Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 58 (2021) 393–400

Transportation Research Procedia

www.elsevier.com/locate/procedia

14th Conference on Transport Engineering: 6th – 8th July 2021

Using Data Analytics & Machine Learning to Design Business Interruption Insurance Products for Rail Freight Operators

John F. Cardona[a], Juliana Castaneda[a], Leandro do C. Martins[a], Mariem Gandouz[a], Angel A. Juan[a], Guillermo Franco[b]*

[a]*IN3-Dept.of Computer Science, Universitat Oberta de Catalunya, Barcelona, Spain*
[b]*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.

Keywords: Data analytics; machine learning; business interruption; insurance; rail freight.

1. Introduction

Data has become essential for the operational development and economic growth of many global organizations. As organizations 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 are welcoming innovative solutions and methods for managing large 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. The data can

* Corresponding author. Tel.: +34-674-50-7793.
E-mail address: jcardonadi@uoc.edu

10.1016/j.trpro.2021.11.053

identify factors which prevent an organization from reaching and maintaining economic growth. Moreover, train freight carriers have encountered unforeseen challenges during transportation of cargo. Increased risks of accidents, loss of goods, and adverse environmental consequences have been the most significant challenges. For this reason, the design of Business Interruption Insurance products for protecting the train freight industry from unforeseen events is imperative. The proposed insurance product or mechanism would parametrically pay when the actual performance falls below the agreed performance thresholds. This mechanism would obviate the need to conduct claims or forensic analysis and therefore permit the insured to promptly recover without legal disputes. This paper will address the use of data analytics and machine learning methods for designing Business Interruption Insurance Products. 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. Currently, data analytic methods could be considered the most efficient and effective solution for managing data (Ji and Wang, 2017). In this regard, 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 be compared with current performance developments. The presented forecast method will support 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 stipulate a payout that considers the real transported goods against expected transported goods over a certain period of time. The policy would behave as a compensation for aggregate drops rather than 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. The forecasts are used with machine learning models for proposing customized insurance products based on identified patterns that analyze current and future freight transportation activity. The Introduction will continue on to: 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. As a result, cargo owners sustain exorbitant monetary losses. 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 (Ganapathy, 2017). 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). According to Gagatsi et al. (2014), the devising of transportation insurance policies is an intricate process requiring extraordinary diligence from all stakeholders. Business interruption insurance products are devised for protecting companies from losses, and stipulate coverage limitations that define three crucial elements: (*i*) the premium or the price paid by the interested company for obtaining insurance coverage; (*ii*) the coverage limit or the maximum amount paid by the insurer in case of a loss; and (*iii*) the deductible or the monetary value of the loss absorbed by the insured company. Beforehand, insurance providers must access prospective insurance holders' facilities for assessing possible loss values (Dong and Tomlin, 2012). Keller et al. (2018) reported that data analytics has played a fundamental role in insurance policies allowing them to evolve from “intuitive bets” to an industry based on logical calculus and decision making. According to Frees (2015), “insurance is a data-driven industry” which is linked to data and models of uncertainty. 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. 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. 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, the non-parametric regression statistical methods, and 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. In Zurich, Switzerland, Mizgier et al. (2018) developed a project between Zurich Insurance and the Swiss Federal Institute of Technology. The results from analyzed data enabled Zurich Insurance to design and implement a business interruption insurance service model for customers. Similarly, 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. Moreover, 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, Tatarinov and Kirsanov (2019) built an information support system aimed at managing the 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. 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. Currently, business interruption insurance research 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 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

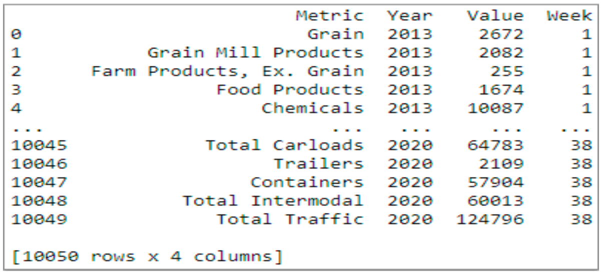
The case study is a recognized train freight transportation company servicing the U.S. railroad industry. Its railways cover thousands of miles across the U.S. eastern contiguous territory. It operates up to 1,300 trains per day, and it transports some 6.5 million carloads of products per year. Train 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. 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 will identify the patterns and will generate a forecast for defining 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 and protecting the rail freight transportation industry.

