on KPIs or metrics that trigger a compensation or payout upon satisfying the previously agreed conditions stipulated in the policies.

To understand how a customized insurance product works, the developments of the novel Coronavirus disease of 2019 or COVID-19 have been considered in this insurance policy “example”. In December 2019, COVID-19 which is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was detected. Nevertheless, it was not until March 2020 that COVID-19 was declared an official world pandemic. Assuming that prior to the COVID-19 outbreak, an insurance policy for the transportation of Automotive (Motor

Vehicle & Equip.) carloads stipulating coverage throughout the entire year of 2020 was devised and bound, and had been in effect noting the following condition:

- *Condition: The policy begins to pay once the transported carloads fall below a threshold of 7,000 carloads.*

Under this condition, will the railway operator receive a payout? As previously illustrated in Figure 5 (a), there was a significant decrease in the transport of Automotive (Motor Vehicle & Equip.) carloads in March 2020 when COVID-19 was officially declared a world pandemic and preventive measures were imposed. It is evident that throughout weeks 10 to 25, the carloads dropped below the 7,000 carloads threshold. In such an event, a payout would be triggered as it satisfies the condition. The insured or policyholder would receive the agreed compensation limit for business losses, approximated by the drop in cargo versus the expectations derived from the forecasting model. Such a policy mechanism, while applied to different cargo types, would offer the railway operator a statistical option for hedging potential business losses due to unexpected events. The fact that the forecasting model is fixed prior to the policy being devised and bound allows for both parties to agree on the algorithm as a feasible trading mechanism.

6. CONCLUSIONS AND FURTHER DEVELOPMENTS

As businesses emerge throughout the world, data has become an essential element in their operational development and their economic growth. In this study, a series of descriptive and predictive data analysis from an existing freight transportation company was performed for devising customized insurance products aimed at identifying existing data values and patterns, examining current data, and for predicting losses that could occur during freight transportation. In this regard, Python scripts for extracting and analyzing raw data were developed. In addition, machine learning strategies for predicting events that could result with losses were adopted.

The revealed information from the analyzed data was subsequently used for adopting resolutions and supporting decision making. Moreover, the data acquired from the 10 most transported carloads revealed informative discrepancies that could prevent businesses from reaching their operational and economical potential.

The employed analytical methods in this study facilitate the designing of customized business interruption insurance products for protecting businesses from losses resulting from unexpected events or disruptions such as the COVID-19 pandemic. For example, the data analysis performed for supporting this study revealed those carload volumes that decreased during transportation and those carload volumes that recovered during the developments of the COVID-19 pandemic. Furthermore, as world economies grow and business operations thrive, data analytics is emerging in a short period of time.

In an effort to protect businesses from sustaining financial losses as a result of unexpected events and disruptions, a daily data analysis rather than a weekly analysis could possibly improve the accuracy in predicting fluctuating data behavior. Data is knowledge, and with knowledge, businesses will be predisposed with the necessary criteria for predicting operational activities, making decisions, and for acquiring reliable insurance products.

ACKNOWLEDGMENTS

This study was collectively completed and supported by Guy Carpenter & Company, LLC, and the Universitat Oberta de Catalunya.

REFERENCES

BREADING, M and GARTH, D. (2014). Big data in insurance. Beyond experimentation to innovation/M. Breading. SMA.

CHATFIELD, C and YAR, M. (1988). Holt-winters forecasting: some practical issues. Journal of the Royal Statistical Society: Series D (The Statistician), 37(2), 129–140.

DONG, L and TOMLIN, B. (2012). Managing disruption risk: The interplay between operations and insurance. Management Science, 58(10), 1898–1915.

FREES, E.W. (2015). Analytics of insurance markets. Annual Review of Financial Economics, 7, 253–277.

GAGATSI, E; GIANNOPOULOS, G; and AIFANDOPOULOU, G. (2014). Supporting policy making in maritime transport by means of MultiActors multi-criteria analysis: A methodology developed for the Greek maritime transport system. Proceedings of the 5th transport research arena (TRA), April, 14–17.

GANAPATHY, V. (2017). A public-Private Partnership Model for Managing Disasters in India. IBMRD's Journal of Management & Research, 6(2),38–65.

JI, W and WANG, L. (2017). Big data analytics based fault prediction for shop floor scheduling. Journal of Manufacturing Systems, 43, 187–194.

KALEKAR, P.S. (2004). Time series forecasting using holt-winters exponential smoothing. Kanwal Rekhi School of Information Technology, 4329008(13), 1–13.

KELLEHER, J.D; MAC NAMEE B; and D'ARCY, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms. Worked Examples, and Case Studies.

KELLER, B; ELING, M; and SCHMEISER, H. (2018). Big data and insurance: implications for innovation, competition, and privacy. Geneva Association-International Association for the Study of Insurance Economics.

LI, Z; XIAO, Z; XU, Q; SOTTHIWAT, E; GOH, R.S.M; and LIANG, X. (2018). Blockchain and IoT data analytics for fine-grained transportation insurance, In 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS). (pp. 1022–1027). IEEE.

MCKINNEY, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. “O’Reilly Media, Inc.”.

MIZGIER, K.J; KOCSIS, O; and WAGNER, S.M. (2018). Zurich Insurance uses data analytics to leverage the BI insurance proposition. Interfaces, 48(2), 94–107.

NETO, A.G; MARUJO, L.G; COSENZA, C.A.N; D´ORIA, F; and LIMA Jr, J.M. (2012). Using fuzzy NPV evaluation to justify the acquisition of business interruption insurance. Expert Systems with Applications, 39(12), 10821–10831.

PARR-RUD, O. (2012). Drive your business with predictive analytics. SAS Institute.

SAMSON, W.D and PREVITS, G.J. (1999). Reporting for success: The Baltimore and Ohio railroad and management information, 1827-1856. Business and Economic History 28(2), 235–254.

TATARINOV, V and KIRSANOV, A. (2019). Information support for safety insurance of road transport of dangerous goods, In IOP Conference Series: Materials Science and Engineering (Vol.492, No. 1, p. 012006). IOP Publishing.

WANG, H; TAN, J; GUO, S; and WANG, S. (2018). High-value transportation disruption risk management: Shipment insurance with declared value. Transportation Research Part E: Logistics and Transportation Review, 109, 293–310.

WU, P.J; CHEN, M.C; and TSAU, C.K. (2017). The data-driven analytics for investigating cargo loss in logistics systems. International Journal of Physical Distribution & Logistics Management.

ZAJDENWEBER, D. (1996). Extreme values in business interruption insurance. Journal of Risk and Insurance, 95–110.

ZHEN, X; LI, Y; CAI, G.G; and SHI, D. (2016). Transportation disruption risk management: business interruption insurance and backup transportation. Transportation Research Part E: Logistics and Transportation Review, 90, 51–68.