4. Methodology

Data analytic methods have been considered as an efficient and effective solution for managing information (Ji and Wang, 2017). Machine learning algorithms were 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 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; (*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.



```
                         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 data was filtered 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.

4.3. Predictive Analysis

The structured data was composed of a sequence of values representing each type of transported products from 2013 to 2020. 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.

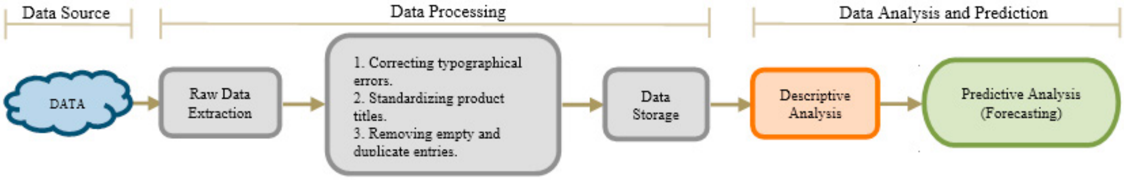

Data Source
Data Processing
Data Analysis and Prediction
DATA
Raw Data Extraction
1. Correcting typographical errors.
2. Standardizing product titles.
3. Removing empty and duplicate entries.
Data Storage
Descriptive Analysis
Predictive Analysis (Forecasting)

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.

5.1. Train Freight Transportation Descriptive Analytics

The descriptive analysis revealed the 10 most transported carloads of 2019 presented in Figure 3. Those carloads representing the most significant carloads were Coal with 27.1%, Chemicals with 17.4%, and Automotive (Motor Vehicles & Equip.) 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.

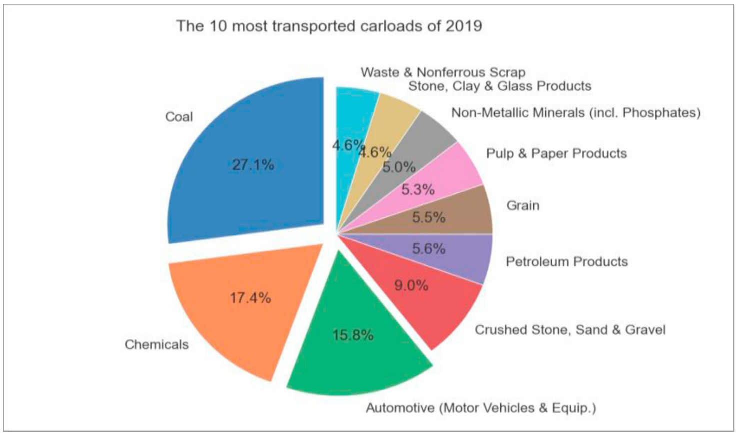


Figure 3 – Volume percentage of the 10 most transported carloads of 2019.

Figure 4 represents the weekly time series of two distinct transported carloads from 2013 to 2020: (a) Automotive which includes (Motor Vehicles & Equip.) and (b) Chemicals. 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. For example, from 2013 to 2019, the Automotive transported carloads maintained the same behavior week by week with approximately 2,000 units when compared to previous years. In the case of Chemicals, transported carload volumes had ascending and descending movements. 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 and Chemicals 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.

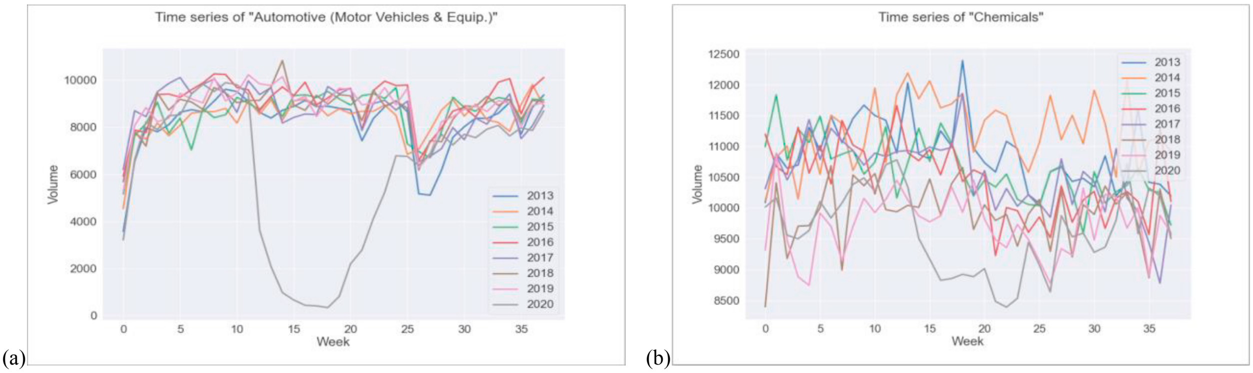


Figure 4 – Time series of transported carloads from 2013 to 2020: (a) Automotive (Motor Vehicles & Equipment) and (b) Chemicals.

5.2. Train Freight Transportation Predictive Analytics

The Predictive analysis provided a transport carload volumes 2021 forecast for (a) Chemicals and (b) Food Products. The forecast can be observed in Figure 5 (a and b). According to the one-year forecast, from week 38 of 2020 to week 38 of 2021, 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. The forecast for Food Products carloads volumes anticipates a descending trend falling under carload volumes from 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.

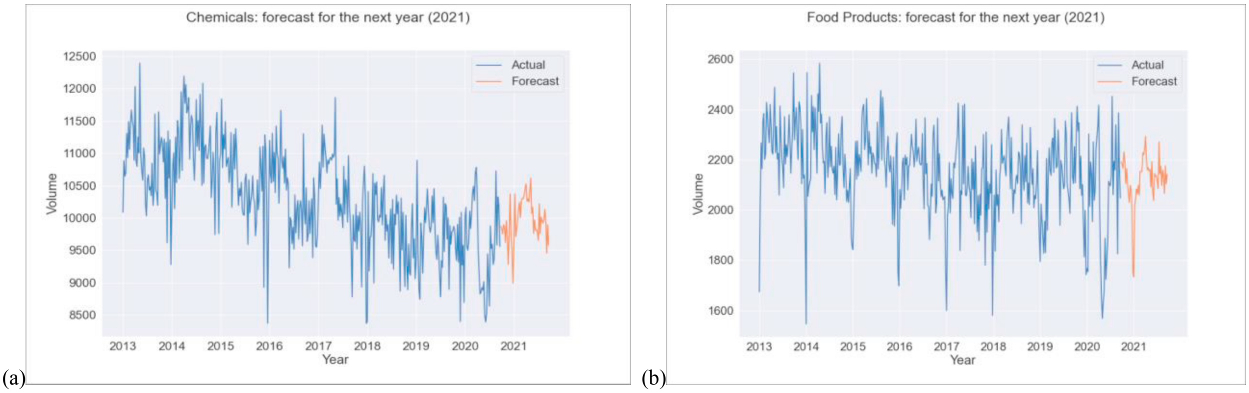


Figure 5 – 2021 Forecast for (a) Chemicals and (b) Food Products.

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 examined data from 2013 to 2019 was subsequently used as training data for this model. Figure 6 (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.

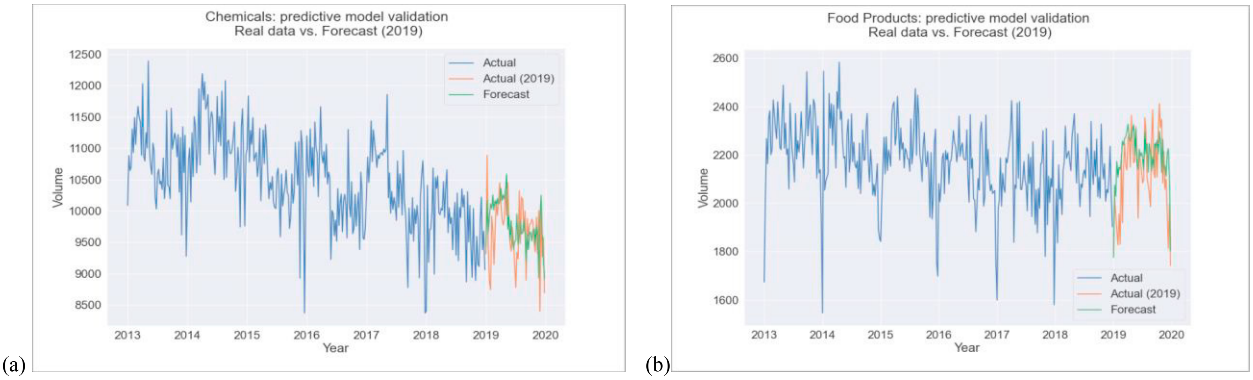


Figure 6 – Predictive model validation: Real data vs. 2019 Forecast for (a) Chemicals and (b) Food Products.

As illustrated in Figure 6 (a) and (b), 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 the functionality of a customized insurance product, 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 carloads stipulating coverage throughout the entire year of 2020 was devised and bound, and had been in effect noting the following 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 4 (a), there was a significant decrease in the transport of Automotive 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 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 supporting decision making. 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.

Acknowledgements

This study was 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).

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.

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.